Identification of host gene expression biomarkers for tuberculosis

Myrsini Kaforou

Section of Paediatrics, Division of Infectious Diseases, Department of Medicine

&

Department of Genomics of Common Disease, School of Public Health

Imperial College London

Submitted for the degree of Doctor of Philosophy

London, 2015

“However many mountains you may climb, from their peaks

You will see yet higher peaks, another landscape beckoning you on;

And when you have scaled the highest of them all,

You will still find yourself under the stars.”

Kostís Palamás

The Twelve Lays of the Gipsy

To my parents,

Abstract

The presence of disease, including infectious disease, has been observed to give rise to specific patterns of gene expression in peripheral whole blood, regardless of disease site. These gene expression signatures allow for distinction between diseases and have the potential to reform diagnostics, particularly in diseases and patient groups for whom current diagnostics are unreliable, like Tuberculosis (TB). Although TB is a treatable infectious disease, it has high morbidity and mortality, especially in low resource countries and HIV infected patients. In this thesis, I propose a bioinformatics toolbox that derives minimal transcriptomic signatures from microarray datasets acquired from heterogeneous groups regardless of underlying co-infections and geographic locations. The transcripts' expression values are then aggregated into a single value disease risk score (DRS) for every patient, that allows for classification between the disease groups in a binary manner. The toolbox was employed to analyse an adult and a paediatric TB transcriptomic study, comprising HIV infected and uninfected patients from sub-Saharan Africa. In the adult study, the DRS based on a 27-transcript signature distinguished culture confirmed TB from latent TB infection (LTBI), while 44 transcripts distinguished TB from other diseases phenotypically similar to TB (OD), with high sensitivity and specificity. Out-of-sample validation was performed using a publicly available dataset. In the paediatric study, a 51-transcript signature distinguished TB from OD and a 42-transcript signature from LTBI. The signatures were validated out-of-sample using an independent cohort and benchmarked against culture-negative TB patients and Xpert® MTB/RIF, currently used for detection of *M. tuberculosis*. This thesis provides proof of principle that minimal host blood transcriptional signatures are able to distinguish TB from LTBI and OD regardless of HIV infection. The subsequent transformation of the signatures into a score for every patient may facilitate disease categorisation and potentially development of diagnostic tools.

Acknowledgements

First and foremost, I would like to thank my main supervisor Prof. Michael Levin, for giving me the opportunity to work on such an interesting project and for his support, trust and encouragement over the past four years. His passion about medical research, his enthusiasm and his vision, which he has instilled upon me, have been invaluable for keeping me going and putting the research questions into their real perspective. I would also like to thank my second supervisor Dr. Lachlan Coin, who has been a very enlightening and motivating teacher for me, as well as a friend. I am grateful for his support and patience, for sharing with me his expertise in the field and for trusting me. I feel the need to thank deeply both my supervisors for helping me to grow as a scientist, for offering me opportunities to present our work in meetings and conferences, for the trips they have generously given me and for genuinely caring about me. Many thanks go to my third supervisor Prof. Giovanni Montana for his crucial guidance during my first steps in the world of bioinformatics and statistical modelling. Furthermore I would like to thank all the members of the ILULU consortium that made this TB work happen and for trusting me with valuable data.

The collaborative atmosphere in the Department of Paediatrics at Imperial College was a great asset for the past four years and for that reason I would like to thank everybody and especially Prof. Beate Kampmann and Prof. Paul Langford for scientific advice and thought-provoking discussions. Special thanks go to Dr. Victoria Wright for her multifaceted support, her biological insight and hard work for the microarray analysis that aided the completion of this thesis, and to Dr. Jethro Herberg for the exciting work we carried out together, for his valuable clinical perspective and for caring for me. Also, I would like to say a very big thank you to all the members of Mike's group for their friendship and support, our endless scientific discussions and the great time we have had together during trips, lunch breaks and nights out. Thank you Shea, Vanessa, Stephanie, Harriet, Meg, Clive and Aubrey.

A big thank you goes to my good friends from Greece and London for being by my side during this journey. Especially to my flatmates Christina, Giota and Ersi, for the great times we have had together and for taking good care of me, when I needed it most. Special thanks also go to my friends with whom we went through the PhD phase of life together: to Adam for being the best listener and for sharing explorations, to Marta and Amalia for cherishing me with an unexpected friendship and for always being there for me and to Evan for all his scientific help, for making the long hours at

Imperial so much more amusing and for standing by me in the ups and the downs in this PhD and in general.

A massive thank you from the bottom of my heart goes to my family. To my parents for their endless unconditional love and support; for raising me with a love of science, encouraging me be curious and be seeking for knowledge. Thank you dad for showing to me that there is no mountain high enough in life if you believe in yourself and thank you mum for your constant care in every possible way, even when you are thousands miles away. My biggest thank you goes to my little sister, Daphne, my best friend, my soul mate. Your constant love, your encouragement and your understanding mean the world to me. Thank you for the big hugs, the smiles and the melodies you bring into my life.

London 2015

Table of Contents

List of Figures

List of Tables

Abbreviations

Abbreviation	Definition
AFB	Acid fast bacilli
AUC	Area Under the Characteristic curve
BCG	Bacillus Calmette-Guérin
bp	base pairs
cDNA	complementary DNA
CFP-10	Culture Filtrate Protein 10
cm	centimetre
CRP	C-reactive protein
CV	Cross Validation
DE	Differential Expression
DNA	Deoxyribonucleic acid
DRS	Disease Risk Score
ELISA	enzyme-linked immunosorbent assay
ESAT-6	Early Secretory Antigen Target 6
EST	expressed sequence tag
FC	Fold change
FDR	False discovery rate
HC	Healthy Controls
HIV	Human immunodeficiency virus
IFN-γ	Interferon Gamma
IGRA	Interferon Gamma Release Assay
IL-12	Interleukin 12
LRTI	Lower respiratory tract infection
LTBI	Latent tuberculosis infection
M.TB	*Mycobacterium tuberculosis*
MDR-TB	Multidrug-resistant tuberculosis
MHC	major histocompatibility complex
mRNA	messenger RNA
MSE	Minimum Square Error
NAAT	nucleic acid amplification tests
NGS	Next generation sequencing
nm	nanometre
NPV	Negative Predictive Value
OD	Other Diseases
OLS	ordinary least squares
PBMC	Peripheral blood mononuclear cell

PCA	Principal component analysis
PCR	Polymerase Chain Reaction
pg	picograms
PJP	*Pneumocystis jirovecii* pneumonia
PPD	Purified Protein Derivative
PPV	Positive Predictive Value
QC	Quality Control
RNA	ribonucleic acid
RNA-seq	RNA sequencing
rRNA	ribosomal RNA
RSN	Robust Spline Normalisation
RT-PCR	polymerase chain reaction
SA	South Africa
SDE	Significantly differentially expressed
SLE	Systemic Lupus Erythematosus
TB	Tuberculosis
TNF-α	tumour necrosis factor alpha
tRNA	transfer RNA
TST	Tuberculin Skin Testing
UTI	Urinary tract infection
WHO	World Health Organisation
µm	micrometre

Declaration of originality

This dissertation describes my work jointly supervised by Prof. Michael Levin at the Department of Paediatrics and Dr. Lachlan Coin at Department of Epidemiology and Biostatistics at Imperial College, London. It is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy. The work described here has not been submitted for any degree, diploma, or any other qualification. This dissertation is the result of my own work and the guidance of my supervisors, except what is described in the text below. Use of the first person plural is for reasons of clarity.

The microarray datasets as well as the clinical information and databases were made available to me via my supervisors. The patient recruitment described in Chapter 3 was part of the EU ILULU study led by Prof. R.J. Wilkinson, Prof. N. French, Prof. H. M. Dockrell and Dr. A. C. Crampin, while the microarray laboratory work was undertaken by Dr. V. J. Wright and Dr. L. Ling under the supervision of Prof. M. L. Hibberd at the Genome Institute of Singapore. The paediatric recruitment described in Chapter 4 was partly led by Dr. S. T. Anderson, Prof. B. Eley, Prof. R. Heyderman funded by the EU, while Kenyan recruitment and clinical classification of cases was conducted through a Wellcome Trust fellowship awarded to Dr Andrew Brent. The microarray laboratory work was undertaken by Dr. L. Ling under the supervision of Prof. M. L. Hibberd at the Genome Institute of Singapore. Measurements of CRP in patients' plasma were undertaken by Dr. S. Hamilton, a Research Associate in the group of Prof. M. Levin.

Myrsini Kaforou

London, January 2015

Copyright Declaration

Publications

Publications' list arising from the work described in this thesis at the time of submission:

1. **Kaforou M.**, Wright V.J., and Levin M., *Host RNA signatures for diagnostics: An example from paediatric tuberculosis in Africa.* (2014) Journal of Infection vol. 69 p. 28-31
2. Anderson S.*, **Kaforou. M.***, Brent A. J.*, Wright V.J.*, Banwell C. M., Chagaluka G., Crampin A. C., Dockrell H. M., French N., Hamilton M. S., Hibberd M. L., Kern F., Langford P. R., Ling L., Mlotha R., Ottenhoff T. H., Pienaar S., Pillay V., Scott J. A., Twahir H., Wilkinson R. J., Coin L. J., Heyderman R. S., Levin M. Eley B., ILULU Consortium and the Kids Tb Study Group *Diagnosis of Childhood Tuberculosis and Host RNA Expression in Africa.* (2014) New England Journal of Medicine 370 (18): 1712
3. **Kaforou. M.***, Wright V.J.*, Oni T.*, French N.*, Anderson S. T., Bangani N., Banwell C. M., Brent A. J., Crampin A. C., Dockrell H. M., Eley B., Heyderman R. S., Hibberd M. L., Kern F., Langford P. R., Ling L., Mendelson M., Ottenhoff T. H., Zgambo F., Wilkinson R. J., Coin L. J., Levin M. *Detection of Tuberculosis in HIV-Infected and -Uninfected African Adults Using Whole Blood RNA Expression Signatures: A Case-Control Study.* (2013) PLoS Medicine 10(10) e1001538
4. Levin M., **Kaforou M.**, Herberg J., Wright V.J., Coin L. J., *Method for Calculating a Disease Risk Score*, Imperial Innovations, PCT/GB2013/05022, 2013.

* Joint first author

Chapter 1

Introduction

Chapter 1 describes the background regarding the biological aspects of the thesis, focusing on how gene expression profiling can help elucidate the host response to external stimuli, such as an infection. I discuss how gene expression can further our understanding of the mechanisms underlying disease and I present the main laboratory methods to measure gene expression, focussing on microarrays and describing how microarrays have been used to study disease, and particularly infectious diseases. I then focus on tuberculosis (TB), a treatable disease that given its current global burden and lack of reliable diagnostics, is in urgent need of better biomarkers that may be used to improve diagnosis. Previous studies that have used blood gene expression profiling to identify TB biomarkers are reviewed. Finally I describe how the research presented in this thesis might overcome some of previous studies' limitations and lead to improved diagnostics tests for TB.

1.1 Host response to infection

The human body has a range of external defence barriers to avoid invasion by pathogens which are able to cause infection and disease. When these external mechanisms fail and the pathogens manage to override them, the immune system - a complex internal system of organs, tissues, cells, compounds, molecules and chemical processes - will be activated to ensure the survival of the host. The immune system is divided into the innate and the adaptive systems, both of which use distinct cells and molecules to fight against pathogens but they also interact and complement each other (Figure 1). The innate immune system and its components (physical and chemical barriers, phagocytic cells, innate leukocytes and plasma proteins) are always present and active in an immunocompetent organism. Upon pathogen invasion, the innate immune system identifies the infectious agents through pattern recognition receptors (proteins produced in the innate immune system cells), which remain either attached to the cell surface, or are soluble in the blood and have the ability to recognise microbe-specific molecules. Upon recognition, the receptors trigger mechanisms responsible for the direct killing of the microbe (including the phagocytic and the natural killer cells) and activate the adaptive immune system as well. The receptors are also central for the induction of cytokine synthesis and secretion. Cytokines are small proteins that act as mediators, allowing communication between cells. The role of cytokines is crucial for containing infections, as they can alter cellular functions of the target cells even if they are located at distant sites.

The activation of the specialised antigen-presenting cells in the innate immune system leads to the triggering of the adaptive immune system. The adaptive immune system is a network of cells (T-cells, B-cells) that are characterised by high affinity when it comes to the recognition and the elimination of microbes and foreign antigens. Different types of T cells are activated depending on the type of antigen (cytotoxic T-cells for major histocompatibility complex class I (MHCI) antigen presentation, T-helper cells for MHCII). The first time that the immune system encounters an infectious agent, the antigen presenting cell present its antigens to T-cells which move from the lymph nodes to the site of infection to assist with the killing of the invading organism. Helper T-cells play central roles in the fight against the microbe. They induce the killing of the microbe by phagocytic cells, they trigger the activation of other T-cells, but most importantly they stimulate the

activation of B-cells to make antibodies against the new invading pathogen (primary response). The antibodies are proteins that consist of constant and variable regions. Their variable regions are antigen-binding sites with high-specificity to the antigen and can either flag a microbe for other cells to attack it, disrupt the entry of the microbe to cells by attaching to it, inactivate its toxins or eradicate it by activating other pathways (i.e. complement). The B-cells divide into "clones" and each "clone" produces antibodies with the same unique binding site. One of the main characteristics of the adaptive immune system is that it exhibits "memory", so while some of the B-cells (plasma cells) will be producing and secreting antibodies against an antigen, a small number of others (memory B-cells) will remain in the body in a dormant form for a long time (days to decades). The memory cells will allow a much more rapid immune response, in subsequent exposures to the same antigen (secondary response) [1].

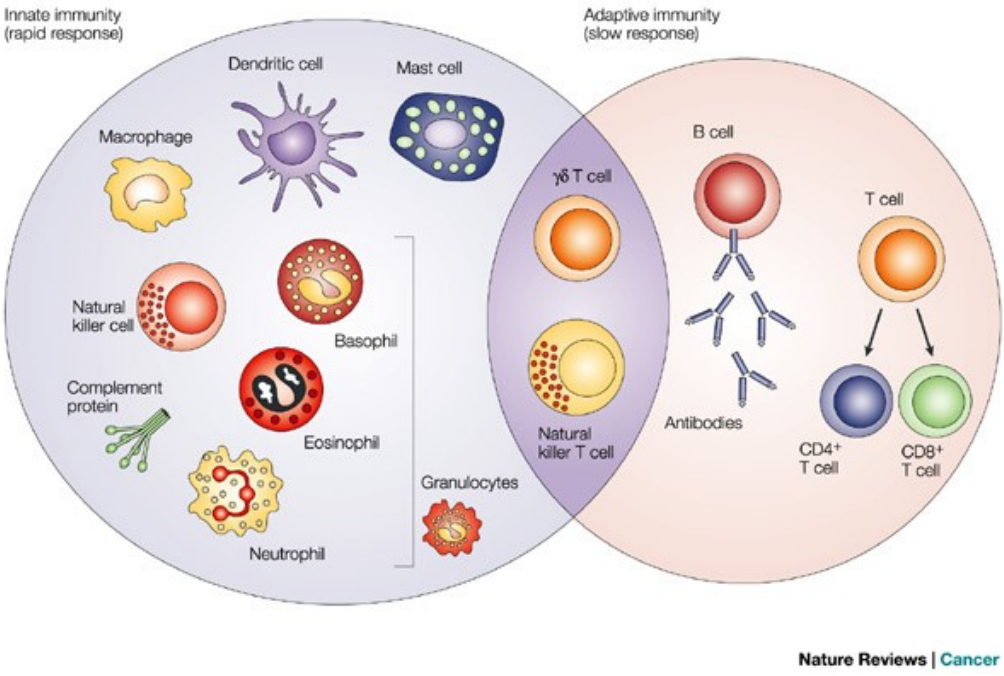


Figure 1: The innate and adaptive immune response. Reprinted with permission from Nature Reviews|Cancer [2].

Induction of the immune response and the intracellular signalling of the immune cells trigger a biochemical chain of events, leading to production of molecules that are needed for defence against infection. In order to undergo activation and release of the chemicals and proteins that are responsible for killing invading pathogens, the immune cells

alter their gene expression and modify their protein synthesising machinery to meet the current needs. There are individual genes that govern these responses. Different pathogens (i.e. viruses, bacteria) or even different species trigger common as well as different parts of the immune system and this trigger drives common as well as unique patterns of gene expression. By measuring the gene expression in either particular cells that are involved in the immune responses [3] or in complex mixtures of cells (such as blood) researchers can (1) understand the immune mechanisms that the body employs to eliminate the infectious agents (2) find biomarkers specific for infection/disease that will aid diagnosis, prognosis or treatment of disease.

1.2 Gene expression

DNA (deoxyribonucleic acid) is a macromolecule shared across all the cells of an organism that holds information about structure, function, growth and development of the cells by coding for functional cellular components [4]. Gene expression is the link between the genotype and the phenotype as it uses the information stored in the DNA to produce functional products via transcription (ribonucleic acid (RNA) species) and translation (proteins) [5]. Even though DNA is the same across the cells of an organism, the particular genes that are expressed, their level of expression and their isoforms differ between cells and between time points, via gene regulation that defines the state of each cell.

DNA's structural unit is the nucleotide and each nucleotide is composed of a monosaccharide sugar, a phosphate group and a nitrogenous base (adenine (A), guanine (G), thymine (T), cytosine (C)). The sugar of the one nucleotide and the phosphate of the next are joined by covalent bonds forming a strand. The bases also bind to each other by hydrogen bonds bringing the two strands together and forming a double helix. The specific order of the nitrogenous bases (gene) holds the information of the genetic code for the creation of the functional products. This encoded information is first transferred to an RNA molecule, via transcription. In eukaryotes, the initiation of transcription requires the presence of specific proteins, namely transcription factors, which are able to activate or suppress the transcription of a gene. Transcription factors can either recruit RNA polymerase or block its binding to the promoter area of the gene of interest, to start or

cease the transcription process, respectively. RNA polymerase is the enzyme required for the synthesis and production of the primary RNA molecule according to the complimentary base pairing rule. If the gene that undergoes transcription encodes for a protein, the product of transcription is messenger RNA (mRNA) which after processing and editing (in eukaryotic cells) will convey the genetic information to the ribosome to produce the protein required by the cell (Figure 2). DNA can also be transcribed to different types of RNA, such as transfer RNA (tRNA) or ribosomal RNA (rRNA), that have other various functional roles.

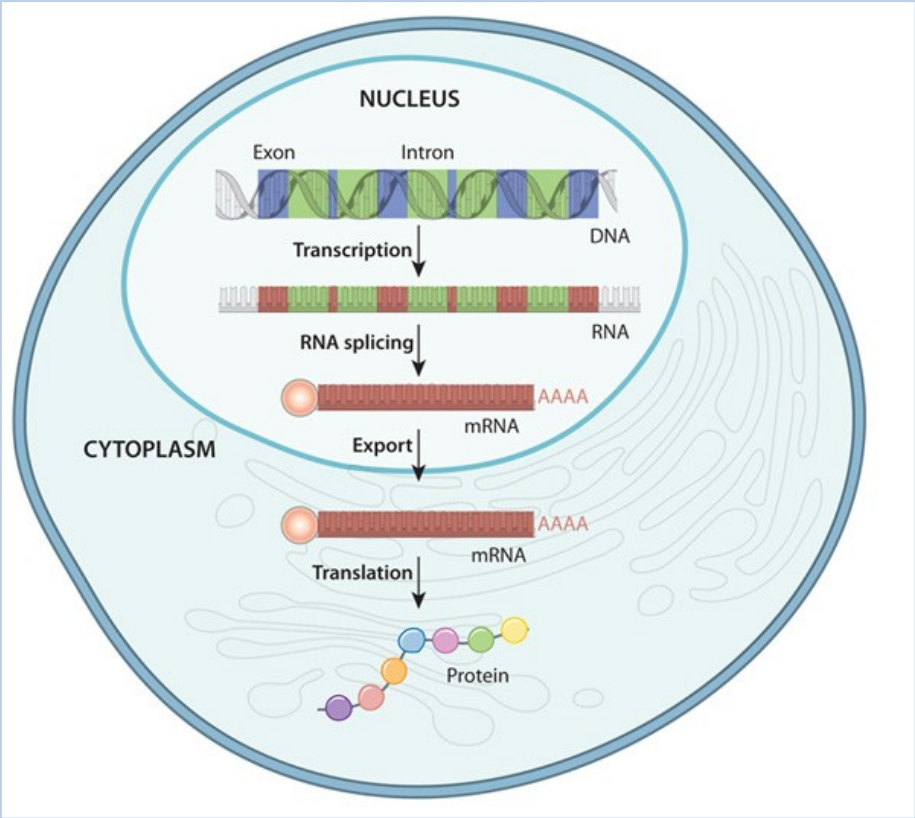


Figure 2: An overview of the flow of information from DNA to protein in a eukaryote. Reprinted with permission from Nature Education [6].

Only a fraction of approximately 30,000 genes in human DNA is expressed in a given cell at a given time, defining the cell's state. Cells employ a variety of mechanisms to define which genes are transcribed into RNA and which mRNAs are translated into proteins. In eukaryotic cells and especially in human cells, the control of gene expression is a very complicated process controlled by the regulome which comprises regulatory proteins, transcription factors, mRNA and metabolites [7]. The expression levels of a specific gene can

be quantified by detecting the presence and measuring the abundance of the final product (protein or functional RNA species) or its precursor (typically measure mRNA for proteins). Measuring the amount of mRNA can act as a "proxy" for the overall cellular activity at the molecular level as protein measurements can be more costly and laborious than mRNA measurements. However, studies in yeast as well as in mice showed that can be discrepancy between the abundance of RNA and proteins [8-10]. Even though protein measurements encompass post-transcriptional modifications as well as the action of miRNAs and are closer to the phenotype, mRNA analysis is a powerful method for monitoring the gene expression of a cell.

1.3 Measuring Gene Expression

1.3.1 Overview

Out of the amount of total RNA in a cell, which is usually 10-30 pg, 80% is rRNA, 15% is tRNA and the remaining 5% is mRNA along with all the other species [11]. The first step in all methods for measuring mRNA abundance is the extraction of RNA and isolation of mRNA from the cells, using specific extraction and isolation kits in order to purify it from the rest of the RNA. There are various methods to estimate the abundance of mRNA, or complimentary DNA (cDNA) synthesized using mRNA as a template, each of which has its own advantages and disadvantages. These methods separate into two broad categories: the candidate gene approaches (i.e. Northern Blotting, serial analysis of gene expression (SAGE), comparative expressed sequence tag (EST), reverse transcriptase followed by polymerase chain reaction (RT-PCR)) and the genome-wide profiling methods (i.e. DNA microarrays, RNA-seq) [12]. The candidate gene approaches focus on measuring expression of a small number of genes and are usually used when the genes of interest are known prior to the experiment for hypothesis testing. The genome-wide profiling methods allow for systematic analysis of thousands of genes simultaneously and provide a global quantitative profile of the cell or tissue of interest. They are more adequate for hypothesis free experiments. Most earlier targeted gene approaches (i.e. Northern Blotting, SAGE, EST) are nowadays superseded by RT-PCR, while the most popular methods for transcriptional profiling are DNA microarrays and RNA-seq, which are discussed in the following section.

1.3.2 RT-qPCR

RT-qPCR is a simple, powerful, widely used method for quantification of mRNA levels that consists of two steps: reverse transcription (RT) and real-time quantitative PCR [13, 14]. Reverse transcription as a methodological step is shared between most protocols for RNA quantification. During this step, the mRNA molecule is used as a template to synthesise a cDNA molecule in a reaction catalysed by the enzymes reverse transcriptase and DNA polymerase. cDNA is much more biologically stable than its precursor and is the starting material for PCR amplification. The PCR part of the RT-qPCR methodology consists of three basic discreet stages (denaturation, annealing and elongation). The three different stages make up a cycle and at the end of each cycle fluorescent dyes emit fluorescent signal that is used for the estimation of the quantity of the starting material [15]. In general, RT-qPCR is a sensitive, accurate, highly reproducible method that can detect very small amounts of RNA. On the other hand, it is still expensive, prone to contamination and not as high-throughput as newer methods (i.e. microarrays and RNA-seq). Nevertheless as the field of gene expression analysis has recently started moving from the candidate gene approaches to profiling methods, RT-qPCR is still considered as benchmark technology and used as the gold standard in many laboratories in the world.

1.3.3 Microarrays

Microarray technology exploits the principles of specific hybridisation between two DNA strands and the emission of fluorescence proportional to the amount of nucleic acid that is bound. Amongst other applications (i.e. detection of mutations, single nucleotide polymorphisms, comparative genomic hybridisation and identification of genomes), DNA microarrays (also referred to as cDNA microarrays / microarray chips) are a rapid, reliable and reproducible way to achieve gene expression profiling. They allow for the simultaneous genome-wide measurement and analysis of gene expression in a high-throughput manner.

While array technology in general originates in the 1950s, when the attachment of biomolecules to solid surfaces was first used to generate immunoassays for antibody binding, further advancements managed to accommodate nucleic acids in a similar way to the Southern Blotting technique in the 1980s [16, 17]. Those advancements led to the first

microarray fabrication in Stanford University for 45 *Arabidopsis* genes in 1995 [18]. Since then microarrays have started revolutionising the field of life sciences. However, it was the completion of the Human Genome Project in 2001 which released an unprecedented amount of genomic information that allowed for the assembly of a cDNA library of the transcripts of the human organism [19, 20].

There are two major types of arrays based on the type of the nucleic acid that is bound to the surface, the cDNA microarrays and the oligonucleotide microarrays. In the cDNA microarrays, the clones are 500-5000 bp long and products of PCR, while the oligonucleotides are 50-70 bp long. Also, oligonucleotides can be synthesised either "*a priori*" or "*in situ*", determining further the type of the array. In this thesis, I am going to focus on spotted oligonucleotide microarrays [21].

Thousands of distinct strands of DNA oligonucleotides, that correspond to segments of every known or putative transcript of the genome are immobilised onto a solid surface (i.e. a glass slide, silicon) in discrete spots with high density (~500 spots/cm^2). Each array contains over 30,000 spots, typically around 100-300 μm in size and each spot contains multiple identical oligonucleotides [22]. The spots are organised in a grid of rows and columns with known location. Due to the micro-scale of the experiment, the fabrication of the array involves robotic machinery of high precision and speed, which prints the oligonucleotide probes on the surface exploiting material science principles. The arrayed probes are complementary to the transcripts of interest and suitable for hybridisation.

The measurement of the gene expression involves RNA extraction from the sample of interest, mRNA isolation and reverse transcription to cDNA. The cDNA is then amplified and labelled with a fluorescent dye. Subsequently, the dyed cDNA is hybridised to the array in specific and controlled conditions, which permit its binding to particular complementary sequences. Then the array is washed in order to eliminate non-specific binding events, followed by insertion into a scanner that both excites the fluorescent dyes and records the emitted intensity. The more abundant the cDNA for a specific probe, the stronger the fluorescence signal it emits. The signals detected by the scanner are processed using imaging algorithms which assign intensity measurements to each probe. The measurements are then transformed into numerical values that can be related to the abundance of the

corresponding RNA transcript in the sample. Prior knowledge of the position of reference transcripts allows for relative quantification of the amount of transcripts that is bound to the specific probe [23, 24]. The output is a highly dimensional dataset that requires bioinformatics analysis to process and understand. The methodology and statistics involved are discussed in Chapter 2.

What is described above is a single-colour array, where the fluorescent dye has one colour. However, as the microarrays are often used for comparative analysis between different cell types, diseases or populations, mRNA from the two different conditions can be made into cDNA with fluorescent dyes of different wavelength emission, usually red (670 nm) and green (570 nm). This process results into probes emitting shades of red, green and yellow fluorescence relative to the amount of starting fluorescent cDNA in each sample. While two-colour arrays can control for technical issues by allowing for direct comparisons in one hybridisation experiment, single-colour arrays allow for more flexibility in analysis [25].

Microarray technology enables transcriptional profiling of large cohort sizes in a high-throughput way. It utilises cutting-edge advances in material science, automation, imaging and data science to quantify gene expression of multiple cells, tissues or organisms in a single experiment. Also, as a technique it is now standardised and highly-commercialised. However, biases in the experimental process due to reverse transcription efficiency and dye affinity are often introduced. Also it is still relatively expensive and requires more complicated post-experimental bioinformatics data analysis compared to RT-qPCR, which will be discussed in detail in Chapter 2.

1.3.3 RNA-seq

Various sequence based approaches have been used to measure gene expression since the 70s, however, it was only recently that next generation sequencing (NGS), which allows for vast parallelisation of deciphering sequences was introduced [26]. RNA-seq is based on the principles of DNA sequencing and allows for mapping and quantifying RNAs [27]. Although different sequencing platforms exist, they all share the same core principles. Following RNA extraction, rRNA depletion and synthesis of a library of short cDNA

fragments, the cDNA fragments are amplified and attached *in situ* on a solid surface. NGS technology is based on detecting and recording light that is emitted when a complimentary nucleotide is added to a particular fragment of cDNA. The light detected will determine the identity of the nucleotide ("base calling") and subsequently the sequence of the whole "read" in single base-resolution. Millions of short reads are either mapped to a reference genome or assembled *de novo* to produce the transcriptome, a base-resolution expression profile [28]. The abundance of each read, along with its alignment coordinates, allows for the quantification of specific genomic regions, in an almost digital way, as the output of RNA-seq produces discreet counts of transcript abundance. RNA-seq methodology has revolutionised the field of transcriptomics allowing for discovery of novel transcripts, as well as splice isoforms. RNA-seq also permits for simultaneous sequencing of pools of transcripts that may come from different organisms that coexist in the same environment, termed as metatranscriptomics [29, 30].

1.3.5 Value and importance

All methods used for transcriptome quantification have their own advantages and disadvantages. Depending on the nature of the experiment and the actual scientific questions posed, one method may be preferable over the other.

RT-qPCR is the method of preference when a relatively small already identified number of transcripts are to be studied. It is still considered the "gold standard" for RNA quantification, making it preferable for validation studies. On the other hand, microarrays and RNA-seq have made whole transcriptome analysis available, with arrays having limited hypothesis space as they examine only the transcripts that are printed *a priori* on a chip. The opportunities for novel transcript discovery and splicing isoform detection along with its higher dynamic range are the main biological reasons why RNA-seq has started superseding microarrays. Despite its superiority as a method, RNA-seq has not replaced arrays yet, as they remain time and cost efficient particularly when it comes to large-cohort studies. On the other hand, RNA-seq studies require ample storage space, high level data management, as well as powerful computational infrastructure [31]. Comparative analyses between RNA-seq and microarray techniques have demonstrated that although a larger proportion of

genes identified as differentially expressed by RNA-seq than microarrays were subsequently validated by RT-PCR, the two methods complement each other in transcriptome profiling [32, 33].

Soon after NGS was established, the third generation sequencing (TGS) methods emerged (also known as single molecule sequencing methods and next-next generation sequencing). These methods are expected to revolutionise the life science field in an unprecedented way as they are expected to reduce the sequencing error rate, the time to results from days to hours, and the overall cost per run [34, 35]. As far as the field of transcriptomics is concerned the impact is anticipated to be particularly high since TGS allow for direct sequencing of RNA molecules whilst omitting the cDNA synthesis and amplification steps.

The aforementioned methods measure gene expression, which is crucial for many normal biological cells processes including cells' response to external stimuli and encountering with pathogens. Identifying differentially expressed genes (DEG) between cells, tissues, diseases, disease states and treatments can shed light into biological processes and detect the key molecules that play crucial roles in the discrimination between the different conditions [36]. Even though sequencing methods confer a natural advantage and offer a wealth of information and possibility for deeper exploration, gene expression microarrays have been addressing the identification of differentially expressed genes successfully for many years. They have enabled the elucidation of a patient's response to external stimuli and further our understanding of the molecular regulatory mechanisms that underlie disease.

1.4 Host gene expression disease biomarkers

The advent of high-throughput gene expression measuring technologies, such as DNA microarrays and RNA-seq, allow for genome-wide level analysis and large cohort sizes. This has enhanced gene expression biomarker discovery and subsequent application to the clinical diagnostics field. A biomarker is defined as "a characteristic that is objectively measured and evaluated as an indicator of a normal biological process, pathogenic process or pharmacologic response to a therapeutic intervention" [37]. Biomarkers provide evidence for classification of different diseases as well as disease subgroups.

1.4.1 Host gene expression biomarkers for infectious diseases

During its early days, gene expression analysis for biomarker discovery found adoption in cancer research [38, 39]. Several studies since then have focused on a wide range of aspects, including classification of tumour types, prediction of metastasis, drug pre-disposition, patient survival or clinical outcomes in general [40, 41]. Infectious diseases were soon to follow, despite the fact that infection by a pathogen and the interaction with the host adds certain levels of complexity into host gene expression profiling [42]. Exposure of an organism to a pathogen can trigger a consistent biological response, for instance in pathways of immune activation. Therefore, disease specific biomarkers that describe the host response can be useful for better understanding of well-known as well as emerging infectious diseases and their diagnosis.

Various models are employed to study response to infection, spanning from cell line *in-vitro* experiments and animal models to study of multi-level human responses [43]. *In-vitro* host response studies monitor cells after the exposure to the pathogen to unravel cell specific mechanisms underlying host response to the pathogen. Cells of preference can either be critical components of the immune system (i.e. NK cells, T cells) or pathogen specific target cell types and cell lines. Even though *in vitro* approaches may not be able to fully describe the transcriptional response of the host, they offer a controlled environment and allow for examining changes in expression over time [44-47]. Furthermore, studies have also focused on analysing the transcriptome of human tissue samples from the site of the primary pathogen infection [48].

The aforementioned approaches can be of high value when the interest is biology and identification of disease stages. However in order to study clinical disease and identify biomarkers that can be of clinical significance, identification of patterns of gene expression in easily accessible bodily fluids such as nasopharyngeal secretions as well as peripheral whole blood from human patient samples is crucial. As distinct patterns of gene expression have been associated with infectious diseases and disease stages, these patterns allow distinction between patients affected by a disease from healthy controls or between patients with different diseases [49, 50].

1.4.2 Host gene expression biomarkers in blood

Diagnostics for infectious diseases have recently started to shift from identification of the causative agents or measuring a specific antibody-mediated response to recognition of host response gene expression patterns specific enough to each pathogen to allow for accurate disease diagnosis [51]. As the pathogen may be present in undetectable numbers or lying in inaccessible sites and as the antibody response is not measurable in the first stages of the disease, identification of gene expression patterns in easily accessible samples - such as blood - is expected to be beneficial in the field of diagnostics.

Blood is not only an accessible tissue which permits investigation of candidate biomarkers, but it has the potential to convey information about many different organs and tissues. Blood cells interact with all tissues of the body and play a key role in transportation of oxygen, nutrients and waste, as well as in immunity, inflammation, signalling and defence. Molecular profiling of circulating blood cells reflects physiological and pathological events occurring in various different tissues of the body as blood connects the systems of the body together and interacts with every organ and tissue. Thus, whole blood expression profiling is not only a means for exploring multiple physiological processes, but also a means for identification of expression patterns that offer a broad picture of the organism's health state and overall immunity. Microarray studies exploring the variation in whole blood gene expression patterns of healthy individuals have shown that differences in gene expression between healthy individuals are far less than the variation between healthy individuals and patients with cancers or bacterial infections [52]. Therefore, whole blood expression

profiling is not only a means for exploring multiple physiological processes, but also for identification of distinct disease signatures, even for diseases that affect different organs and tissues of the body. Peripheral blood cells share more than 80% of the transcriptome with brain, colon, heart, kidney, liver, lung, prostate, spleen and stomach [53]. Hence, it has been feasible to derive distinct host response signatures for a variety of diseases from transcriptional profiling of peripheral blood. Gene expression signatures have been reported for several diseases including bacterial and viral infections [54, 55] as well as pathogen specific diseases including malaria [56], dengue virus infection [57], staphylococcus [58], tularemia [59] and TB [60].

However, the use of whole-blood gene expression patterns to derive diagnostic biomarkers capable of distinguishing phenotypically similar diseases remains challenging. Differential gene expression (DGE) analysis is used to identify the key genes that undergo changes in expression relative to healthy individuals in each disease state as well as to identify changes between comparator groups of different diseases. These key genes can act as diagnostic, prognostic and predictive markers of disease. Once the gene expression "signatures" in the blood are identified, they can potentially be used for diagnosis of infectious diseases, where current diagnostics are unreliable or of limited potential.

1.5 Tuberculosis

Tuberculosis (TB) is among the most important causes of mortality and morbidity from infection in the human population. In 2012, it was estimated that 8.6 million people developed TB and 1.3 million died, with Africa, South-East Asia and Western Pacific regions accounting for 75% of the TB cases worldwide, according to the World Health Organisation (WHO) [61].

The causative agent of TB is *Mycobacterium tuberculosis* (M.TB), an intracellular acid fast bacterium first described in 1882 by Robert Koch, also known as "Koch's bacillus". The lungs are the major infection site for *M. tuberculosis* (pulmonary TB) but it can also affect other parts of the body such as the central nervous system, the bones, the lymph nodes and the heart (extra-pulmonary TB). TB is an air-borne disease as *M. tuberculosis* is carried in infectious droplets generated by adult patients, who have pulmonary or laryngeal TB

disease. Infection by the pathogen can either be eliminated by the host (rare to absent event) (i), progress to active disease immediately (ii) or controlled as a latent infection (LTBI). Approximately 90% of immunocompetent adults infected with *M. tuberculosis* will contain the disease in a dormant form and will remain persistently infected with a 10% lifetime chance of progressing to active disease [62, 63]. However, the probability of developing active TB rises with age and is much higher among people infected with human immunodeficiency virus (HIV) and in children (Figure 3).

After inhalation of *M. tuberculosis*, alveolar macrophages and dendritic cells engulf bacteria or their components (Figure 3). Dendritic cells circulate to the draining lymph nodes and prime T cells, which then travel to the lungs where they recruit and activate macrophages in order to contain the initial phase of bacterial multiplication. Macrophages play a crucial role in the infection as they phagocytise *M. tuberculosis*, which either multiply within the infected macrophages, or are contained and/or killed, depending on the state of the host immunity. T cells enhance the antibacterial activity of macrophages by releasing cytokines, such as interferon gamma (IFN-γ) and tumour necrosis factor alpha (TNF-α), which generally results in arrest or clearance of the infection [64].

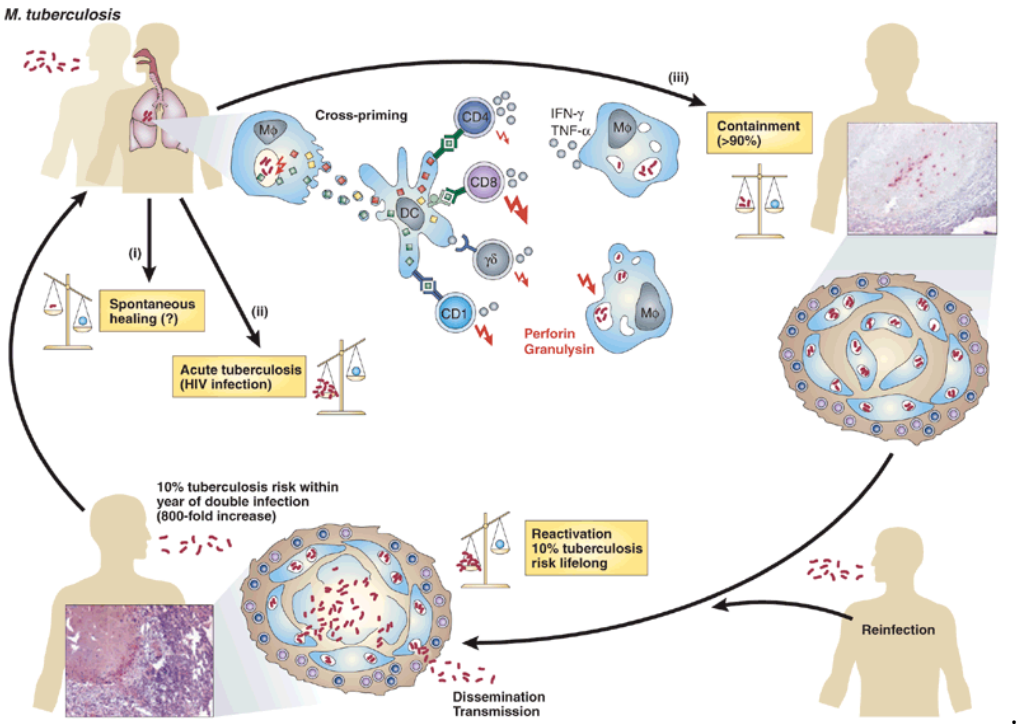


Figure 3: Different outcomes of *M. tuberculosis* infection and underlying immune mechanisms. Reprinted with permission from Nature Medicine [65].

Antigen-specific T cells induce the formation of granulomas around infected macrophages to contain bacterial spread [65]. Formation of granulomas prevents mycobacteria growth and spread in the body, but does not eliminate the possibility of subsequent development of TB in the future.

Co-infection of TB and HIV increases disease progression, and complicates diagnosis and treatment resulting in a dual-epidemic [66]. The lifetime risk of progressing from latent to active disease rises from 10% lifetime chance to 10% annually for HIV infected individuals [67]. Amongst the 8.6 million TB cases worldwide in 2012, 1.1 million (13%) are estimated to be co-infected with HIV but in some countries with high HIV prevalence, such as southern Africa, the proportion of patients with dual-infection rises to 80% (Figure 4) [61]. An estimated 25% of the 1.3 million deaths from the disease in 2012, approximately 320,000 deaths, were due to HIV-associated TB.

Children account for 500,000 to 1,000,000 of the new TB cases per year, while estimated deaths from TB amongst the HIV-negative children only were 74,000. The global estimated burden varies markedly, while the true global burden of childhood TB is unknown due to difficulties in diagnosis and underreporting [63, 68]. Among children infected with *M. tuberculosis*, a higher proportion in comparison to adults will progress to active disease, with the majority of them developing disease within a year after infection. Young age, severe malnutrition and HIV co-infection are the most important risk factors for developing disease. Worldwide, diagnosis and management of childhood TB is mostly based on clinical features as the available tests are unreliable in children due to the paucibacillary nature of the disease and high frequency of extra-pulmonary TB presentation [69, 70]. Whereas sputum microscopy is the mainstay of adult pulmonary diagnosis and control, sputum is difficult to obtain from young children, and is frequently negative on both microscopy and culture due to the low numbers of bacteria, absence of cavitatory lesions and frequent lympho-haematogenous spread without major pulmonary involvement.

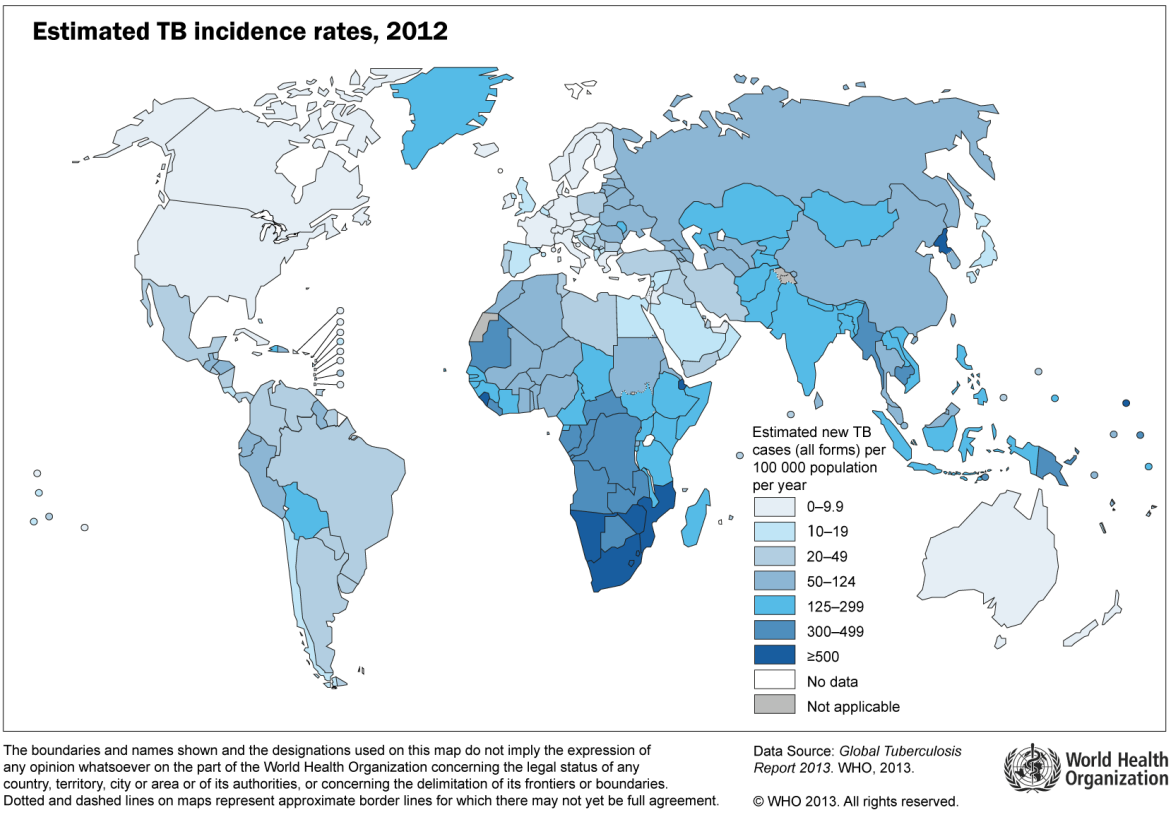


Figure 4: Worldwide estimated TB incidence rates from WHO in 2012.

The introduction of the TB vaccine, the bacillus Calmette-Guérin (BCG), in the 1920s marked a milestone in the fight against TB. The vaccine consists of an attenuated strain of *Mycobacterium bovis* and has been part of the national childhood immunisation programme for neonates and infants in numerous countries [71]. BCG vaccine has a documented protective effect against TB meningitis and disseminated TB disease in children but unfortunately, it has only moderate efficacy in preventing disease and decreasing mortality in children, and it is not able to control chronic infection or to protect from recurrent episodes or re-activation of LTBI in adults particularly if co-infected with HIV [72].

Although the incidence of new TB cases has been falling worldwide, the rate of decline remains slow (2%) [61]. TB is considered a curable disease requiring therapy with antibiotics, and most disease remains readily treated with available drugs, apart from multi-drug resistant TB (MDR-TB) cases that account for an estimated 3.2% of the new cases annually. Nevertheless controlling the global TB epidemic has been impeded by the absence

of an effective vaccine and the lack of rapid and sensitive diagnostics for both adults and children.

1.5.1 Diagnosis of Tuberculosis

Active TB disease is often diagnosed clinically, based on symptoms and clinical examination. Symptoms of the disease include fever, weight loss, night sweats and general malaise. Pulmonary tuberculosis usually presents with cough, lung inflammation and abnormalities often detectable on chest X-rays [73]. However, these symptoms and physical findings are non-specific. The clinical features of the disease are also affected by various factors including age, severity of disease, malnutrition, underlying conditions and genetic factors [74].

The gold standard laboratory test for diagnosing pulmonary TB as well as extrapulmonary disease is the isolation of *M. tuberculosis* in culture. Spontaneous or induced sputum samples -if produced- or other specimens are cultured for mycobacteria using specific growth medium. Due to the slow growth of *M. tuberculosis,* detection of mycobacterial growth usually takes weeks. Sputum smear microscopy to detect acid-fast bacilli (AFB) following a staining procedure and an acid wash is a quick method, and available even in low resource countries but it is not specific for the species of *M. tuberculosis.* AFB testing requires the presence of 5,000 to 10,000 bacteria per millilitre of specimen to allow for detection of bacteria under the microscope, while a positive culture requires 10 to 100 bacteria.

A number of *in-vitro* and *in-vivo* immunodiagnostic tests can provide indirect evidence of current or past exposure to *M.* tuberculosis. The Tuberculin Skin Test (also known as Mantoux test or purified protein derivative (PPD) skin test) is a method that has been used since 1890, and involves a subcutaneous injection of a small amount of *M. tuberculosis* antigens (i.e. PPD). The extent of the subsequent reaction to the antigens, measured in millimetres of induration, if present, is an indicator of TB exposure. However, the TST cannot discriminate between active and latent TB infection; has poor specificity upon prior BCG immunisation and exposure to environmental mycobacteria; can evoke

false-positive results and needs a second visit to the clinic 48 - 72 hours later for reaction reading [75]. These limitations were overcome by the introduction of Interferon γ–release assays (IGRAs). The IGRAs are based on the fact that sensitised host cells, when stimulated *in vitro* with *M. tuberculosis* specific antigens, will release the cytokine IFN-γ. IGRAs are more specific than TST and do not have cross-reactivity due to prior BCG vaccination or non-tuberculous mycobacterial infections [76, 77]. Currently commercially available tests employ similar enzyme-linked immunosorbent assays (ELISAs) to detect production of IFN-γ by circulating T cells in response to *M. tuberculosis* specific antigens (i.e. Early Secretory Antigen Target 6 (ESAT-6), Culture Filtrate Protein 10 (CFP-10) and TB7.7). Assays may be undertaken in either whole blood, or using separated peripheral blood mononuclear cells (PBMCs) that produce IFN-γ in response to ESAT-6 and CFP-10 [78, 79].

Sequencing of the *M. tuberculosis* genome [80] allowed for remarkable progress in the field of TB diagnostics by developing nucleic acid amplification tests (NAAT). Research in the field of molecular diagnosis for TB using various PCR-based assays has led to commercially available kits for both TB diagnosis and the identification of drug resistance-associated gene mutations. In 2010, WHO recommended the use of the Xpert® MTB/RIF assay, a quick, automated, PCR-based assay, which has the capacity to detect *M. tuberculosis* in the samples as well as its resistance to rifampicin [81]. Even for the simplest methods though, some experience in PCR methods and availability of laboratory facilities are required.

It is suggested that lab based TB diagnostic tools should be interpreted alongside with chest radiography, clinical and laboratory examinations to lead to accurate diagnosis and adequate treatment. Currently less than half of all cases of active TB are detected and treated, particularly in low resource countries that account for 95% of all TB cases and 98% of all TB deaths [61]. This proportion is even higher for children and HIV infected patients. In these groups, current technologies to detect TB are ineffective. Smear microscopy, chest radiography and TST are unreliable in children, extra-pulmonary cases and BCG vaccinated individuals respectively, whilst culture of *M. tuberculosis* is currently considered as the gold standard [20]. Sensitivity of the NAAT tests depends on the DNA abundance in the sample. Although molecular diagnosis has improved detection of *M. tuberculosis* DNA in sputum, the sensitivity of this approach is lower in sputum negative, culture positive samples [82], and

the method does not detect solely extra-pulmonary disease. IGRAs are not able to discriminate TB from LTBI [83] and are of limited utility in African countries where LTBI is highly prevalent in the healthy population [84, 85].

As a simple, inexpensive point of care test for TB is still not available, there is an urgent need for improved diagnostic methods, particularly in African countries where TB and HIV persists as a dual epidemic as well as in paediatric populations. Transcriptional approaches using blood signatures have mostly been used to gain insight into the host response to mycobacteria infection and disease, but can also assist to the discovery of diagnostic biomarkers.

1.5.2 Host Gene Expression in Tuberculosis

The earliest studies using blood transcriptional profiling by microarray analysis in TB patients aimed at establishing the feasibility of identification of discriminatory biomarkers for a pulmonary disease using PBMCs or whole blood. The initial focus was on distinguishing groups of active TB patients, patients with recurrent TB, post-treatment TB patients and *M. tuberculosis* latently infected individuals using gene expression [86, 87]. These studies described the first blood transcriptional signatures of TB, detected even from small sample sizes, using gene expression microarray technology. Subsequent studies identified biomarkers which discriminated between latent TB infected individuals, active TB patients, BCG vaccinated individuals as well as healthy controls [88-91]. No differences in gene expression have been reported between the LTBI and healthy control groups *in vivo*, unless the samples were stimulated with PPD *in vitro* prior to microarray analysis [88]. However, signatures correlating with extent and progression of the disease, as well as disease treatment, have been identified [92-94]. Although these signatures would be very useful in discriminating active and latent TB infection, they would be of limited use for diagnostic purposes in the clinic, as the early studies did not compare TB with the spectrum of other diseases with similar presenting symptoms.

In order to identify a TB specific signature, a study from the UK and South Africa compared transcriptional profiles from adult active TB patients with those from patients with other bacterial and inflammatory diseases namely Still's disease, Staphylococcus, group

A Streptococcus and adult & paediatric systemic lupus erythematosus) [60]. A signature of 86 transcripts dominated by over-representation of interferon-inducible transcripts, able to discriminate between TB patients and patient with other diseases was reported. However, further studies have observed a significant overlap of this signature with whole blood transcriptional signatures for sarcoidosis and melioidosis diseases, highlighting the need for appropriate control groups and highly specific biomarker sets [95-97]. A different set of studies have used hypothesis driven approaches based on previous microarray experiments or biological pathway data and tested the expression of candidate biomarkers in the blood of TB infected individuals [98-100].

Most notably, HIV co-infected patients have been excluded from all the above studies. Given the disruptive effect that HIV infection has on the immune system [101] as well as the impediments it adds to TB diagnosis and outcome of the disease, HIV patients should be considered as a priority group to include in TB biomarker discovery studies. A recent study compared the transcriptional profiles of PBMCs of HIV mono-infected patients to HIV/TB co-infected patients, and identified a 251-transcript signature able to discriminate between the two groups. However, the study excluded any HIV infected patients with other underlying diseases or co-infections [102]. In terms of signature identification, these studies employ various statistical methods for model fitting, signature identification and patients' classification (i.e. t-test, linear discriminant analysis, support vector machines, decision trees, k-nearest neighbor, and random forests). The machine learning methodology is discussed extensively in the following chapter. The published literature on host RNA expression regarding TB disease and infection available prior to publication of the studies described in this thesis is summarised in Table 1 below.

Due to the different nature of the disease between adults and children and the particular challenges involved in diagnosis of paediatric TB, gene expression biomarkers in children should be sought for separately. Only one recent study has compared the whole blood transcriptional profiling of HIV uninfected native American children with active TB, latent TB infection and healthy controls to identify a discriminatory transcriptional signature [103]. However, the study excluded HIV infected children as well as children with other diseases and infections apart from latent TB infection.

Previous studies in TB have suggested that blood transcriptional signatures can be used to distinguish TB from healthy controls, latent TB infection, as well as other conditions [104]. The major challenge that remains is how to distinguish TB disease from the range of other conditions that show similar symptoms to it and represent the “real world” disease spectra from which TB must be distinguished in the clinic. This challenge is amplified for biomarkers that might be used in settings where TB and HIV are co-endemic, as well as in paediatric populations.

Study	Region	Year	Sample	Groups	RT-PCR Validation	HIV+	OD	OD/HIV+	Paediatric
Jacobsen *et al.* [86]	Germany	2007	PBMC	TB LTBI	+	-	-	-	-
Mistry *et al.* [87]	SA	2007	whole blood	TB LTBI	+	-	-	-	-
Berry *et al.* [60]	UK & SA	2010	whole blood	TB LTBI HC OD & TB treatment	-	-	+	-	-
Lesho *et al.* [89]	USA & Brazil	2011	whole blood	TB LTBI HC (TST+)	-	-	-	-	-
Lu *et al.* [88]	China	2011	PBMC (PPD)	TB LTBI HC	-	-	-	-	-
Maertzdorf *et al.* [90]	Gambia	2011	whole blood	TB LTBI HC	-	-	-	-	-
Maertzdorf *et al.* [91]	SA	2011	whole blood	TB LTBI HC	-	-	-	-	-
Bloom *et al.*[92]	SA	2012	whole blood	TB treatment	-	-	-	-	-
Maertzdorf *et al.* [96]	Germany	2012	whole blood	TB OD	-	-	+	-	-
Ottenhof *et al.*[93]	Indonesia	2012	PBMC	TB HC & TB treatment	+	-	-	-	-
Bloom *et al.* [97]	UK	2013	whole blood	TB & TB treatment	-	-	+	-	-
Cliff *et al.*[94]	SA	2013	whole blood	TB treatment	-	-	-	-	-
Dawany *et al.* [102]	SA	2014	PBMC	TB LTBI HC & TB treatment	-	limited	-	-	-
Verhagen et al.[103]	Venezuela	2014	whole blood	TB LTBI HC	+	-	-	-	+

Table 1: Summary of the published literature on RNA expression TB studies available prior to publication of the studies described in this thesis in chronological order, indicating the location of cohort, the year, the sample, the patient groups and the RNA quantification method.
TB: tuberculosis, LTBI: latent tuberculosis, OD: Other Diseases, HC: healthy controls, HIV+/-: HIV infected/uninfected

1.6 Pitfalls of employing gene expression for biomarker discovery

Although measuring transcript abundance in patient cohorts has been used to identify genes involved in disease pathways, as well as genes that can act as biomarkers for disease, there are limitations that may undermine the power of gene expression profiling if not appropriately addressed [105].

Firstly, the correct patient groups and relevant tissue samples need to be identified and collected to enable a robust approach to the scientific question posed. As the majority of gene expression studies are analysed in a comparative way, relevant age-matched controls (other diseases or healthy) need to be recruited as well. Before patient recruitment starts, accurate and detailed clinical definitions of the groups need to be decided. The minimum number of samples required, so that an effect of a certain size is likely to be detected, should be calculated (power analysis). Factors such as age, gender and race have a strong impact on gene expression and need to be considered, both during the early stages of setting up a study as well as during the data analysis stage. As for disease gene expression studies, heterogeneity within disease groups may have a strong influence on the results; time from onset of symptoms until the sample is taken, different disease subtypes, severity, and treatment protocols.

Sample-related bias can also be introduced during different experimental stages from collection and handling (RNA degradation, number of freeze-thaw cycles), to processing (RNA extraction) and analysis. Another factor that complicates gene expression studies is the large heterogeneity of the tissues. For example in whole blood samples, the observed gene expression is a weighted average of the gene expression in all the different cell types, reflecting the cell types' proportions.

Quality control at various steps of gene expression analysis should be carefully assessed. Additionally, alternative technologies and out-of-sample validation should be employed to assess the results of analysis and further in vivo and in vitro studies should be conducted to address the functional relevance of the findings.

1.7 Summary

This introductory chapter sets the background for the rest of the thesis and describes the biological basis and questions which are addressed in the remaining chapters. It depicts the importance of host gene expression as a marker of biological processes, and in response to particular pathogens or infections. Measuring gene expression can shed light on mechanisms underlying complicated biological processes as well as identifying transcriptional patterns specific for different diseases and infections. In particular whole blood can be used for transcriptional profiling of host response to infection and has been shown to be able to distinguish between different diseases and disease types.

The existing challenges in biomarker discovery using gene expression microarrays are then highlighted. Former studies in the field of whole blood transcriptional profiling for infectious diseases and in particular tuberculosis disease are also reviewed, presenting their advantages and limitations. Previous studies' findings support that recruitment of appropriate control groups is crucial for biomarker discovery and that results should be out-of-sample validated using independent datasets. In the next section we list the objectives of this thesis and how it aims to further the research in the field of identification of whole blood RNA biomarkers for Tuberculosis disease.

1.8 Objectives of the thesis

The aims of my thesis are:

1. To develop an analysis pipeline applicable to whole blood gene expression microarray data to derive diagnostic biomarkers capable of distinguishing phenotypically similar diseases.
2. To employ variable selection methods to identify the smallest sets of biomarkers that would allow best discrimination between comparator groups.
3. To develop a risk score for each individual to indicate class of disease and evaluate the predictive performance of the score.
4. To apply the diagnostic pipeline to cohorts of adult and paediatric patients with whole blood gene expression measurements, including TB patients as well as patients with other diseases that phenotypically mimic TB, while including HIV co-infected patients.
5. To assess the performance of the biomarker sets and validate their classificatory power in independent cohorts.
6. To compare the signatures with published signatures as well as current diagnostic methods employed in the clinical setting.

Chapter 2

Methodology

In Chapter 2, the statistical pipeline used to analyse gene expression microarrays and identify minimal gene expression signatures that can accurately classify patients with different diseases is described. High-dimensionality of microarray data, along with the particular challenges introduced by gene expression datasets derived from African settings was taken into consideration. The driving force behind this analysis was to develop a diagnostic test for TB that could be used in clinical practice. For that reason we employed parsimonious regression models of transcripts that are biologically interpretable. Subsequently, these models are translated into a single value disease score for each patient that reflects the risk of the patient having the disease.

2.1 Overview

In terms borrowed from machine learning, the scientific questions that gene expression microarray application aims to address fall in three categories: class discovery, class comparison and class prediction [106]. The classes represent groupings of samples with shared characteristics (for example treated or untreated patients, healthy versus disease or differentiation between subtypes of the same disease). In brief, class discovery aims at the identification of subgroups of patients that share common gene expression patterns in an unsupervised manner. Class comparison aims at the derivation of genes that are differentially expressed between conditions in a supervised manner, since the class membership is known. Lastly, class prediction seeks for multivariate predictors (i.e. set of genes) that would allow for the best classification of samples of unknown class membership. The breakthrough for the clinical application of microarrays from a machine learning perspective was the molecular classification of different types of leukaemia cancer employing both class discovery and class prediction [107]. In this thesis we employ techniques that fall in these three categories while analysing gene expression microarrays from patients that have TB disease and LTBI as well as other conditions.

2.2 Pre-processing

The microarray technology allows for simultaneous measuring of thousands of transcripts on a dense array in a high-throughput manner that generates high-dimensional datasets. These data require dedicated management and statistical analysis in order to produce biologically meaningful results. An additional level of complexity arises from the experimental process itself, which involves multiple steps that can be prone to error. From sample collection to RNA extraction to microarray hybridization and imaging, every step can introduce variability and bias and thus affect the end result.

Microarray analysis can serve as a tool for addressing various biological questions, which require distinct experimental design followed by specialised statistical approaches. However, the initial pre-processing steps such as quality control, background subtraction and normalisation are common across most experiments. In this thesis, we will focus on the

analysis of single-colour Human HT-12 V4 Expression BeadChip arrays (Illumina). As described in Chapter 1, after scanning the arrays, the image files are converted into a numeric table with measures of the intensity for each bead. This automated digitalisation procedure employs image processing algorithms that assign a signal intensity value to each pixel representing the fluorescence of the corresponding microarray bead [102]. By using spatial information for the array, a numerical value for every probe is recovered.

The raw dataset is initially processed using the Illumina Genome Studio software, which provides summary results to assess the quality of the experimental procedures, as well as the sample quality. This quality control step measures the fidelity of the hybridization of RNA to the array, by assessing non-specific binding and the hybridisation stringency and estimates the efficiency of Cy3 binding using biotin tagged probes. In addition, the customary inclusion of housekeeping genes in transcriptomic experiments allows for a consistency check. This is achieved by comparing the intensities of housekeeping genes to background intensity levels and examining the distribution of the average housekeeping genes intensity across the samples.

Following this first inspection, the three main steps of the pre-processing procedure are as follows: background adjustment, logarithmic transformation and normalisation. Background adjustment eliminates the effects of noise due to image scanning, non-specific hybridisation of some probes as well as auto-fluorescence of the chip surface. A specific function (forcePositive) adds an offset to all the values of the array, to address any negative values.

The data then undergoes a logarithmic transformation (base 2) in order to reflect more intuitively the amplitude of transcriptional change, both the induction and the repression of gene expression. Even though ratios are often used as a measure of expression change, they are not easily interpretable for under-expression, as this is described by numbers between 0 and 1. The logarithmic transformation outputs a continuous spectrum of differences of gene expression and treats the up- and down-regulated expression values in a similar way.

Normalisation is used to eliminate the systematic technical variation between arrays that is due to non-biological factors. As it has been shown that normalisation processes affect the outcome of the analysis, many microarray normalisation methods have been developed, having their own advantages and disadvantages. According to a systematic comparison, the most appropriate method for normalising Illumina Beadchip microarray data is Robust Spline Normalisation (RSN) [108, 109]. RSN combines the advantages of two other popular methods, the quantile and loess normalisation [110, 111]. The quantile normalisation method is a simple non-parametric normalisation procedure that has low computational burden and only uses the assumption that intensities of different arrays are derived from the same underlying distribution. It is a quick and efficient way of reducing technical variation, while maintaining the ranks of genes. Despite its advantages, quantile normalisation has a tendency towards over-correction and may thus reduce existent subtle biological differences between the samples. Loess normalisation, on the other hand, is based on fitting a smoothing curve to the microarray intensities, in a weighted way which penalises the contributions of data points that are far from the curve. It is based on principles of multiple localised linear regressions applied to a window of the data each time. So, the RSN methodology first employs quantile normalisation to estimate the fold change of the probes and then fits a monotonic spline curve on the data. This information is used to down-weight the contribution of probes that are highly differentially expressed presumably due to biological factors [109].

The final stage of the pre-processing is summarisation, which reflects the combining of multiple pre-processed probe level intensities that correspond to the same gene into gene abundance estimates. This step was skipped in our pipeline, as we analysed the data at the transcript level, rather than at the gene level, in order to avoid conflating values from transcripts with important biological differences (for instance reflecting alternative splicing events).

2.2 Microarray Analysis Toolbox

The output of pre-processing is an expression data matrix of logarithmically transformed values, which stands as the base of subsequent statistical analysis. The expression matrix is two-dimensional with rows that correspond to transcripts and columns that correspond to patient samples. A database file containing the phenotypic characteristics of the samples was included in our analysis pipeline.

2.2.1 Quality Control

Since microarray data are of high dimensionality, as part of the quality control process we employed multidimensional scaling. This strategy allows us to summarise and visualise the dataset, explore any remaining variance in the expression level even after normalisation and to detect outliers. Principal Component Analysis (PCA) is used to perform dimensionality reduction (no. of arrays x ~48,000 probes, down to no. arrays x no of principal components) [112]. Since the analysis toolbox is designed for gene expression data from heterogeneous disease groups, various geographic locations, and different underlying endemic infections with samples handled by different laboratories, we needed to detect any diagnostically irrelevant subclasses of the data. PCA transforms the data into a new set of variables, the principal components, which are linear combinations of the original variables. The number of the principal components can be less than or equal to the number of original variables in the orthogonal space. The first component accounts for as much of the variability in the data as possible and the succeeding components show the highest possible variability of the data, while being orthogonal to the previous components. A Hotteling T^2 test statistic, the 2-dimensional equivalent of the t-statistic, was calculated to measure the multivariate distance of each observation from the centre of the data set. The Hotelling T^2 distance is a measure that accounts for the covariance structure of a multivariate normal distribution. The higher the T^2 value, the more distant the observation is from the mean [113, 114].

2.2.2 Differential Expression

In this thesis, we set out to identify biological differences and discover biomarkers able to discriminate between two predetermined sample classes. For the rest of the thesis (unless otherwise stated) we assume that the two classes correspond to accurately phenotyped groups of patients. In order to compare the expression of the transcripts between classes, the average expression for each examined transcript is summarised by its mean value across samples in the class. Differential expression can be assessed by the log fold change (logFC) of the transcript expression values, i.e. the difference between the means of the log-transformed expression values. However, examining the logFC by itself can lead to false positive results. To address this concern, we also employed more rigorous statistical tests, such as the Student's t-test, which evaluates the equality of the means, while comparing the effect size with its variance among subjects and is widely used in microarray studies. The t-statistic is the difference between two means divided by the standard error of the difference between the two means. The null hypothesis of the two sample t-test is that the means of the expression values are equal. Thus, the t-test can be used to calculate p-values that define the probability of rejecting the equality hypothesis, based on a predefined level of significance (α).

However, within any microarray experiment there is a component of the transcript intensity variation that is shared amongst them. Consequently, information from other transcripts can be used to aid in estimation of the variance of a single transcript. There are many approaches developed that employ Bayesian approaches to leverage information across multiple transcripts. In our pipeline, we assessed differential expression between patient groups using LIMMA, a method that combines linear models, moderated t-test statistics and an empirical Bayes approach [115]. First, a linear model is fitted to each transcript aiming to model any systematic effect on the expression.

$$\begin{pmatrix} \boldsymbol{Y}^{(1)} \\ \vdots \\ \boldsymbol{Y}^{(j)} \end{pmatrix}_{j\,x\,n} = \begin{pmatrix} a^{(1)} \\ \vdots \\ a^{(j)} \end{pmatrix}_{j\,x\,p} \boldsymbol{X'}_{p\,x\,n} + \begin{pmatrix} \varepsilon^{(1)} \\ \vdots \\ \varepsilon^{(j)} \end{pmatrix}_{j\,x\,n} \tag{2.1}$$

The response matrix **Y** for every transcript (j) across the samples (n) is modelled. A design matrix **X** is used for every transcript corresponding to the comparison grouping. It establishes the relationship between observations and the parameters of the model. The regression coefficient matrix **α** represents the average difference in expression between the classes, with some of the coefficients corresponding to differences of biological interest, which are indicated by a contrast matrix **C**.

$$\boldsymbol{\beta}_{jp} = \boldsymbol{C}^T \boldsymbol{\alpha}_{jp} \tag{2.2}$$

A moderated t-statistic with augmented degrees of freedom (allowing for more than two contrasts) is calculated to test the null hypothesis $H_0 : \beta_{jp}=0$. This statistic employs an empirical Bayes approach where the parameters of the prior distribution are estimated from the data) permitting estimation and testing of the effects. An assumption underlying the use of this statistical model is that there is no interaction of one gene with another. Lastly, the moderated t-test statistics are translated into p-values as for the ordinary t-test.

In microarray expression analysis, where tens of thousands of comparisons are conducted simultaneously, the chance of incorrect rejection of the null hypothesis is elevated (type I error). However, this type of error can be controlled using methods that correct for multiple testing either by adjusting the level of significance or by reducing the proportion of false positives transcripts amongst all significant results. Bonferroni correction method modifies the significance cut-off (α/n), by dividing the level of significance (α) by the number of samples (n), but it is highly conservative, particularly for microarray data [116]. Another popular method is the Benjamini-Hochberg False Discovery Rate (FDR) method that reduces the false positive results and is less conservative [117]. It is based on assigning ranks (i) to the p-values of the transcripts, multiplying them by their number (n) and dividing them by their rank (i). In this way the adjusted p-values for transcripts with low ranking are deflated.

$$B-H\ Adjusted\ p-value_j = \frac{p-value_j \times n_j}{n_j + 1 - i}, i = 1, \dots, n \tag{2.3}$$

Long lists of transcripts that are significantly differentially expressed between conditions of interest are very informative, especially when the aim is biological exploration of the data. However, smaller probe lists may achieve better prediction between conditions, as noise and redundant information is removed. Also smaller numbers of biomarkers enhance the feasibility of diagnostic development based on gene expression. In the pipeline, I employ a class comparison step to filter out transcripts that are not adequate candidates for biomarker selection (i.e. low $\log_2$FC and high p-values). For each comparison of interest, only the transcripts with a $|\log_2 FC| > 0.5$ were taken forward to variable selection. The FC threshold was chosen in order to ensure that differential expression for selected variables is within the resolution limits of qRT-PCR, in case of subsequent analysis.

2.3 Feature Selection

There are various methods of implementing feature selection, but regression is an intuitive way to identify contribution of gene expression to a particular phenotype [118]. Regression is a statistical model that reveals underlying relationships between a dependent variable (disease state) and various independent variables (transcripts). The contributions of the independent variables are reflected by the regression coefficients. Regression is a popular method in systems biology as it constructs models that are simple and easy to interpret in a biological context. However, predictor variables are usually highly correlated, meaning that multiple variables may explain the same component of the independent variable. This holds particularly for gene expression data where many genes can be co-expressed as part of the same biological process.

As the aim is to identify the smallest possible transcript set containing key predictive variables that will have a good predictive performance, a dimension reduction strategy was employed to acquire a parsimonious regression model. There are two broad families of model selection in the context of linear models; subset selection methods and shrinkage methods.

Subset selection methods (i.e. forward and backward stepwise selection) have a series of limitations amongst which the inability to handle appropriately multicollinearity

between the variables and the "large p, small n". Data derived from microarray experiments are typical of the "large p, small n" problem, meaning that the number of features **(p)** is much larger than the number of observations **(n)**. Consequently, in order to address this class classification problem we resorted to shrinkage methods.

Shrinkage methods (also called regularisation methods) shrink the regression coefficients -potentially to zero to obtain sparse models- by imposing a penalty on their size, and are extensions of the least squares method. That leads to a selection of a small subset of variables that exhibits stronger effects on the independent variable than the rest of the variables, while maintaining prediction accuracy. It was decided to address this problem using elastic net [119], which is an intermediate approach between ridge regression [120] and lasso [121].

If we assume a linear model with n observations and p predictors:

$$\mathrm{y} = \beta_{\mathrm{o}} + \mathrm{x}\hat{\beta} + \varepsilon \tag{2.4}$$

where:

$\mathrm{y_i} = (\mathrm{y_1}, \dots, \mathrm{y_n})^{\mathrm{T}}$ is the response variable

β_{o} is the intercept term

$\mathrm{x_j} = \left(\mathrm{x_{1j}}, \dots, \mathrm{x_{np}}\right)^{\mathrm{T}}$ are the predictors

$\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_{\mathrm{p}})$ is the vector of the coefficients given by the model fitting procedures

$\varepsilon_{\mathrm{i}} = (\varepsilon_1, \dots, \varepsilon_{\mathrm{n}})^{\mathrm{T}}$ is the vector of the coefficients given by the model fitting procedures

The most popular method to estimate the coefficients of the model $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_{\mathrm{p}})$ is ordinary least squares (OLS). In this method the coefficients are determined by minimising a penalised residual sum of squares.

OLS:

$$\hat{\beta}^{least\ squares} = argmin\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\left(x_{ij}\beta_j\right)\right)^2\right\} \tag{2.5}$$

In shrinkage methods, a non-negative complexity parameter ($\boldsymbol{\lambda} \geq 0$) is introduced to control the amount of shrinkage and the coefficients are determined by minimising the penalised residual sum of squares (2.6). The parameter $\boldsymbol{\lambda}$ determines how important the penalty on coefficient weights is.

Ridge:

$$\hat{\beta}^{ridge} = argmin\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}(x_{ij}\beta_j)\right)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2\right\} \tag{2.6}$$

Although ridge regression shrinks the coefficients of the model, it does not produce sparse models as it always retains all the predictors. Lasso, on the other hand, uses a modified penalty that allows both continuous shrinkage and automatic variable selection simultaneously, as well as sparsity exploitation.

Lasso:

$$\hat{\beta}^{lasso} = argmin\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}(x_{ij}\beta_j)\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{2.7}$$

Elastic net is a compromise between ridge regression and lasso that selects variables like the lasso, but jointly shrinks the coefficients of correlated predictors like ridge regression.

Elastic Net:

$$\hat{\beta}^{enet} = argmin\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}(x_{ij}\beta_j)\right)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right\}, \tag{2.8}$$

The second term of the elastic net penalty encourages highly correlated variables to be averaged while the first one encourages a sparse solution in the coefficients of these averaged features. Compared to lasso, elastic net has the ability to result in more than **n** non-zero coefficients, where **n** is the sample size. This is especially advantageous for microarray experiments with small sample sizes, as it overcomes the **p>n** problem. Simulations as well as real data show that the elastic net often outperforms the lasso in predictive performance [122].

For the above reasons, we selected the elastic net as the shrinkage method used for the analysis. In our pipeline, the $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ parameters of elastic net, which indirectly control the size of the transcript signature as well, were selected via ten-fold cross-validation (CV) to avoid overfitting and promote generalisation. We varied both parameters over a two dimensional grid and acquire the full solution path for every combination. Values that produced the lowest mean squared error (MSE) determined the final parameterisation of the model.

2.4 Disease Risk Score implementation

The constructed elastic net models based on the microarray data include a specific set of transcripts along with their regressors, and they are used to perform class annotation of the patient samples. However, the analysis pipeline was created with the ultimate goal of developing a simple, affordable TB diagnostic test suited for use in resource poor settings. In order for the gene expression signatures information to be routinely utilised for clinical diagnosis even in low resource settings, a transformation is required to enable the potential development of a suitable for clinical use test that would not rely on sophisticated equipment.

Since a regression model does not hold the potential to be translated into a near patient assay test, we therefore developed a general method for converting complex weighted multiple transcript RNA signatures into a disease risk score (DRS). The simplification technique that we propose enables microarray signatures of disease to be

converted into a single value score for every individual. The single value classification could aid the development of a simple, low cost, diagnostic test requiring basic laboratory facilities and minimal bioinformatics analysis. The calculated DRS is able to classify the patients without significant loss of prediction accuracy and could form the basis of a diagnostic test (see Results chapter) [123].

Transcripts from the signatures were separated into up-regulated and down-regulated based on their in-sample calculated differential expression. The score is subsequently derived by adding the total intensity of the up-regulated transcripts and subtracting the total intensity of the down-regulated transcripts. The disease risk score for an individual *i* can be calculated using the formula:

$$Disease\ Risk\ Score^{i} = \sum_{k=0}^{n} expr.value_{k}^{i} - \sum_{l=0}^{m} expr.value_{l}^{i} \tag{2.12}$$

where:

n is the number of up-regulated transcripts in the signature in disease of interest (TB) compared to comparator group(s).

m is the number of down-regulated transcripts in the signature in disease of interest (TB) compared to comparator group(s).

The threshold for the classification was calculated as the weighted average of the risk score within each class (group of patients), with weights given as the inverse of the standard deviation of the score within each class. The threshold for the classification between group *u* and *v* is shown below:

$$threshold(u,v) = \frac{\frac{\mu_u}{\sigma_u} + \frac{\mu_v}{\sigma_v}}{\frac{1}{\sigma_u} + \frac{1}{\sigma_v}} \tag{2.13}$$

where:

μ is the average of the disease risk score in the group.

σ is the standard deviation of the disease risk score in the group.

We further explored the effect of adjusting the threshold of the disease risk score for assigning individual patients to groups. By accepting a percentage of patients as "not classified" by our test and thus requiring further investigation, the majority of patients under investigation are accurately assigned. This suggests that the score might be used as an initial screening "rule in/rule out" approach to TB diagnosis. We define the indeterminate zone, by calculating the lower and upper DRS thresholds as a weighted average given by $\frac{w}{\sigma_u}$, $\frac{2-w}{\sigma_v}$ respectively:

$$weighted\_threshold(u,v) = \frac{w * \frac{\mu_u}{\sigma_u} + (2-w) * \frac{\mu_v}{\sigma_v}}{\frac{w}{\sigma_u} + \frac{2-w}{\sigma_v}}, \quad 0 \leq w \leq 2 \qquad (2.14)$$

For w=1 the formula is equivalent to the main threshold formula (2.13).

The information that the DRS requires for classification (i.e. the expression values of the transcripts in an non-weighted manner) can be derived from the dataset itself, which allows its unbiased application using expression data acquired using other array platforms or non-array technologies. The classification on a single value holds diagnostic potential and could be used for the development of rapid diagnostic tests suitable for clinical use for the determination of the presence of an infection or disease in a host. It can readily be characterised as positive for a disease, disease state or infection.

2.5 Classification metrics

In order to evaluate the predictive accuracy of the classificatory models derived from our variable selection analysis as well as the performance of the DRS, we calculated different point and interval metrics typically used for diagnostic tests [124-127]. We compared the results obtained using the elastic net model and the DRS to the clinical diagnosis of the patients, which is considered the "gold standard" diagnosis. For this benchmark, we calculated area under the characteristic curve (AUC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and likelihood ratios. Confidence intervals at 95% were calculated to measure the reliability of our estimates ($CI_{95\%}$).

Group	Definition
True positive	gold standard "positive" individuals that the test classifies them as "positive"
False positive	gold standard "negative" individuals that the test classifies them as "positive"
True negative	gold standard "negative" individuals that the test classifies them as "negative"
False negative	gold standard "positive" individuals that the test classifies them as "negative"

Sensitivity (True Positive Rate): the proportion of true positive correctly identified by the classifier. It reflects the probability of being test positive with disease present.

$$sensitivity = \frac{true\ positive}{true\ positive + false\ negative} \tag{2.15}$$

Specificity (1 - False Positive Rate): the proportion of true negatives correctly identified by the classifier. It reflects the probability of being test negative with disease absent.

$$specificity = \frac{true\ negative}{true\ negative + false\ positive} \tag{2.16}$$

Positive predictive value (PPV): the proportion of the outcome positives that are true positives. It reflects the probability of a patient having the disease when the test is positive.

$$PPV = \frac{true\ positive}{true\ positive + false\ positive} \tag{2.17}$$

Negative predictive value (NPV): is the proportion of the outcome negatives that are true negatives. It reflects the probability of a patient not having the disease when the test is negative.

$$NPV = \frac{true\ negative}{true\ negative + false\ negative} \tag{2.18}$$

For case-control studies the ratio of gold standard positive and negative individuals does not reflect the real prevalence of the disease in a community or hospital setting, as in

observational studies. Given the dependency of NPV/PPV on the prevalence of the disease in the population, it is important to provide estimates of these values specific to scenarios in which such a diagnostic test would be applied. In this case, prevalence can be interpreted as "the probability before the test is carried out that the subject has the disease" as suggested by D. Altman [128]. The formulas for PPV and NPV can be modified based on Bayes' theorem as follows:

$$PPV = \frac{sensitivity * prevalence}{sensitivity * prevalence + (1 - specificity) * (1 - prevalence)} \quad (2.19)$$

$$NPV = \frac{specificity * (1 - prevalence)}{(1 - sensitivity) * prevalence + specificity * (1 - prevalence)} \quad (2.20)$$

Positive likelihood ratio (LR+): is the probability that an individual with disease has a positive test divided by the probability that an individual without disease has a positive test. The ratio indicates the effect of the increase on the probability of disease if the test is positive.

$$LR+= \frac{sensitivity}{1 - specificity} \quad (2.21)$$

Negative likelihood ratio (LR-): is the probability that an individual with the disease has a negative test divided by the probability that an individual without the disease has a negative test. The ratio indicates the effect of the decrease on the probability of disease if the test is negative.

$$LR-= \frac{1 - sensitivity}{specificity} \quad (2.22)$$

We report positive and negative likelihood ratios along with their confidence intervals employing the method described in [129].

Area Under the Receiver Operating Characteristic Curve (AUC): is a global assessment of the diagnostic accuracy of the test, summarising both sensitivity and specificity for a varying test discrimination threshold (Figure 5). A Receiver Operating Characteristic (ROC) plots the true positive rate (sensitivity) of the test against the false positive rate (1-specificity) at various thresholds. A random classifier has an AUC of 0.5, while a perfect classifier has an AUC of 1.

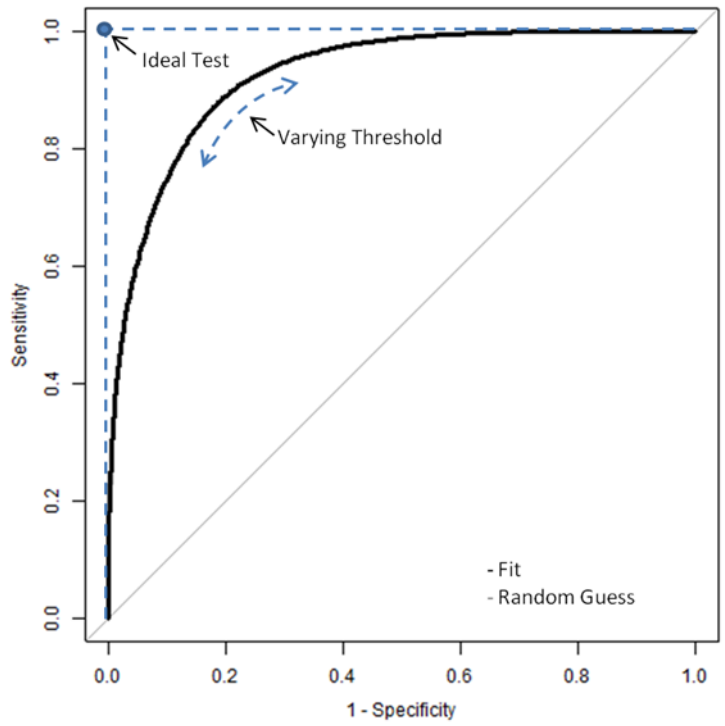


Figure 5: Characteristics of a ROC curve

The calculation of the confidence intervals at 95% for the aforementioned metrics was based on a non-parametric stratified bootstrap resampling. Each replicate contained the same number of cases and controls as the original sample and we ran 2000 bootstraps, as recommended by Carpenter et al. [130].

2.6 The pipeline

In order to ensure the generalisation of our models, avoid overfitting and estimate the accuracy of the classification, we employed the hold-out method as a cross-validation technique. The initial microarray dataset was randomly partitioned into two independent sets, the training set and the test set, comprising of 80% and 20% of the samples respectively. When different prominent subclasses were present, we maintained a balanced representation in both sets.

The training set was used to derive the transcriptional signatures, which were subsequently employed to classify the samples in the test set and independent validation cohorts. The process begins by identifying the transcripts that exhibit the most substantial variation in the training set and selecting them as candidate predictors. An elastic net model was then fitted on the subset of the training data defined by the candidate predictors. To avoid overfitting, the model's parameters were tuned using 10-fold cross validation. The test set was held out for the accuracy estimation of the elastic net model as well as the DRS. The derived signatures were also applied to independent validation cohorts, which can comprise either publicly available data or independently collected and processed data (Figure 6).

The pipeline is implemented in the 'R' Language and Environment for Statistical Computing, version 2.12.1. [131]. R integrates a multitude of packages that are developed and contributed by researchers worldwide. In addition there is a specific collection of packages named BioConductor, which provides tools for the analysis, visualization and comprehension of high-throughput biological data [132]. In Table 2 below the key BioConductor packages that were used in the analysis pipeline are summarised:

Task	**Tool**
pre-processing of microarray data	lumi [109]
differential expression analysis	limma [133]
elastic net model	glmnet [134], enet [122]
ROC curves	pROC [135]

Table 2: BioConductor packages used in the pipeline

Clustering tools, such as hierarchical clustering methods (i.e. heatmaps) were used to visualise the groups of patients and groups of genes with similar expression profiles throughout the pipeline.

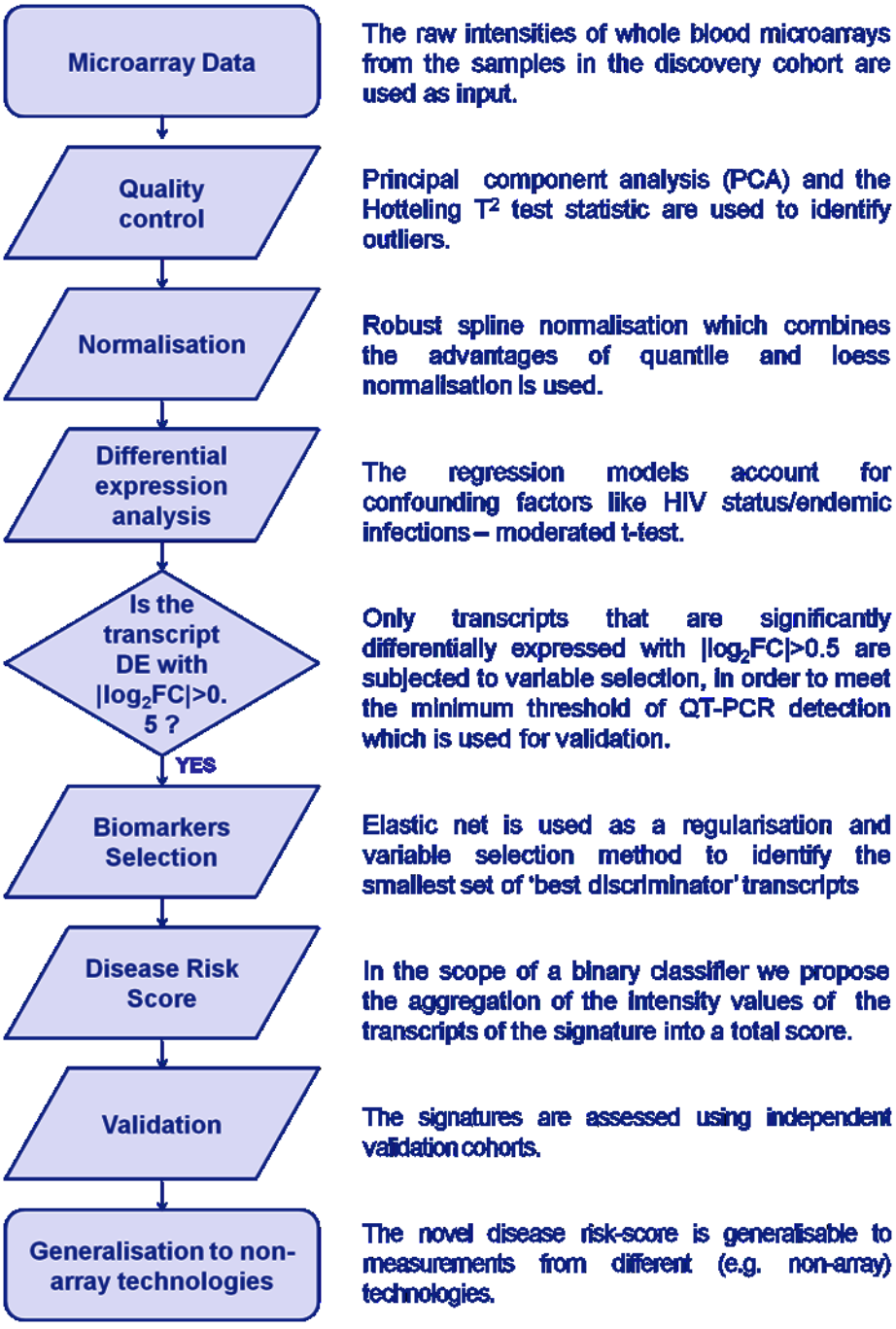


Figure 6: Methodological pipeline implemented using R and BioConductor

2.7 Summary

The use of whole-blood gene expression to derive diagnostic biomarkers capable of distinguishing phenotypically similar diseases remains challenging. In this chapter I proposed a pipeline for discovering biomarker signatures from whole blood transcriptomic data. The Disease Risk Score (DRS), which provides a new generalisable approach to reduce the dimensionality of multi-transcript signatures, was also introduced.

The pipeline was created with the particular challenges of acquiring gene expression data in an African setting in mind. Those include deriving a signature robust to geographical heterogeneity, to the presence of endemic underlying co-infections and to measurement variability from different assays. In the following chapters I will describe the application of the pipeline to two large gene expression studies from Africa.

Although the main steps of the bioinformatics pipeline were described in this chapter, statistical plug-ins will be introduced in the results chapters, tailored to the needs of the particular datasets.

Chapter 3

Identification of host gene expression biomarkers for TB in adults

In Chapter 3, I present the analysis of a large, prospective case-control study to identify host biomarkers of TB. Patient recruitment and laboratory work was carried out by colleagues in the ILULU consortium as stated in the Declaration of Originality. The study compared patients with culture proven TB to patients with a wide spectrum of other diseases as well as LTBI, including both HIV-infected and -uninfected subjects. After data exploration and pre-processing, the bioinformatics pipeline described in Chapter 2 is employed in order to identify the minimal transcriptional signatures able to discriminate between diseases and disease states. The performance of the signatures is first evaluated using the above cohort and then validated in an independent publicly available dataset via the calculation of patient disease risk scores.

3.1 Cohorts

To investigate the hypothesis that TB can be distinguished from LTBI and other diseases phenotypically similar to TB prevalent in HIV-infected and -uninfected African populations, a multicentre international study was set up. In order to enable generalisation of the findings to African countries with differing prevalence of malaria and other parasitic infections, as well as other environmental exposures that might affect transcription, two highly contrasting study sites were chosen (one urban, one rural) in two African countries with differing co-endemic diseases: Cape Town, South Africa (SA) and Karonga district, Northern Malawi. South Africa has one of the highest TB incidence rates in Africa (860 per 100,000) [61]. Out of the total number of TB cases, 62% are HIV-positive, while 90% are actually tested. In Malawi, the incidence of new TB cases is lower (156 per 100,000) while 56% of the TB cases are HIV-positive, out of the 90% that are tested. In Malawi, malaria and helminth infection are hyperendemic.

3.1.1 Recruitment

Patients presenting to hospitals in South Africa and Malawi with suspected tuberculosis were recruited and biological samples were collected, amongst which peripheral blood in Qiagen PAXgene RNA tubes for the microarray study. In addition, patients with latent tuberculosis infection (LTBI) were recruited in the community or from contact tracing clinics. Overall, 311 patients from South Africa and 273 from Malawi were recruited. All patients underwent examination adequate for suspected TB cases: chest radiographs and serological testing for HIV, along with cultures of blood, cerebrospinal fluid and urine, and biopsies for histological examination including TB culture. Two sputum samples obtained after induction or coughing were examined by standard microscopy for acid fast bacilli (AFB) and cultured for TB.

Healthy LTBI controls were recruited by random community selection (Malawi) and from HIV screening clinics (South Africa) from the same areas as patients with TB. TST and/or *in vitro* IGRAs was undertaken to validate LTBI categorisation. Individuals were

assigned into diagnostic groups based on examination results at enrolment and follow-up, as well as HIV status.

3.1.2 Diagnostic Process and data generation

Following the diagnostic work-up, patients were assigned to groups using the following definitions shown in Table 3 below.

Group	Definition
Definite TB case (TB)	Clinical condition consistent with TB and microbiological evidence (*M.tuberculosis* was cultured from sputum or tissue samples)
Latent TB infection (LTBI)	Clinically assessed as healthy without any clinical syndrome in which tuberculosis is likely. TST positive (>10mm or more if HIV-uninfected, or >5mm or more if HIV-infected) and/or a positive IGRA and negative sputum culture (if cough was productive)
Other disease case (OD)	Disease syndrome that on presentation included tuberculosis in the differential diagnosis, but following clinical investigation and management, tuberculosis was excluded and a firm alternative diagnosis was established

Table 3: Definitions for "gold standard" classification of patient samples of the adult study from South Africa and Malawi.

Individuals were either assigned to one of the diagnostic groups above or excluded once the results of investigations and follow-up were available (Figure 7). After investigation, any case with an established alternative diagnosis to TB, no microbiological evidence of TB and an absence of TB symptoms at the time of follow-up, or with an observed improvement of clinical symptoms on follow-up without TB treatment, was

recruited as an OD case. If TB could not be reliably ruled out of the differential, the patient was excluded.

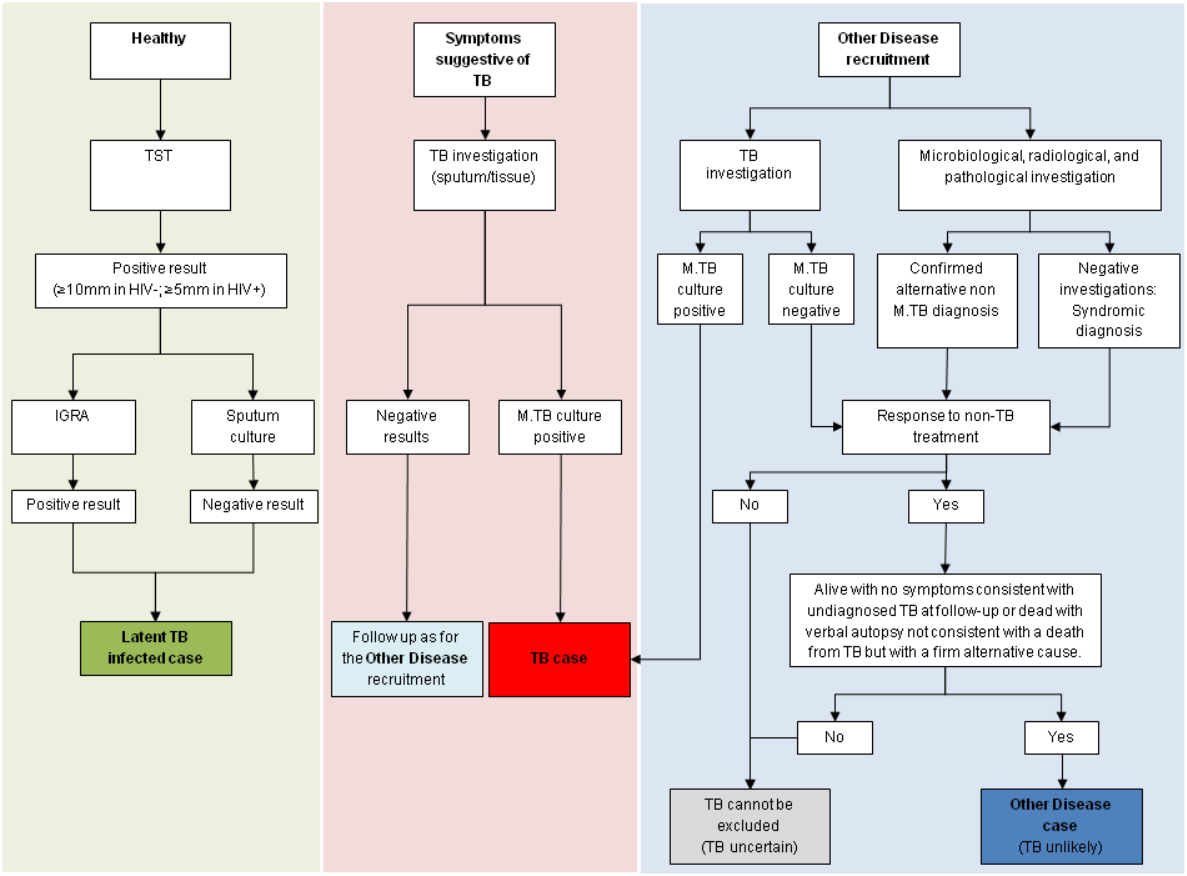


Figure 7: Diagnostic algorithm for patient classification into LTBI, active TB and other diseases groups.

In the RNA expression study 96 TB patients, 117 patients with OD and 98 LTBI individuals were included from Cape Town. From Malawi, the first 119 patients with TB, 77 patients with OD and 77 LTBI individuals were included (Figure 8).

Amongst the culture confirmed TB cases, there were some patients with negative smear results. In Cape Town, 36.7% (18/49) HIV-infected patients with TB were smear-negative and 8.5% (4/47) HIV-uninfected patients were smear-negative. In Malawi, 13.3% (8/60) HIV-infected patients with TB were smear-negative and 10.2% (6/59) HIV-uninfected patients were smear-negative. The Other Diseases (OD) group reflected the range of conditions with similar clinical manifestations to TB at each site. Patients were followed up

26 weeks post diagnosis to confirm that those with other diseases remained TB-free. The spectrum of infectious and malignant diseases in the OD cohort is shown in Table 4.

Disease Groupings	**HIV infected**		**HIV uninfected**		
	SA	**Malawi**	**SA**	**Malawi**	**Total**
Pneumonia/LRTI/PJP	24 (35%)	19 (50%)	5 (10%)	13 (33%)	**61 (31%)**
Malignancy & other neoplasia* (other than Kaposi's sarcoma)	2 (3%)	4 (11%)	17 (35%)	5 (13%)	**28 (14%)**
Pelvic inflammatory disease / UTI	4 (6%)	1 (3%)	15 (31%)	5 (13%)	**25 (13%)**
Bacterial, viral meningitis or meningitis of uncertain origin	4 (6%)	4 (11%)	0 (0%)	6 (15%)	**14 (7%)**
Hepatobiliary disease	6 (9%)	0 (0%)	7 (14%)	0 (0%)	**13 (7%)**
Febrile syndromes of uncertain origin	1 (1%)	3 (8%)	1 (2%)	6 (15%)	**11 (6)%**
Kaposi's sarcoma	9 (13%)	1 (3%)	0 (0%)	0 (0%)	**10 (5%)**
Cryptococcal meningitis	6 (9%)	4 (11%)	0 (0%)	0 (0%)	**10 (5%)**
Non TB pleural effusion / Empyema	5 (7%)	0 (0%)	2 (4%)	0 (0%)	**7 (4%)**
Gastroenteritis	5 (7%)	0 (0%)	0 (0%)	0 (0%)	**5 (3%)**
Peritonitis	0 (0%)	1 (3%)	0 (0%)	3 (8%)	**4 (2%)**
Other**	0 (0%)	1 (3)%	2 (4%)	1 (3%)	**4 (2%)**
Gastric ulcer or gastritis	2 (3%)	0 (0%)	0 (0%)	0 (0%)	**2 (1%)**
Total	**68**	**38**	**49**	**39**	**194**

Table 4: Major clinical diagnoses of the patients in the other diseases (OD) group.

* (14) Bronchial carcinoma, (4) Lymphoma, (1) Cervical carcinoma, (1) Ovarian carcinoma, (1) mesothelioma, (1) gastric carcinoma, (4) metastatic carcinoma of unknown origin, (1) benign salivary tumour, (1) Dermatological tumour

** (1) HIV-related lymphadenopathy, (1) Crohn's, (1) Orchitis, (1) Pyomyositis

LRTI: Lower Respiratory Tract Infection, UTI: Urinary tract infection, PJP: *Pneumocystis jirovecii* pneumonia

Uniform protocols for sample collection, separation, recovery of RNA and storage were utilised in both sites. Whole blood was collected in PAXgene RNA tubes, frozen at -20°C within 3 hours, and then stored at -80°C. RNA was extracted using PAXgene blood RNA kits at one site (Cape Town) to minimise any sample handling bias. The integrity and yield of the total RNA was assessed using an Agilent 2100 Bioanalyser and a NanoDrop 1000 spectrophotometer respectively. Total RNA was then shipped to the Genome Institute of Singapore (GIS). After quantification and quality control, biotin-labeled cRNA was prepared using Illumina TotalPrep RNA Amplification kits (Applied Biosystems) from 500ng RNA. Labelled cRNA was hybridised overnight to Human HT-12 V4 Expression BeadChip arrays (Illumina). After washing, blocking and staining, the arrays were scanned using an Illumina BeadArray Reader. Using Genome Studio software the microarray images were inspected for artefacts and QC parameters were assessed (Appendix). During sample handling for RNA expression analysis 48 samples were excluded: 45 due to low RNA concentration prior to amplification, 2 failed amplification and 1 was excluded during the data QC. A total of 536 microarrays were used for gene expression analysis (Figure 8). The expression data has been submitted to Gene Expression Omnibus (record GSE37250).

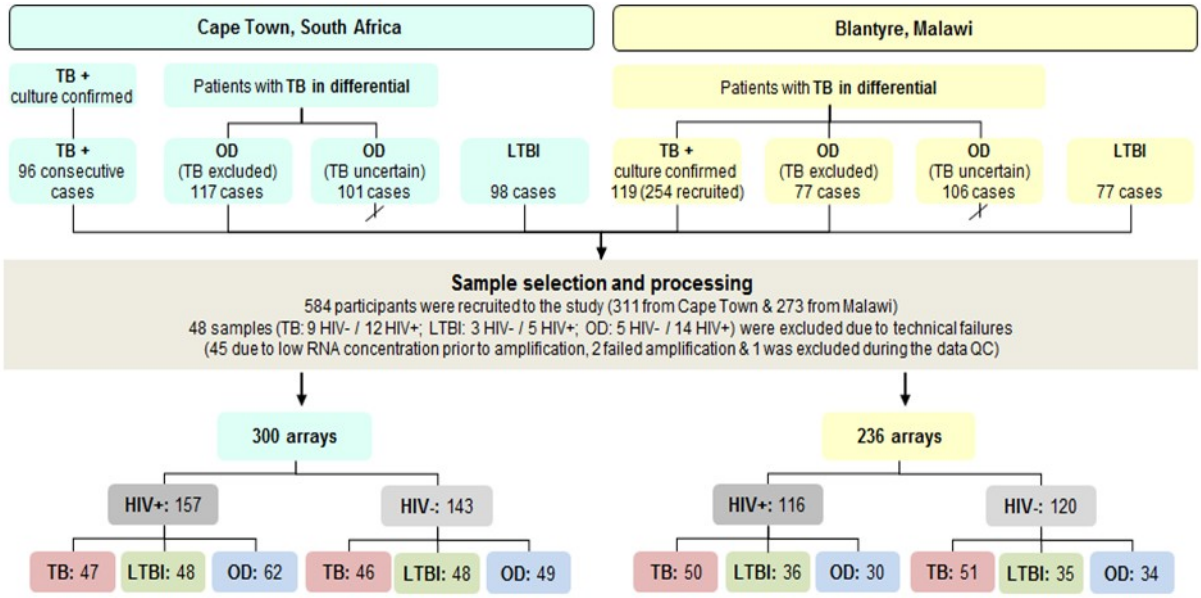


Figure 8: South Africa/Malawi Patient cohorts and microarrays
Prospective cohorts of patients undergoing investigation for TB were established in two sub-Saharan African countries (South Africa, Malawi) with differing patterns of endemic diseases. TB uncertain: patients for whom TB diagnosis could not be excluded were not included in the study.

3.2 Dataset exploration

First we pre-processed the microarray data as described in Chapter 2. Then, after background adjustment and normalisation, quality control was performed in order to examine the effect of disease state on transcript expression and to check for possible assignment errors. PCA analysis was employed and the first three principal components were examined, based on all transcripts on all samples (Figure 9). Inspection revealed that the primary clustering was based on disease state (TB, LTBI or OD) rather than geographical location or HIV status. The sample highlighted (categorised as active TB HIV+ from Malawi) was removed from the analysis. The elliptic rings represent levels of confidence (0.9 inner circle; 0.9999 outer circle).

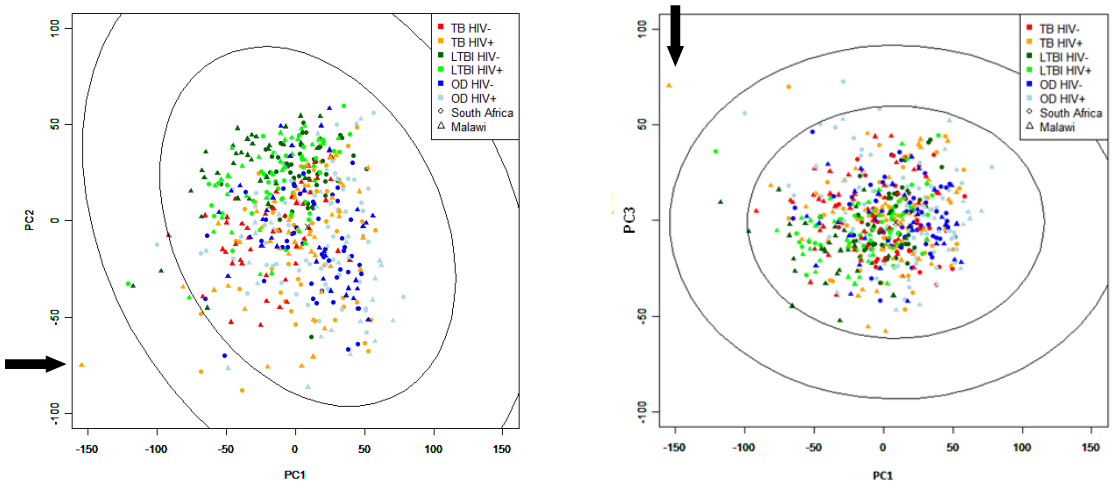


Figure 9: Principal components analysis (PCA) plot of the microarray samples.
The PCA plot is based on all transcripts and all samples after background adjustment and normalisation. On the left PCA1 & PCA2 are plotted and on the right PCA1 & PCA3. The sample highlighted was removed from the analysis. Rings are levels of confidence (0.9 inner circle; 0.9999 outer circle).

The six first principal components were examined further. The ratio of variance that can be predicted independently by the PCs (R^2 values) shows the importance of each component. In this analysis the R^2 values for the principal components are: R^2 (PC1) = 0.128; R^2 (PC2) = 0.108; R^2 (PC3) = 0.049; R^2 (PC4) = 0.045; R^2 (PC5) = 0.036; R^2 (PC6) = 0.031. In order to interpret the principal components, a correlation analysis was used to identify which variables are strongly correlated with each component. A threshold of 0.5 for

correlation was used to indicate strong correlation. As for disease status, only the LTBI samples had a correlation with PC2 over 0.5 (0.52), indicating that PC2 is probably capturing a "healthy" component of gene expression (Figure 9). Location (South Africa) was correlated with PC5 (-0.57) and HIV status was correlated with PC6 (0.55) (Figure 10).

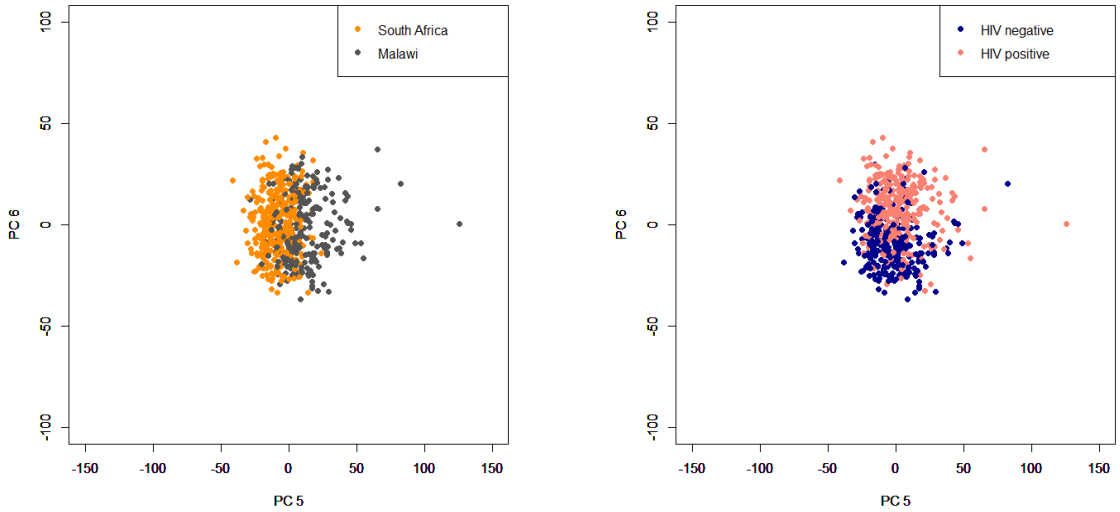


Figure 10: Principal components correlated with location and HIV status.
Principal components 5 and 6 are shown. On the left plot colour coding depicts the location (South Africa or Malawi), and on the right plot colour coding depicts the HIV status (negative or positive).

As previous gene expression TB studies have not considered different locations and HIV prevalence, we explored if there was evidence of a robust underlying signature of TB, independent of HIV status and geographical location. Two pairwise comparisons were of particular interest as they reflect the real world diagnostic questions: TB vs. LTBI and TB vs. OD. In contact screening, the aim is to distinguish active TB from latent infection, whereas distinguishing TB from other phenotypically similar diseases is the challenge in the clinical context of a sick patient.

Separate linear models were fitted to the data for the different HIV groups and the different locations using the suitable design matrices: TB vs. LTBI in HIV-uninfected patients from South Africa, TB vs. LTBI in HIV-uninfected patients from Malawi, TB vs. LTBI in HIV-infected patients from South Africa, TB vs. LTBI in HIV-infected patients from Malawi. The same comparisons were performed for TB vs. OD. The outputs were lists of transcripts with adjusted p-values reflecting the level of confidence for evidence of differential expression.

The negative logarithm of the corrected p-values was plotted and their Pearson correlation coefficient was calculated to assess how location and HIV status affect the signatures. Substantial correlation was observed for TB vs. LTBI differential expression across different geographic locations regardless of HIV co-infection (Figure 11, Table 5). Less correlation was observed for TB vs. OD between the different locations which may reflect the different spectra of conditions that TB has to be differentiated from in the two different sub-Saharan regions (Figure 11, Table 5).

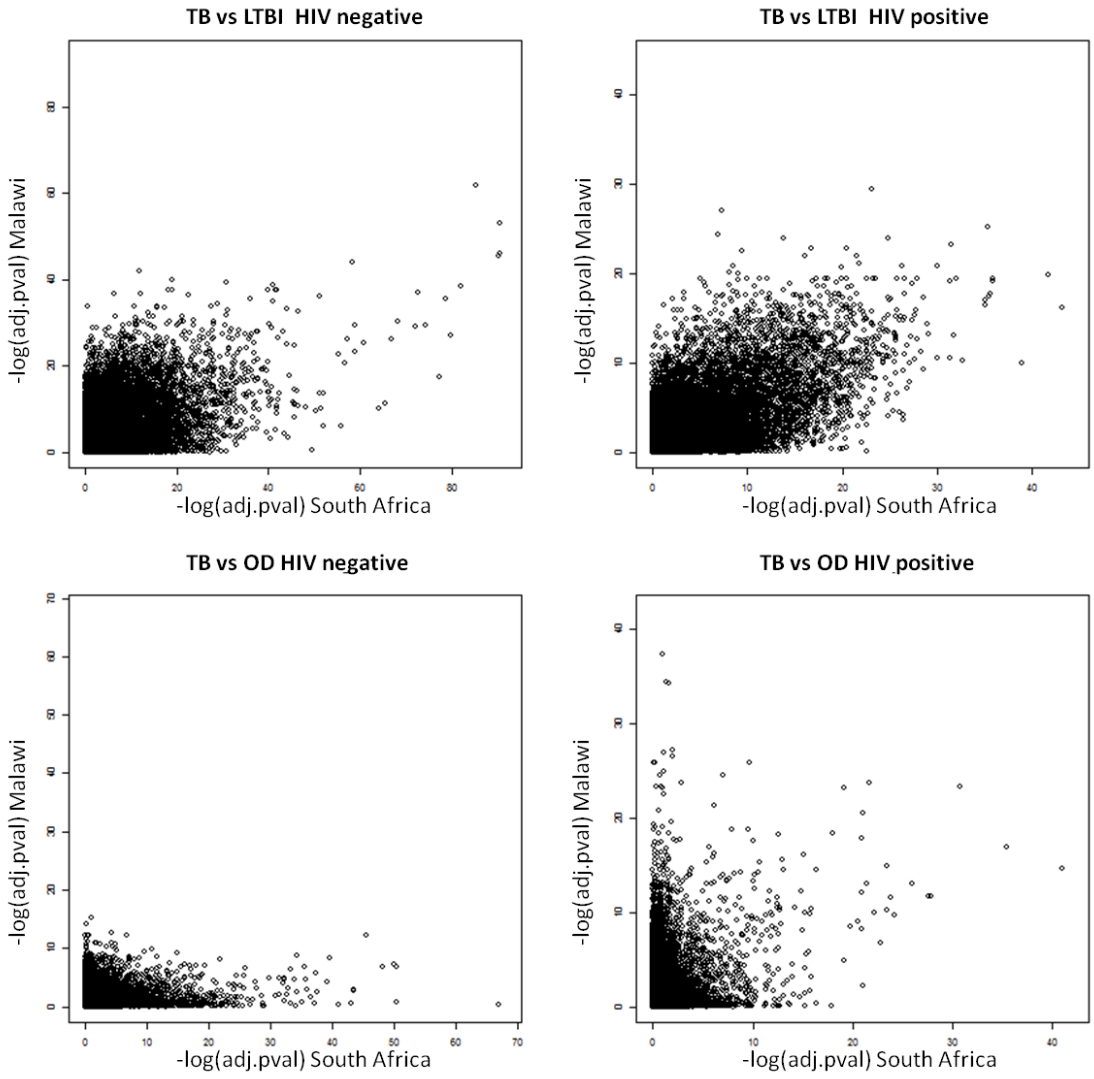


Figure 11: Concordance of differential expression across sites for the TB vs. LTBI comparison (top) and TB vs. OD comparison (bottom), by HIV status (uninfected: left, infected: right).

Each circle corresponds to a transcript and the negative logarithm of the corrected p-values is plotted for each comparison.

Comparison	**Correlation Coefficient for Locations**	
TB vs. LTBI HIV negative	0.56 (0.55-0.56)	p-value < 2.2e-16
TB vs. LTBI HIV positive	0.70 (0.69-0.71)	p-value < 2.2e-16
TB vs. OD HIV negative	0.11 (0.10-0.12)	p-value < 2.2e-16
TB vs. OD HIV positive	0.26 (0.25-0.27)	p-value < 2.2e-16

Table 5: Pearson correlation coefficient for the negative logarithm of the corrected p-values for the different pairwise comparisons.

Regarding the different HIV groups, the correlations for all the comparisons were positive (Figure 12, Table 6). However, the correlations for TB vs. OD HIV-infected and -uninfected individuals were less strong than for TB vs. LTBI, reflecting the different spectrum of conditions among the HIV-infected individuals that mimic the symptoms of TB disease.

Comparison	**Correlation Coefficient for HIV status**	
South Africa TB vs. LTBI	0.66 (0.65-0.67)	p-value < 2.2e-16
South Africa TB vs. OD	0.55 (0.54-0.56)	p-value < 2.2e-16
Malawi TB vs. LTBI	0.78 (0.77-0.79)	p-value < 2.2e-16
Malawi TB vs. OD	0.34 (0.33-0.36)	p-value < 2.2e-16

Table 6: Pearson correlation coefficient for the negative logarithm of the corrected p-values for the different pairwise comparisons.

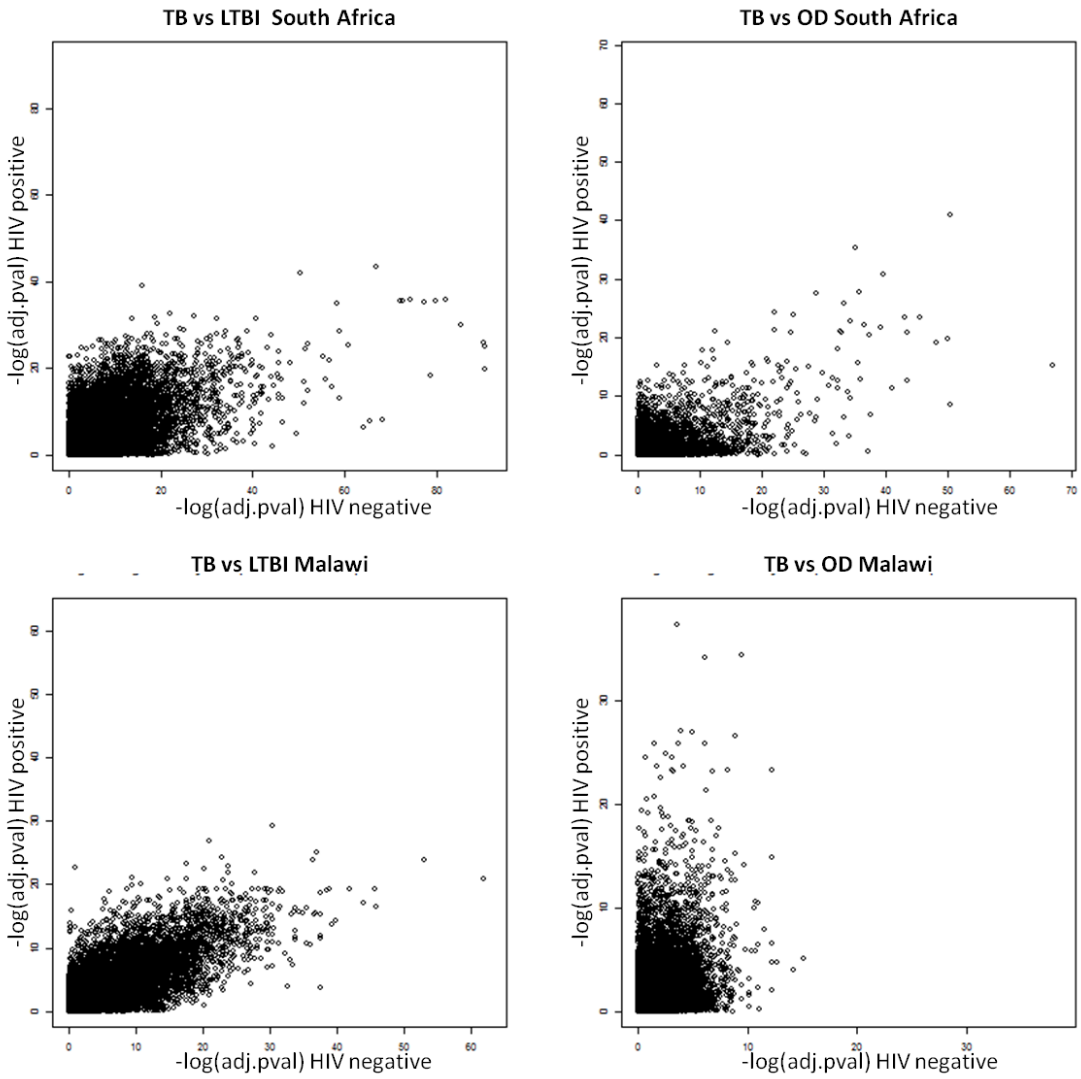


Figure 12: Concordance of differential expression for HIV status for the TB vs. LTBI comparison (left) and TB vs. OD comparison (right), by location (SA: top, Malawi: bottom). Each circle corresponds to a transcript and the negative logarithm of the corrected p-values is plotted for each comparison is plotted.

This analysis provides evidence that a substantial number of transcripts are significantly differentially expressed irrespective of location and HIV status, with the same direction of expression. However, HIV status and location should be accounted for in the linear models, and in any partitioning of the dataset, to ensure the identification of robust TB signatures independent of HIV status or geographical location.

3.3 Power calculation

In order to evaluate the suitability of the sample cohorts to detect effects that would be appropriate for biomarker use in this study, a power calculation was used. As the study has two diagnostic arms (TB vs. LTBI and TB vs. OD) different calculations were implemented for each comparison.

For an anticipated sensitivity of 90% and a specificity of 90% for the TB vs. LTBI comparison in adults, a sample size of 194 TB patients and 167 LTBI patients will provide 95% confidence intervals (CI) spanning 4.2% from the point estimate of sensitivity and 4.7% from the point estimate of specificity. In an 80-20 split of the cohort, the 95% CI will span 10.3% for sensitivity and 10.9% for specificity for the test set.

For an anticipated sensitivity of 90% and a specificity of 90% for the TB vs. OD comparison in adults, a sample size of 194 TB patients and 175 OD patients will provide 95% CI spanning 4.2% from the point estimate of sensitivity and 4.5% from the point estimate of specificity. In an 80-20 split of the cohort the 95% CI will span 10.4% for sensitivity and 10.4% for specificity for the test set. The adjusted Wald method was used for the calculations.

3.4 Results

3.4.1 Overview

To identify transcript signatures applicable across geographic locations and in patients with differing HIV status, we combined HIV-infected and -uninfected patient cohorts from South Africa and Malawi. The recruited participants were randomly assigned into a training set (80% of the participants) and a test set (20%) with no overlap, using the "sample()" function without replacement in R which obtains a subset of a given set [136]. We used the training set to determine the transcript signatures (80% of the samples to train elastic net and 20% to test elastic net and the disease risk score) and an independent published cohort to validate the findings [60].

3.4.2 Signatures Discovery

To find minimal transcript sets required to discriminate TB from other patient groups, we applied the pipeline described in Chapter 2 to the training cohort, for the two comparisons of interest: TB vs. LTBI and TB vs. OD (Figure 13).

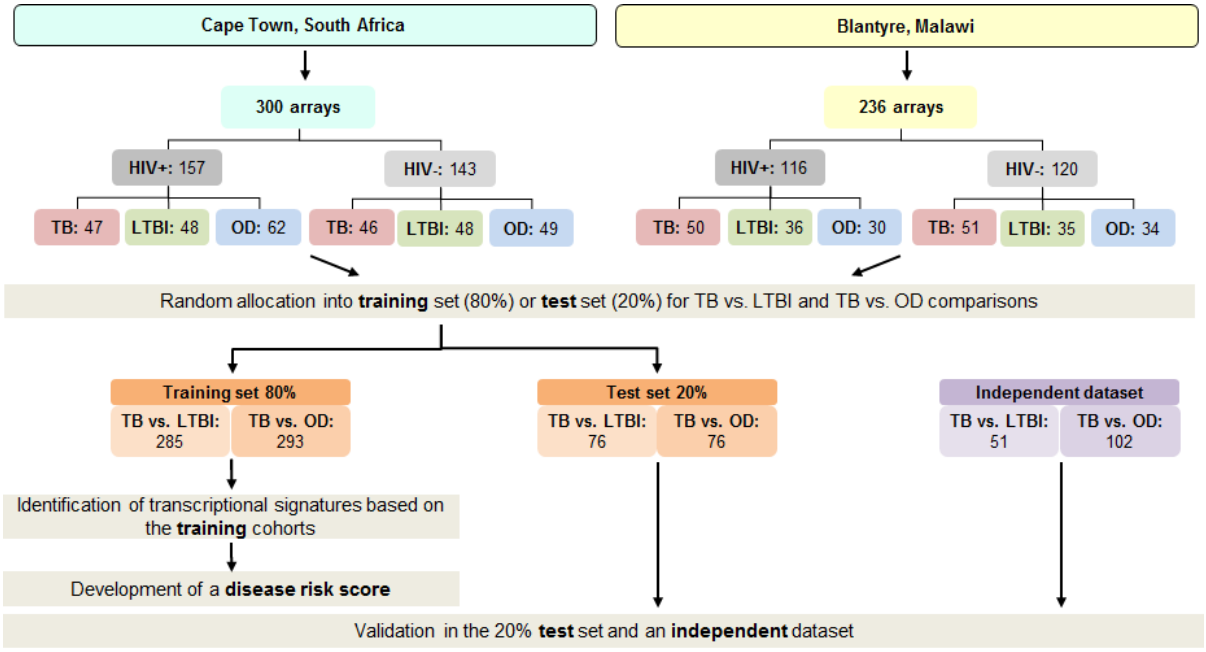


Figure 13: Microarray analysis plan for the adult dataset from South Africa and Malawi.

3.4.2.1 TB vs. LTBI signature

The training set for the TB vs. LTBI comparison comprised of 285 individuals while the test set comprised of 76 individuals. After fitting a linear model using the geographical location and HIV status as factors, 2054 transcripts passed the $|\log_2 FC|>0.5$ criterion for the TB vs. LTBI comparison and were used as candidate predictors for elastic net modelling. The tuning parameters of the model was cross-validated in the two dimensional space. For every combination of the lasso (λ_1) and ridge regression penalty parameters (λ_2), the entire solution path is calculated and the 10-fold cross-validated mean squared prediction error was computed. The different values of lasso parameter are reflected by the s parameter, which corresponds to λ_1 standardised by the full solution obtained via Ordinary Least Squares (0 to 1). For the minimum square error (MSE) λ_2=0.5 and s=0.6 (Figure 14).

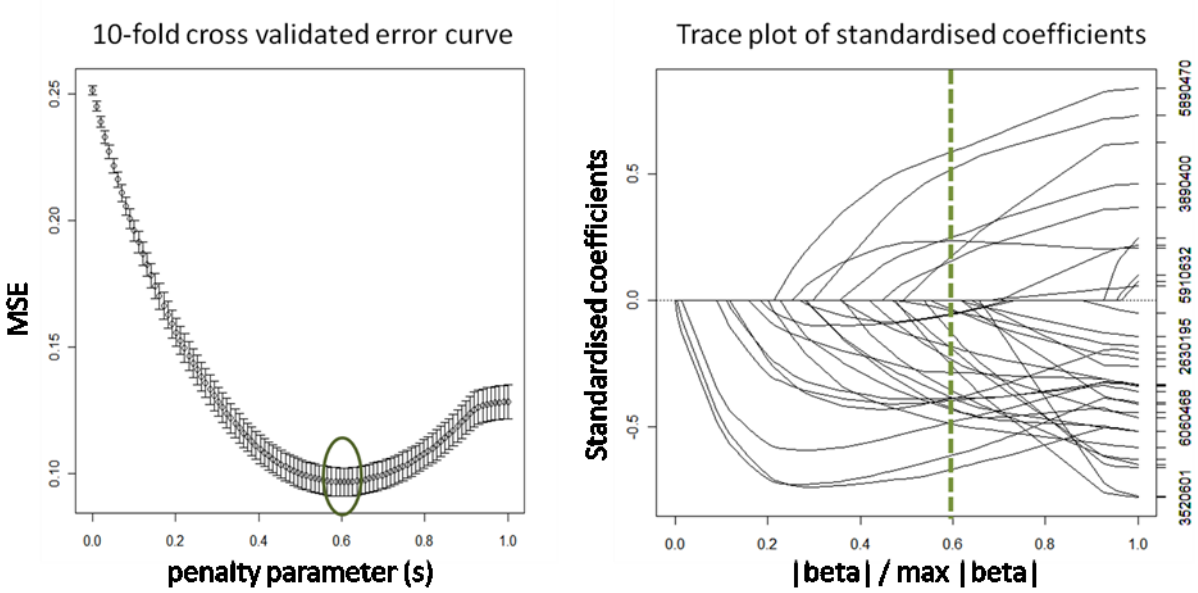


Figure 14: Mean Squared Prediction Error (MSE) plot and trace plot of the standardised coefficients for the TB vs. LTBI comparison (s=0.6).

For these parameters the model was fitted to the training data to estimate the weights (the standardised coefficients). The plot shows the number of zero and non-zero coefficients across the solution path. The vertical line represents the value with minimum mean squared error (acquired from the CV), which was obtained through model training. A

27 transcript model was identified for discriminating TB from LTBI which was subsequently used for discrimination of the samples of the training, as well as the test set.

The elastic net model had very good classification accuracy in the training as well as the test set, with sensitivity and specificity of 89% $CI_{95\%}$(78-97) and 90% $CI_{95\%}$(80-97) respectively (Table 7). The accuracy of the classifier was also assessed separately for the HIV-infected and -uninfected groups, with the classifier performing better on the HIV-uninfected individuals, AUC 100% $CI_{95\%}$(100-100) vs. AUC 96% $CI_{95\%}$(91-100). This might be explained by the lower cell counts that HIV infected patients have and/or the alteration in gene expression caused by the presence of the underlying HIV infection.

	HIV+/-		**HIV-**		**HIV+**	
	Training	**Test**	**Training**	**Test**	**Training**	**Test**
	(95% CI)	(95% CI)	(95% CI)	(95% CI)	(95% CI)	(95% CI)
AUC	**97%**	**97%**	**99%**	**100%**	**95%**	**96%**
(95% CI)	(95-98%)	(94-99%)	(98-100%)	(100-100%)	(91-98%)	(91-100%)
Sensitivity	**87%**	**89%**	**84%**	**84%**	**89%**	**94%**
(95% CI)	(81-92%)	(78-97%)	(75-92)	(68-100%)	(81-96%)	(83-100%)
Specificity	**91%**	**90%**	**99%**	**100%**	**84%**	**80%**
(95% CI)	(86-96%)	(80-97%)	(96-100%)	(100-100%)	(75-91%)	(60-95%)

Table 7: Classification of the South Africa/Malawi training and test sets using the 27 TB/LTBI transcript signature elastic net model.
The elastic net model was applied to the South Africa/Malawi training and test set. Metrics were calculated for the HIV-infected and -uninfected patients combined, as well as separately.

In addition, unsupervised clustering analysis was performed on the data using the 27 TB/LTBI transcript signature identified by the elastic net. We calculated the Euclidean distance measure for the expression values of the transcripts of the samples in the test set to create a heatmap. In the hierarchical clustering we observed two distinct clusters corresponding to TB and LTBI individuals, regardless of HIV co-infection (Figure 15). The rows correspond to transcripts; the red colour scheme indicates up-regulation, and green down-regulation. Columns correspond to patients regardless of HIV status; purple colour

indicates patients with TB and green indicates individuals with LTBI. Dendrograms are added for both the transcripts and the patients to indicate the level of similarity in the groupings. Good clustering of the groups based on the signature reflects its classificatory potential.

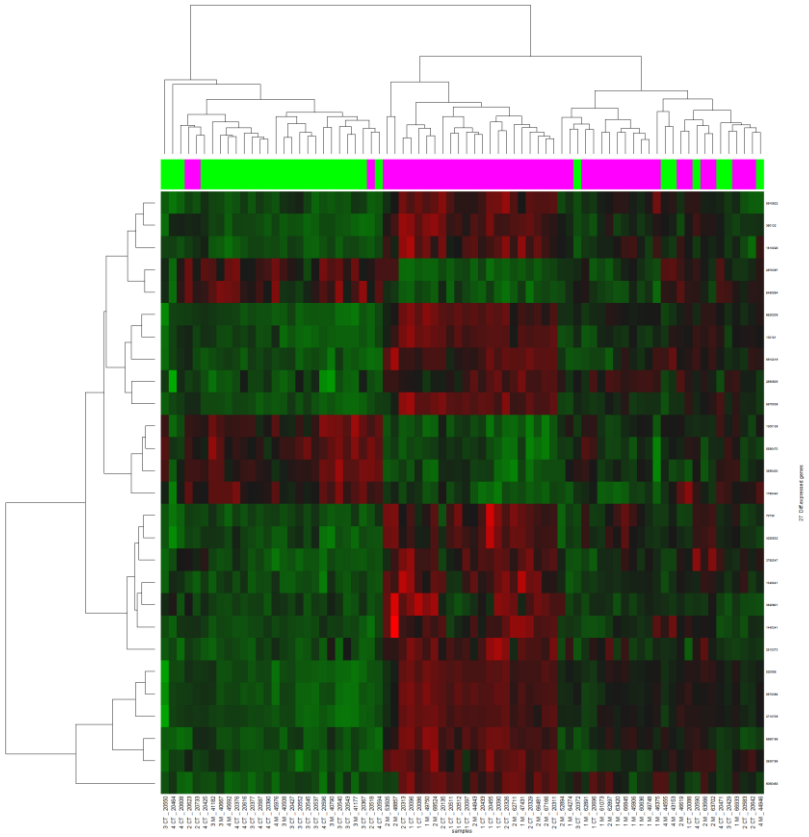


Figure 15: Heatmap showing clustering of the South African/Malawi test set using the 27 TB/LTBI signature.
The rows correspond to transcripts (red colour indicates up-regulation; green colour indicates down-regulation). Columns correspond to patients regardless of HIV status (purple=TB; green=LTBI).

The disease risk score based on the 27 transcript signature was calculated using equation 2.12 for the patients of the training and the test set. In the combined HIV-infected and -uninfected test set, the 27 transcript disease risk score discriminated TB from LTBI with sensitivity and specificity of 95% $CI_{95\%}$(87-100) and 90% $CI_{95\%}$(80-97) respectively (Figure 16, Table 8). Its classificatory power was also assessed for the HIV-infected and uninfected individuals separately (Figure 17). It achieved perfect classification in the HIV-uninfected group (AUC=100%, $CI_{95\%}$(100-100)) while it had a slightly reduced accuracy in the HIV-infected group (AUC=97%, $CI_{95\%}$(95-100)) (Table 8).

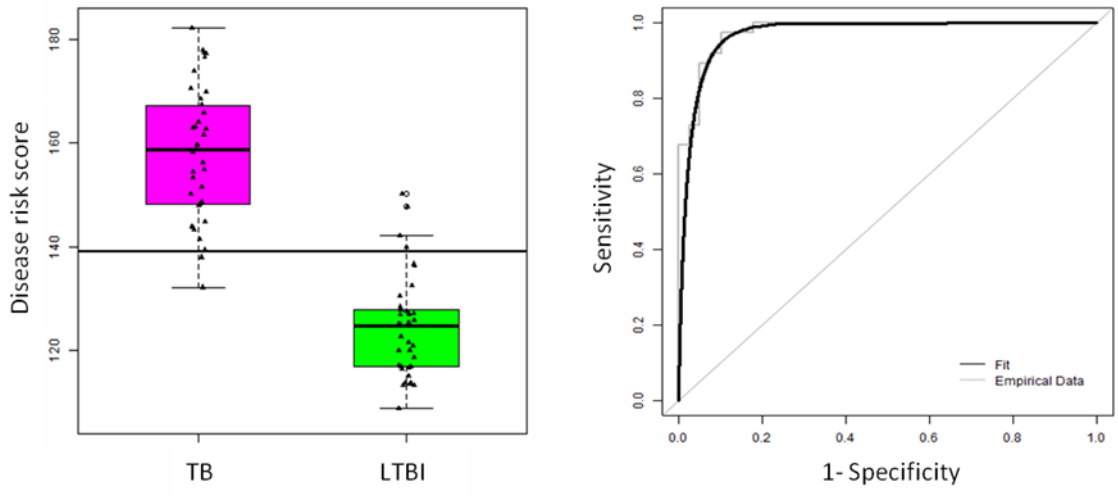


Figure 16: Classification of the South African/Malawi test set using the Disease Risk Score based on the 27 TB/LTBI transcript signature.
Disease risk score (left) and receiver operating characteristic curve (right) based on the TB/LTBI 27 transcript signature applied to the combined HIV infected and uninfected South African/Malawi test cohort. Sensitivity, specificity and AUC are reported in Table 8.

	HIV+/-		**HIV-**		**HIV+**	
	Training	**Test**	**Training**	**Test**	**Training**	**Test**
	(95% CI)	(95% CI)	(95% CI)	(95% CI)	(95% CI)	(95% CI)
AUC	**95%**	**98%**	**98%**	**100%**	**92%**	**97%**
(95% CI)	(93-97%)	(95-100%)	(97-100%)	(100-100%)	(88-96%)	(95-100%)
Sensitivity	**87%**	**95%**	**91%**	**100%**	**81%**	**94%**
(95% CI)	(81-92%)	(87-100%)	(85-97)	(100-100%)	(72-90%)	(83-100%)
Specificity	**87%**	**90%**	**89%**	**100%**	**86%**	**90%**
(95% CI)	(81-92%)	(80-97%)	(81-95%)	(100-100%)	(77-94%)	(75-100%)

Table 8: Classification of the South Africa/Malawi training and test sets using the Disease Risk Score based on the 27 TB/LTBI transcript signature.
The DRS based on the 27 TB/LTBI transcript signature was applied to the South African/Malawi training and test sets. Metrics were calculated for the HIV-infected and -uninfected patients combined, as well as separately.

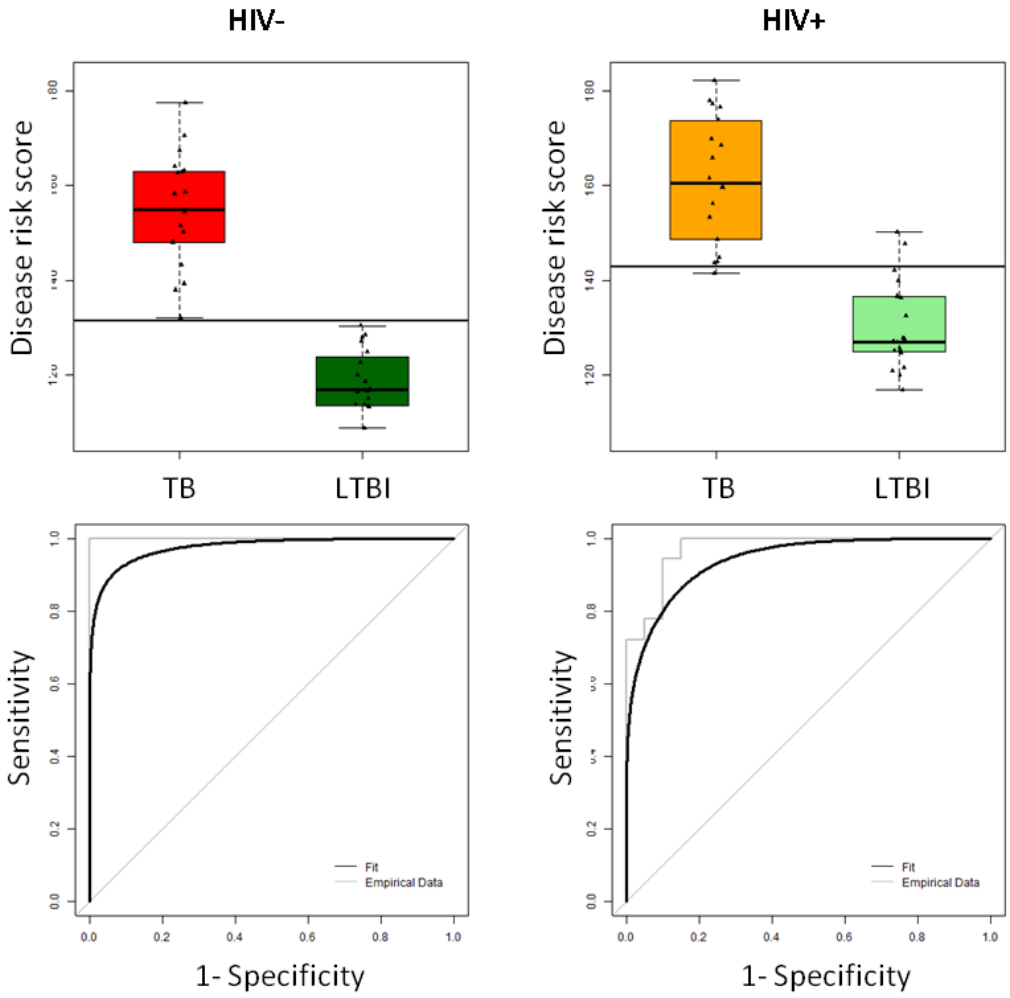


Figure 17: Classification of the South African/Malawi test set by HIV status using the Disease Risk Score based on the 27 TB/LTBI transcript signature.
Disease risk score (top panels) and receiver operating characteristic curves (bottom panels) based on the 27 TB/LTBI transcript signature applied to the HIV-uninfected and -infected patients of the South African/Malawi test set separately. Sensitivity, specificity and AUC are reported in Table 8.

Overall, the disease risk score has a very similar classificatory performance to the elastic net classifier both in the training and the test set, even though it is not using any weighting for the expression values of the transcripts, apart from their direction of regulation.

3.4.2.2 TB vs. OD signature

In the TB vs. OD comparison 379 transcripts were used as candidate predictors for elastic net modelling. The training set comprised of 293 patients, while the test set comprised of 76 patients. Cross-validation of the elastic net tuning parameters in the training set resulted in a model with 44 variables for λ_2=0.5 and s=0.013 (Figure 18).

The predictive performance of the 44 transcript elastic net model on the test set was estimated. The elastic net model had a sensitivity and specificity of 83% $CI_{95\%}$(71-93) and 97% $CI_{95\%}$(91-100) respectively, when applied to the test set (Table 9). The accuracy of the classifier was also assessed separately for the HIV-infected and -uninfected groups, with its performance being 1% better for the HIV-uninfected individuals (AUC=95%, $CI_{95\%}$(88-100) in the HIV-uninfected vs. AUC=94%, $CI_{95\%}$(84-100) in the HIV-infected group).

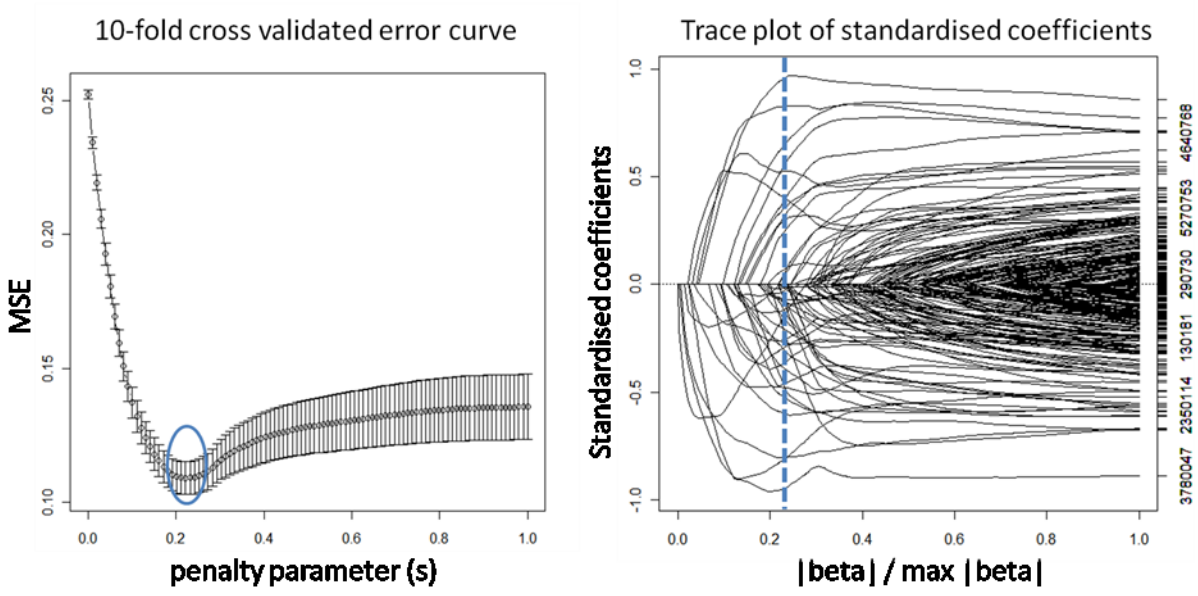


Figure 18: Mean Squared Prediction Error (MSE) plot and trace plot of the standardised coefficients for the TB vs. LTBI comparison (s=0.013).

	HIV+/-		HIV-		HIV+	
	Training (95% CI)	**Test** (95% CI)	**Training** (95% CI)	**Test** (95% CI)	**Training** (95% CI)	**Test** (95% CI)
AUC	**97%**	**94%**	**97%**	**95%**	**97%**	**94%**
(95% CI)	(95-98%)	(88-99%)	(94-100%)	(88-100%)	(94-99%)	(84-100%)
Sensitivity	**93%**	**83%**	**95%**	**82%**	**92%**	**85%**
(95% CI)	(90-97%)	(71-93%)	(89-99)	(64-96%)	(86-97%)	(70-100%)
Specificity	**89%**	**97%**	**90%**	**100%**	**88%**	**95%**
(95% CI)	(83-94%)	(91-100%)	(82-96%)	(100-100%)	(80-95%)	(84-100%)

Table 9: Classification of the South Africa/Malawi training and test sets using the 44 TB/OD transcript signature elastic net model.
The elastic net model was applied to the South Africa/Malawi training and test set. Metrics were calculated for the HIV infected and uninfected patients combined, as well as separately.

Unsupervised cluster analysis on the data using the identified 44 TB/OD transcript signature was performed. Based on the Euclidean distance measure, two distinct clusters were observed corresponding to TB and OD patients regardless of HIV co-infection (Figure 19). The rows correspond to transcripts; red colour scheme indicates them being up-regulated, and green down-regulated. Columns correspond to patients regardless of HIV status; purple colour indicates patients with TB while blue indicates patients with other diseases.

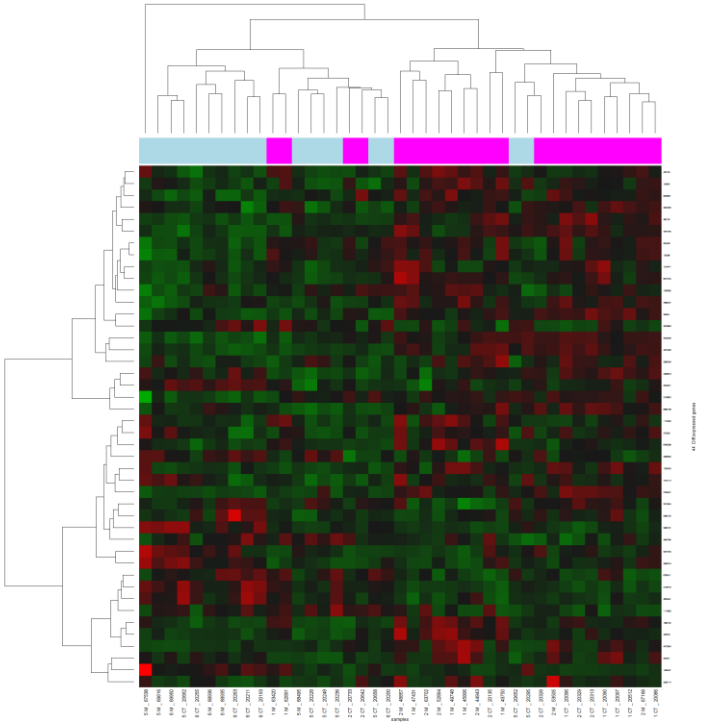


Figure 19: Heatmap showing clustering of the test set using the 44 TB/OD signature.
The rows correspond to transcripts (red colour indicates up-regulation; green colour indicates down-regulation). Columns correspond to patients regardless of HIV status (purple=TB; blue=OD).

The disease risk score based on the 44 transcript signature for every patient on the training and the test set was then calculated. In the test set, DRS had sensitivity and specificity of 93% $CI_{95\%}$(83-100) and 88% $CI_{95\%}$(74-97) respectively (Figure 20, Table 10). Its accuracy remained consistent in the HIV-infected and -uninfected test cohorts, with AUC 96% $CI_{95\%}$(89-100) and 94% $CI_{95\%}$(83-100) respectively (Figure 21, Table 10). Similar metrics were obtained when the DRS was evaluated in the training dataset, demonstrating its robustness.

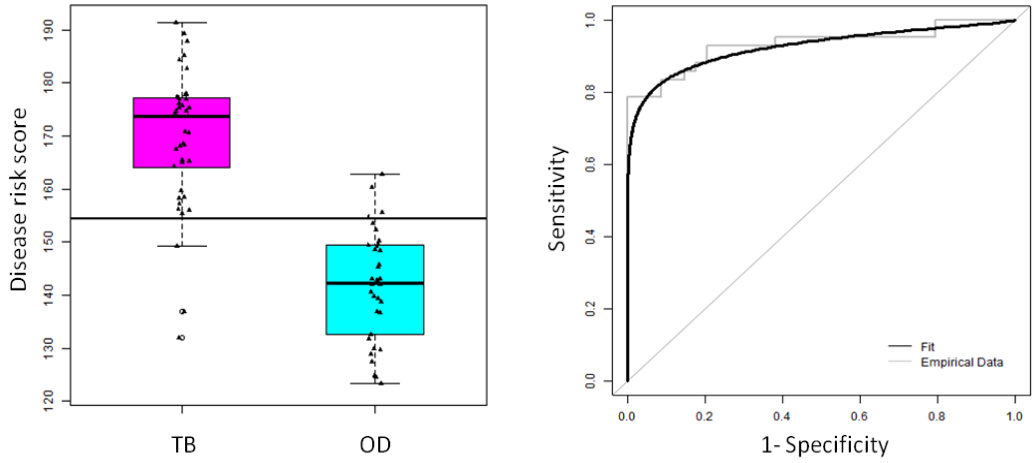


Figure 20: Classification of the South African/Malawi test set using the Disease Risk Score based on the 44 TB/OD transcript signature.

Disease risk score (left) and receiver operating characteristic curve (right) based on the TB/OD 44 transcript signature applied to the combined HIV-infected and -uninfected South African/Malawi test cohort. Sensitivity, specificity and AUC are reported in Table 10.

	HIV+/-		**HIV-**		**HIV+**	
	Training	**Test**	**Training**	**Test**	**Training**	**Test**
	(95% CI)	(95% CI)	(95% CI)	(95% CI)	(95% CI)	(95% CI)
AUC	**96%**	**95%**	**97%**	**96%**	**95%**	**94%**
(95% CI)	(94-98%)	(89-99%)	(94-99%)	(89-100%)	(92-98%)	(83-100%)
Sensitivity	**88%**	**93%**	**89%**	**91%**	**86%**	**95%**
(95% CI)	(82-93%)	(83-100%)	(83-96)	(77-100%)	(78-92%)	(85-100%)
Specificity	**87%**	**88%**	**88%**	**93%**	**86%**	**84%**
(95% CI)	(82-92%)	(74-97%)	(79-96%)	(80-100%)	(78-95%)	(68-100%)

Table 10: Classification of the South Africa/Malawi training and test sets using the Disease Risk Score based on the 44 TB/OD transcript signature.

The DRS based on the 44 TB/OD transcript signature was applied to the South African/Malawi training and test sets. Metrics were calculated for the HIV-infected and -uninfected patients combined, as well as separately.

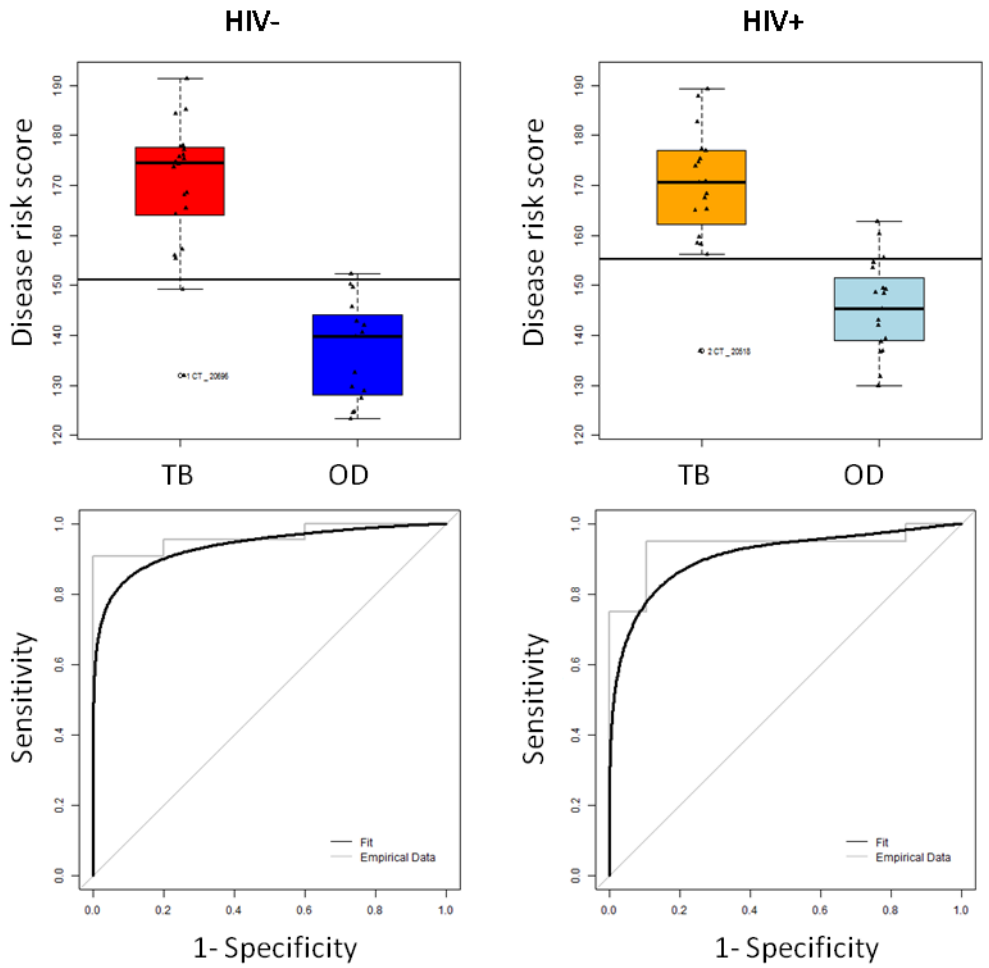


Figure 21: Classification of the South African/Malawi test cohort by HIV status using the Disease Risk Score based on the 44 TB/OD transcript signature.
Disease risk score (top panels) and receiver operating characteristic curves (bottom panels) based on the 44 TB/OD transcript signature applied to the HIV-uninfected and -infected patients of the South African/Malawi test set separately. Sensitivity, specificity and AUC are reported in Table 10.

The test set (20% of the samples) was used to evaluate the performance of the signatures and the DRS and the good accuracy of the classification showed that we did not over fit on the training data. However, further benchmarking of both the signatures and the DRS technique using an independent validation cohort, as well as estimating their performance on subgroups of interest of the data was crucial.

3.5 Benchmarking

3.5.1 Performance of the DRS on the Berry *et al.* dataset

In order to further assess the performance of the disease risk score based on the TB/LTBI 27 transcript signature and TB/OD 44 transcript signature, we employed a publicly available microarray dataset: the whole blood expression dataset of Berry *et al.* [60]. The dataset was generated using Illumina HT12 V3 arrays comparing TB with LTBI and other infections. The dataset comprised HIV-uninfected TB, LTBI and OD individuals, as no TB HIV-infected adult cohort was publically available. For each testing dataset (accession series SA TB GSE19491 SA GSE19442, OD GSE22098), both quantile and robust spline normalisation were applied separately to the arrays and the data was base 2 log transformed. However the results were the same regardless of normalisation method.

We evaluated the performance of the TB/LTBI 27 transcript signature, using the TB and LTBI patients in the South African testing set (TB n=20, LTBI n=31). Out of the 27 transcripts of the signature, one had no corresponding transcript on the HT12 V3 array (ILMN_3247506 - *FCGR1C*). For the evaluation of the performance of the 44 TB/OD transcript signature, we used TB patients from the South African testing set (TB n=20) and OD patients that did not include patients with systemic lupus erythematosus as they were judged to be a rare disease in an African setting (n=82). Three of the transcripts in the TB/OD signature had no corresponding transcripts on the HT12 V3 array: ILMN_3287952 (LOC100133800), ILMN_3215715 (LOC389386) and ILMN_3308961 (*MIR1974*).

First we employed hierarchical clustering on the validation dataset using the transcripts present in the signatures. Distinct clusters of the samples were observed from the dendrograms for both TB vs. LTBI and TB vs. OD comparisons (Figure 22).

Subsequently we calculated the DRS and assessed its classificatory performance on the validation cohort (Figure 23). For the TB vs. LTBI comparison the DRS achieved a sensitivity of 95% $CI_{95\%}$(85-100) and a specificity of 94% $CI_{95\%}$(84-100). Regarding the discrimination between TB and OD, the 44 transcript DRS's sensitivity and specificity in the validation dataset were 100% $CI_{95\%}$ (100-100) and 96% $CI_{95\%}$(93-100) respectively (Table 11).

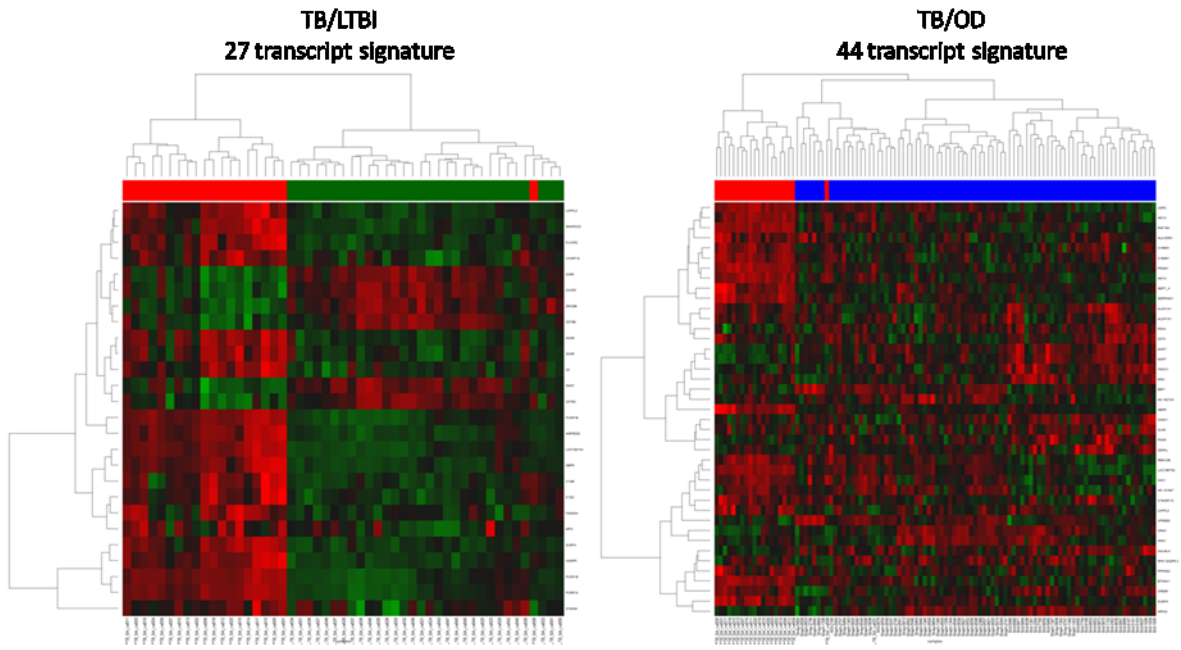


Figure 22: Heatmap showing clustering of the validation cohort.
The 27 TB/LTBI and the 44 TB/OD transcript signatures were used. The rows correspond to transcripts (red colour indicates up-regulation; green colour indicates down-regulation). Columns correspond to patients (red=TB; green=LTBI; blue=OD)

	Validation Dataset	
	TB vs. LTBI	**TB vs. OD**
AUC	**99%**	**100%***
(95% CI)	(97-100%)	(100-100%)
Sensitivity	**95%**	**100%**
(95% CI)	(85-100%)	(100-100%)
Specificity	**94%**	**96%**
(95% CI)	(84-100%)	(93-97%)

Table 11: Classification of the Berry *et al.* validation set using the Disease Risk Score based on the 27 TB/LTBI and the 44 TB/OD transcript signatures
The DRS based on the 27 TB/LTBI and 44 TB/OD transcript signatures was applied to the validation set from Berry *et al.* [60]. *99.94

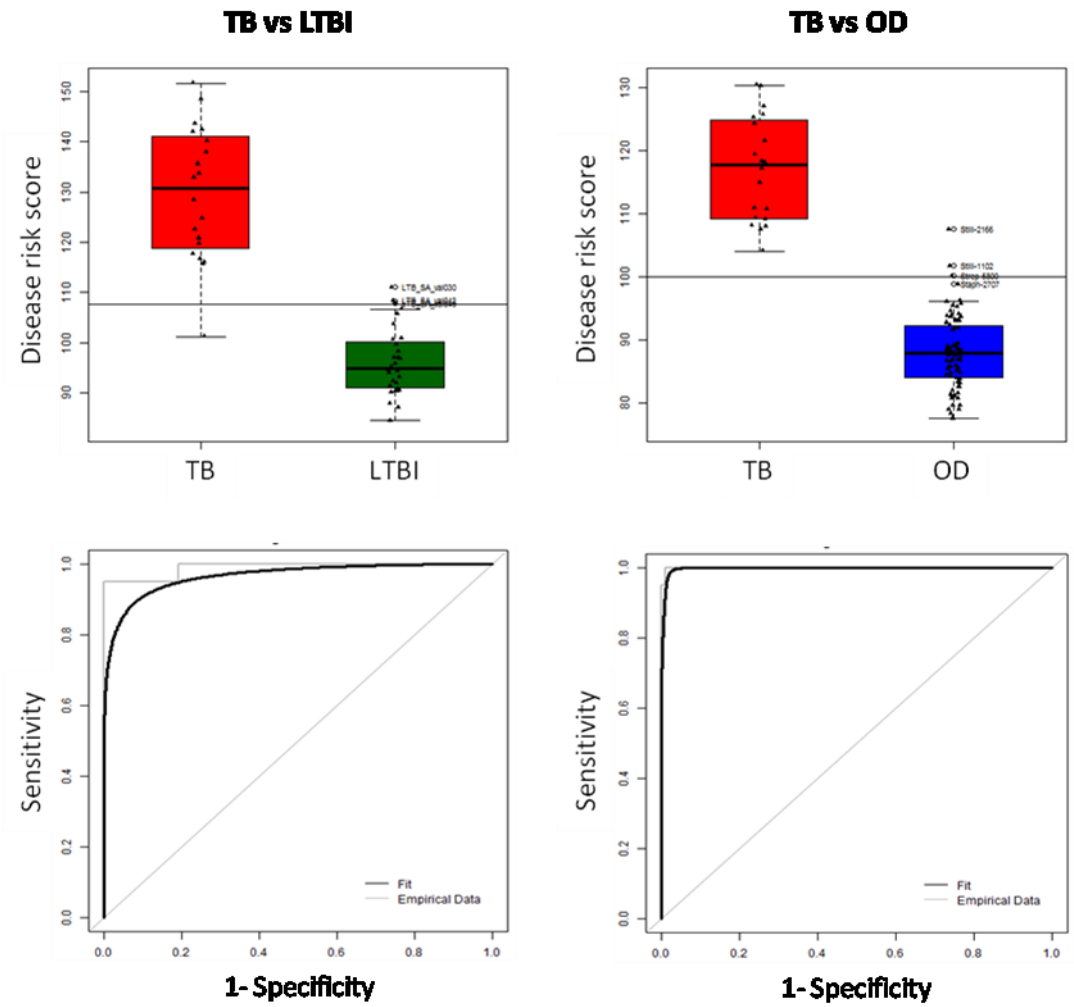


Figure 23: Classification of the Berry *et al.* validation set using the Disease Risk Score based on the 27 TB/LTBI and the 44 TB/OD transcript signatures.
Disease risk score (top panels) and receiver operating characteristic curves (bottom panels) based on the 27 TB/LTBI and the 44 TB/OD transcript signatures applied to the dataset from Berry *et al.* [60]. Sensitivity, specificity and AUC are reported in Table 11.

For a comparison between the performance of the DRS on our test set and the independent validation dataset, we present a summary statistics table (Table 12). The test set is separated by HIV status to provide a direct comparison of the performance with the validation dataset which included only HIV –uninfected individuals. We also calculated the positive and negative likelihood ratios to reflect the effect of the increase/decrease on the probability of disease if the test is positive/negative. The actual numbers of patients per group and the calls of DRS classification are also presented in Table 13.

	South Africa/Malawi test cohort			Validation dataset
	HIV+/-	**HIV-**	**HIV+**	**HIV-**
TB vs. LTBI				
Number of patients	n=76	n=38	n=38	n=51
AUC (95% CI)	**98%** (95-100%)	**100%** (100-100%)	**97%** (95-100%)	**99%** (97-100%)
Sensitivity (95% CI)	**95%** (87-100%)	**100%** (100-100%)	**94%** (83-100%)	**95%** (85-100%)
Specificity (95% CI)	**90%** (80-97%)	**100%** (100-100%)	**90%** (75-100%)	**94%** (84-100%)
Likelihood ratio+ (95% CI)	**9.23** (3.63-23.4)	**NA**	**9.44** (2.52-5.34)	**14.73** (3.84-56.47)
Likelihood ratio- (95% CI)	**0.06** (0.02-0.23)	**0**	**0.06** (0.01-0.42)	**0.05** (0.01-0.36)
TB vs. OD				
Number of patients	n=76	n=37	n=39	n=102
AUC (95% CI)	**95%** (89-99%)	**96%** (89-100%)	**94%** (83-100%)	**100*%** (100-100%)
Sensitivity (95% CI)	**93%** (83-100%)	**91%** (77-100%)	**95%** (85-100%)	**100%** (100-100%)
Specificity (95% CI)	**88%** (74-97%)	**93%** (80-100%)	**84%** (68-100%)	**96%** (93-100%)
Likelihood ratio+ (95% CI)	**7.89** (3.13-19.89)	**14.3** (2.15-95.12)	**6.02** (2.1-17.08)	**27.67** (9.11-84.03)
Likelihood ratio- (95% CI)	**0.08** (0.03-0.24)	**0.05** (0.01-0.35)	**0.06** (0.01-0.41)	**0**

Table 12: Classification of the South African/Malawi test set and the Berry *et al.* validation set using the Disease Risk Score based on the 27 TB/LTBI and the 44 TB/OD transcript signatures.

The DRS based on the 27 TB/LTBI and 44 TB/OD transcript signatures was applied to the South African/Malawi test and the validation set from Berry *et al.* [60].Metrics were calculated for the HIV-infected and -uninfected patients combined, as well as separately. *99.94

	South Africa/Malawi test cohort			Validation dataset
	HIV+/-	**HIV-**	**HIV+**	**HIV-**
TB vs. LTBI				
Number of patients	n_{ALL}=76; n_{TB}=37; n_{LTBI}=39	n_{ALL}=38; n_{TB}=19; n_{LTBI}=19	n_{ALL}=38; n_{TB}=18; n_{LTBI}=20	n_{ALL}=51; n_{TB}=20; n_{LTBI}=31
DRS positive calls/gold standard positive	[35/37]	[19/19]	[17/18]	[19/20]
DRS negative calls/gold standard negative	[35/39]	[19/19]	[18/20]	[29/31]
TB vs. OD				
Number of patients	n_{ALL}=76; n_{TB}=42; n_{OD}=34	n_{ALL}=37; n_{TB}=22; n_{OD}=15	n_{ALL}=39; n_{TB}=20; n_{OD}=19	n_{ALL}=102; n_{TB}=20; n_{OD}=83
DRS positive calls/gold standard positive	[39/42]	[20/22]	[19/20]	[20/20]
DRS negative calls/gold standard negative	[30/34]	[14/15]	[16/19]	[80/83]

Table 13: Number of patients per group and calls for the classification using Disease Risk Score per group.

In order to assess the potential predictive value of DRS in a cohort of patients undergoing investigation for persistent symptoms such as cough, fever and weight loss i.e. where TB was included in the differential diagnosis, we calculated the Positive and Negative Predictive Values (PPV, NPV). As the recruitment in Malawi was conducted in a prospective manner, we used the prevalence of TB in the Malawi cohort for the calculations. Out of the 437 patients with suspected TB, 254 ended up with a confirmed TB diagnosis, giving a prevalence of 58%. The DRS for TB vs. OD had a PPV of 92% $CI_{95\%}$(84-99) and a NPV of 90% $CI_{95\%}$(80-100) (Table 14). We also employed a scenario of 20% prevalence which may be more reflective of a general primary care setting in a high-burden African country [61]. NPV for TB vs. OD was higher: 98% $CI_{95\%}$(96-100), but PPV decreased: 66% $CI_{95\%}$(46-87). These metrics suggest that the DRS could be used as a rule-out test, with patients with positive DRS selected for further investigation.

	South Africa/Malawi test cohort						Validation dataset	
	HIV+/-		HIV-		HIV+		HIV-	
	PPV	**NPV**	**PPV**	**NPV**	**PPV**	**NPV**	**PPV**	**NPV**
TB vs. LTBI								
20% prevalence (95% CI)	**70%** (50-89%)	**99%** (97-100%)	**100%** (100-100%)	**100%** 100-100%)	**70%** (42-99%)	**98%** (96-100%)	**79%** (56-100%)	**99%** (96-100%)
58% prevalence (95% CI)	**93%** (86-99%)	**92%** (83-100%)	**100%** (100-100%)	**100%** (100-100%)	**93%** (84-100%)	**92%** (78-100%)	**95%** (89-100%)	**93%** (81-100%)
TB vs. OD								
20% prevalence (95% CI)	**66%** (46-87%)	**98%** (96-100%)	**77%** (44-100%)	**98%** (94-100%)	**60%** (35-85%)	**99%** (96-100%)	**87%** (75-100%)	**100%** (100-100%)
58% prevalence (95% CI)	**92%** (84-99%)	**90%** (80-100%)	**95%** (86-100%)	**88%** (74-100%)	**89%** (79-99%)	**92%** (79-100%)	**97%** (95-100%)	**100%** (100-100%)

Table 14: Positive and negative predictive values for the classification of the South African/Malawi test set and the Berry *et al.* validation set using Disease Risk Score based on the 27 TB/LTBI and the 44 TB/OD transcript signatures.
The DRS based on the 27 TB/LTBI and 44 TB/OD transcript signatures was applied to the South African/Malawi test set and the validation set from Berry *et al.* [60]. PPV and NPV were calculated for the HIV-infected and -uninfected patients combined, as well as separately for different scenarios of prevalence of the disease. *99.94

3.5.2 Performance of Berry *et al.* signatures on the SA/Malawi dataset

In contrast to the South African/Malawi TB cohort, recruitment in previous studies for TB host transcriptomics research has excluded HIV-infected patients, while other disease controls have not been recruited concurrently with TB cases or from the same population of patients undergoing investigation for TB. We wanted to estimate how these differences in biomarker study design may affect the discovery and performance of transcriptomic signatures, with potential for TB diagnostics. We compared the performance of the 27 transcript TB/LTBI signature and the 44 probe TB/OD signature with the performance of the signatures reported in Berry *et al.* [60]. The authors reported a 393-transcript signature for TB vs. LTBI and a 86-transcript signature for TB vs. OD that were derived using a Kruskal–Wallis test for the transcript comparisons across all study groups, with α=0.01 and after Benjamini–Hochberg adjustment for multiple testing (1%). These signatures were tested on the South African/Malawi cohort (Figure 24, Figure 25).

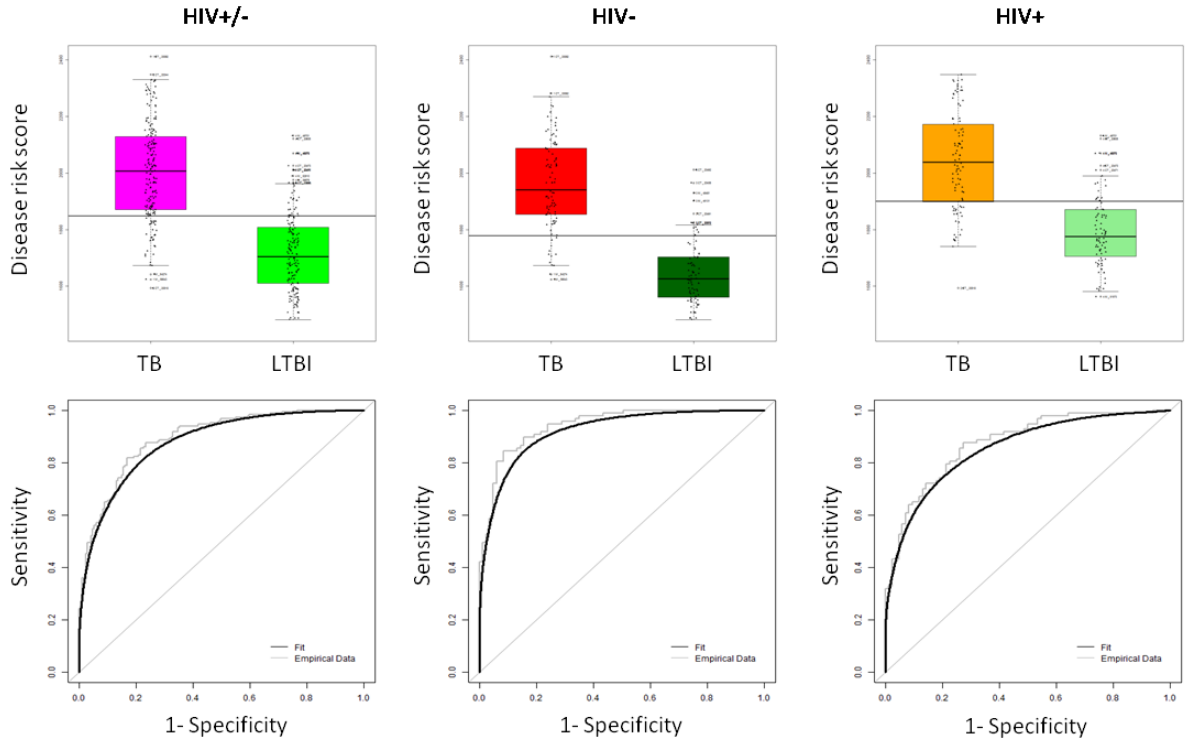


Figure 24: Classification of the combined South Africa and Malawi cohort using the transcript signature for TB vs. LTBI from Berry *et al.* [60].
DRS (top panels) & ROC curves (bottom panels) based on the TB/LTBI transcript signature from Berry *et al.* [60] applied to the combined training & test South African and Malawi cohort, calculated for the HIV-infected and -uninfected patients combined, as well as separately. Sensitivity, specificity and AUC are reported in Table 15.

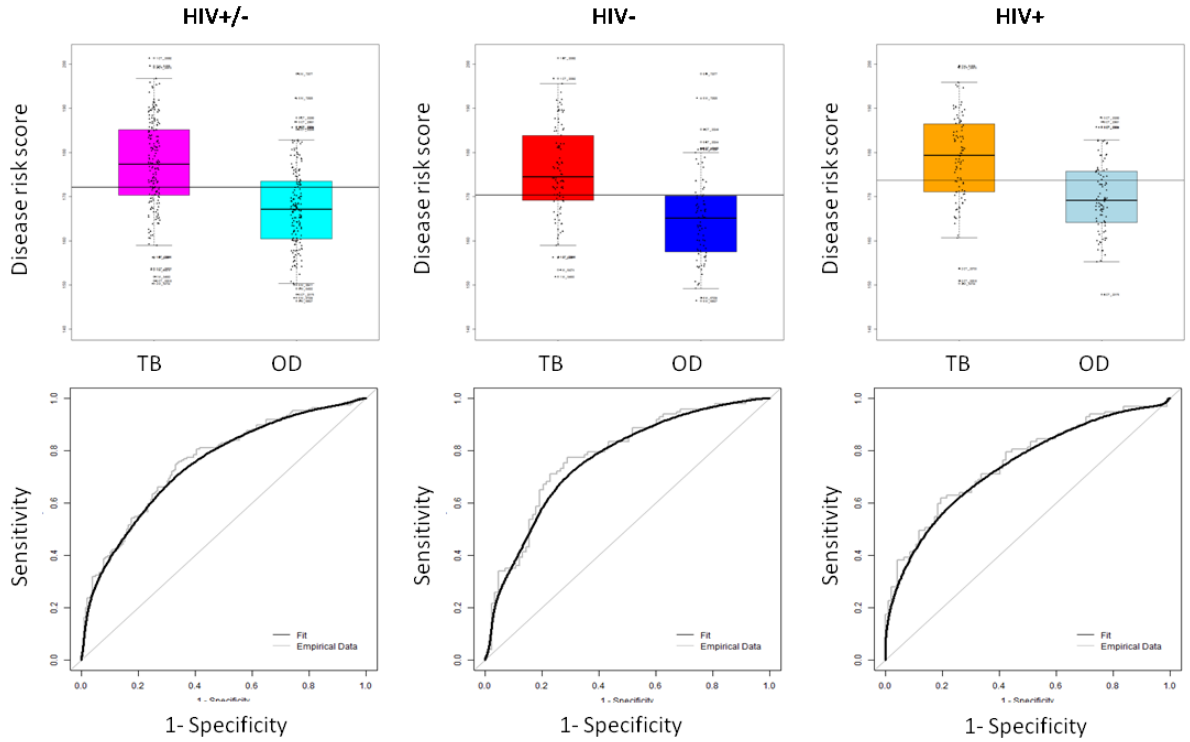


Figure 25: Classification of the combined South Africa and Malawi cohort using the transcript signature for TB vs. OD from Berry *et al.* [60].
DRS (top panels) & ROC curves (bottom panels) based on the TB/OD transcript signature from Berry *et al.* [60] applied to the combined training & test South African and Malawi cohort, calculated for the HIV infected and uninfected patients combined, as well as separately. Sensitivity, specificity and AUC are reported in Table 15.

For testing the performance of the reported 393 TB vs. LTBI signature and the 86 TB vs. OD signature on the African dataset, the disease risk score was calculated with these signatures as previously described, although 7 transcripts in the reported signatures were not present on the HT-12 V4 array (TB vs. LTBI 6 transcript, TB vs. OD 1 transcript).

The 393 TB/LTBI signature achieved a sensitivity of 82% $CI_{95\%}$(76-87) and a specificity of 81% $CI_{95\%}$(75-87) on the entire South Africa/Malawi TB/LTBI cohort regardless of HIV. In the HIV-uninfected individuals the 393 TB/LTBI signature achieved a sensitivity of 88% $CI_{95\%}$(80-94) and a specificity of 84% $CI_{95\%}$(76-92) while the performance on the HIV-infected group was lower with a sensitivity of 74% $CI_{95\%}$(65-82) and a specificity of 80% $CI_{95\%}$(71-87) respectively (Table 15). Furthermore, the Berry *et al.* TB/OD 86 transcript signature had a sensitivity of 68% $CI_{95\%}$(61-73) and a specificity of 70% $CI_{95\%}$(62-76) on the entire South Africa/Malawi TB/OD cohort regardless of HIV. In the HIV-uninfected group it achieved 71%

$CI_{95\%}$(62-80) sensitivity and 76% $CI_{95\%}$(67-84) specificity, while its classificatory performance dropped in the HIV-infected group (sensitivity 67% $CI_{95\%}$(58-75), specificity 69% $CI_{95\%}$(59-78)) (Table 15).

	South Africa/Malawi cohort (test & training)		
	HIV+/-	**HIV-**	**HIV+**
TB vs. LTBI signatures Berry *et al.*			
Number of patients	n=361	n=180	n=181
AUC (95% CI)	**89%** (86-92%)	**94%** (91-97%)	**88%** (82-92%)
Sensitivity (95% CI)	**82%** (76-87%)	**88%** (80-94%)	**74%** (65-82%)
Specificity (95% CI)	**81%** (75-87%)	**84%** (76-92%)	**80%** (71-87%)
TB vs. OD signatures Berry *et al.*			
Number of patients	n=369	n=180	n=189
AUC (95% CI)	**76%** (70-80	**78%** (70-84%)	**75%** (68-82%)
Sensitivity (95% CI)	**68%** (61-73%)	**71%** (62-80%)	**67%** (58-75%)
Specificity (95% CI)	**70%** (62-76%)	**76%** (67-84%)	**69%** (59-78%)

Table 15: Classification of the combined South African/Malawi training and test sets using the Berry *et al.* published signatures.
The DRS based on the TB/LTBI and TB/OD Berry *et al.* transcript signatures was applied to the South African/Malawi training and test sets. Metrics were calculated for the HIV-infected and -uninfected patients combined, as well as separately.

Thus the minimal transcript signatures and the DRS method show better classificatory performance in distinguishing TB from LTBI and OD (especially in the HIV-infected cohorts), than the much larger number of probes identified by Berry *et al.* [60].

In order to compare directly the differences of the performance of our signatures to the signatures presented by Berry *et al.* [60], we calculated the subtraction of the means of the measures of classification (namely the AUC, the sensitivity and the specificity) on our

test set along with their 95% confidence intervals using the following mathematical formulas:

$$(a, b) = \widehat{p_1} - \widehat{p_2} \pm z_{a/2} \times SE \quad (3.1)$$

$$SE = \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}} \quad (3.2)$$

Where:

α , b are the lower and upper confidence bounds

$\widehat{p_1}$, $\widehat{p_2}$ are the sample proportions of successes in population 1 and 2.

n_1 , n_2 are the sizes of each sample.

SE is the standard error

z = 1.96 for a 95% CI

The predictive performance of the 27 TB/LTBI transcript signature on the test set was higher in comparison to the 393 transcript (minus 6 transcripts) TB/LTBI Berry *et al.* signature (Table 16). Comparison of the statistical measures of performance of disease classification using the 44 TB/OD transcript signature and the Berry *et al.* 86 transcript (minus one transcript) signature suggested a better performance of the 44 TB/OD signature on the test set as well. The difference was actually larger for the TB vs. OD comparison, reflecting the need for real world control sets when deriving diagnostic signatures.

	South Africa/Malawi test cohort								
	HIV+/- (95% CI)			HIV- (95% CI)			HIV+ (95% CI)		
	Our signatures	Berry et al. signatures	Difference*	Our signatures	Berry et al. signatures	Difference*	Our signatures	Berry et al. signatures	Difference*
TB vs. LTBI comparison									
AUC	98%	88%	10%	100%	91%	9%	97%	89%	9%
(95% CI)	(95-100%)	(85-97%)	(2-18%)	(100-100%)	(88-100%)	(0-18%)	(92-100%)	(83-98%)	(-3-20%)
Sensitivity	95%	84%	11%	100%	90%	11%	94%	78%	17%
(95% CI)	(87-100%)	(73-95%)	(1-21%)	(100-100%)	(74-100%)	(1-20%)	(83-100%)	(61-94%)	(2-32%)
Specificity	90%	87%	3%	100%	79%	21%	90%	85%	5%
(95% CI)	(80-97%)	(77-97%)	(-8-13%)	(100-100%)	(58-95%)	(8-34%)	(75-100%)	(65-100%)	(-10-20%)
TB vs. Other Diseases comparison									
AUC	95%	73%	22%	96%	76%	20%	94%	72%	21%
(95% CI)	(89-99%)	(63-86%)	(10-33%)	(89-100%)	(62-91%)	(5-35%)	(82-100%)	(57-89%)	(5-37%)
Sensitivity	93%	74%	19%	91%	77%	14%	95%	70%	25%
(95% CI)	(83-100%)	(60-86%)	(8-31%)	(77-100%)	(59-96%)	(-3-30%)	(85-100%)	(50-90%)	(9-41%)
Specificity	88%	74%	15%	93%	67%	27%	84%	74%	11%
(95% CI)	(74-97%)	(59-88%)	(2-27%)	(80-100%)	(40-87%)	(9-44%)	(68-100%)	(53-90%)	(-7-28%)

Table 16: Classification of the South African/Malawi test set using the 27 TB/LTBI, 44 TB/OD and Berry *et al.* transcript signatures.
The DRS based on the 27 TB/LTBI and 44 TB/OD transcript signatures identified by elastic net and the TB/LTBI and TB/OD Berry *et al.* transcript signatures was applied to the South African/Malawi training and test sets. Metrics were calculated for the HIV-infected and -uninfected patients combined, as well as separately, and the differences along with their CIs are reported.

3.5.3 Performance of the DRS on the smear negative TB patients

A small proportion of the TB culture positive patients in the study were smear-negative; 31 TB patients had a definite negative smear status. The performance of the signatures by smear-status is important for estimating the potential utility of the signature in the clinical arena. We evaluated the performance of the signatures in the smear-negative sub-group of patients with TB, the majority of whom were HIV-infected (7 TB HIV-uninfected and 24 TB HIV-infected). In these patients, the DRS showed a sensitivity for detecting TB of 68%, $CI_{95\%}$ (52–84), when using the TB versus LTBI signature and a sensitivity of 90%, $CI_{95\%}$(81–100), with the TB/OD signature (Table 17). The results for the performance of the signatures are comparable to the results obtained in the larger HIV-infected cohort of smear-positive and -negative patients. As we used the same LTBI and OD patients from the test set, the specificity was unchanged (90%, $CI_{95\%}$ (80–97), for TB versus LTBI and 88%, $CI_{95\%}$ (74–97), for TB versus OD).

	South Africa/Malawi smear negative TB and controls from the test set	
	Sensitivity	**Specificity**
TB vs. LTBI comparison		
n_{ALL}=70; $n_{TB\ smear\ neg}$=31; n_{LTBI}=39		
Calls by DRS	[21/31]	[35/39]
Measures	68% (52-84)	90% (80-97)
TB vs. OD comparison		
n_{ALL}=65; $n_{TB\ smear\ neg}$=31; n_{OD}=34		
Calls by DRS	[28/31]	[30/34]
Measures	90% (81-100)	88% (74-97)

Table 17: Classification of the South Africa/Malawi smear negative TB patients and controls from the test set using Disease Risk Score based on the 27 TB/LTBI and the 44 TB/OD transcript signatures.
The DRS based on the 27 TB/LTBI and 44 TB/OD transcript signatures was applied to the smear negative TB cases from South Africa and Malawi as well as the respective controls from test set.

3.5.4 Smaller Signatures

Although elastic net suggested the smallest possible model able to achieve the best classification, we also explored whether a smaller number of transcripts could be used to distinguish TB from LTBI and from OD.

Instead of optimising the parameters of elastic net (which control the size of the selected model) via ten-fold cross-validation (CV), we used penalty settings that run the lasso method. Lasso constructs more parsimonious models than elastic net, by setting more β coefficients to zero. Then, within the cross-validation step of choosing s, we forced the penalty to be such that the error would remain within one standard deviation of the minimum error. This analysis resulted in a 21 transcript signature for the TB versus LTBI comparison (12 overlapping with the 27 transcript signature) and a 29 transcript signature for the TB versus OD comparison (14 overlapping with the 44 transcript signature). The sensitivity of the smaller models was 6% - 10% lower than the original models, while retaining almost the same specificity (Table 18).

After calculation of the DRS for every patient using the TB/LTBI signature, sensitivity and specificity were 89%, $CI_{95\%}$ (78-97) and 89%, $CI_{95\%}$(79-97) respectively. As for the TB versus OD comparison, DRS had a sensitivity of 83% $CI_{95\%}$ (69-93) and a specificity of 88% $CI_{95\%}$(75-97) respectively. In conclusion, smaller models have reduced sensitivity in comparison to the elastic net models.

	South Africa/Malawi test set	
	Sensitivity	**Specificity**
TB vs. LTBI comparison		
27 transcript signature	95% (87-100%)	90% (80-97%)
21 transcript signature	89% (78-97%)	89% (79-97%)
TB vs. OD comparison		
44 transcript signature	93% (83-100%)	88% (74-97%)
29 transcript signature	83% (69-93%)	88% (75-97%)

Table 18: Classification of the South African/Malawi test set using Disease Risk Score based on the elastic net and lasso transcript signatures.
The DRS based on the 27 TB/LTBI and 44 TB/OD transcript signatures identified by elastic net, as well as the 21 TB/LTBI and 29 TB/OD transcript signatures identified by lasso was applied to the South African/Malawi test set.

3.5.5 Missing Rate

We also explored the option of allowing the test to classify some patients as having "inconclusive diagnosis" with the thought that any patients falling into this category can be sent for further testing, as the result is not considered definitive. As expected, the introduction of a missing rate impacts on the error rate of the classification as well.

This was achieved by studying the effect of adjusting the threshold for the DRS in assigning individual patients to TB or LTBI/OD. By accepting a percentage of patients as "non-classifiable", the majority of patients under investigation were accurately assigned. These "non-classifiable" patients could then be selected for more detailed investigation or could be followed-up a later (Figure 26). This strategy seems to be better applicable to TB vs. LTBI patients.

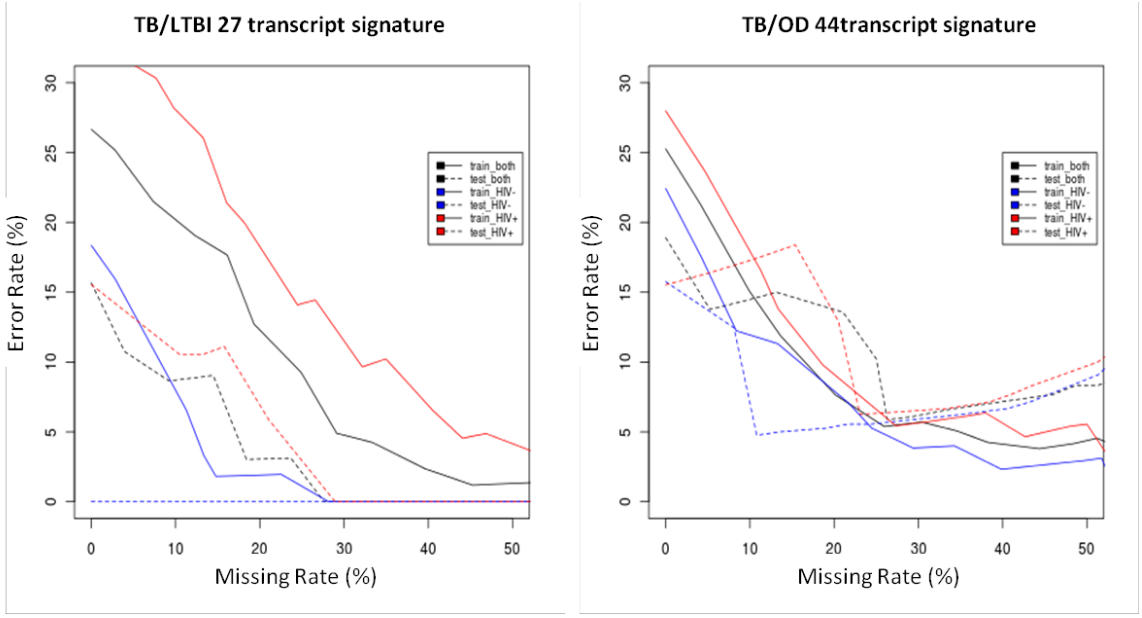


Figure 26: The error rate of classification in relation to the percentage of unclassified samples (missing rate).
The solid lines correspond to the training set while the dotted lines correspond to the test set. Black colour depicts the HIV-infected and -uninfected patients combined; blue colour the HIV-uninfected only and red colour the HIV-infected only.

3.5.6 TB vs. LTBI and OD simultaneously

As it would be advantageous from a clinical perspective to have a single signature that can distinguish patients with active TB from individuals with latent TB infection as well as patients with diseases other than TB, we assessed the performance of a signature in distinguishing TB from both LTBI and OD. A 53 transcript signature was identified by elastic net (Figure 27). The DRS based on the 53 transcript signature distinguished TB from both LTBI and OD with sensitivity of 91% and specificity of 82%, a lower performance than TB/LTBI and TB/OD signatures alone (Figure 28).

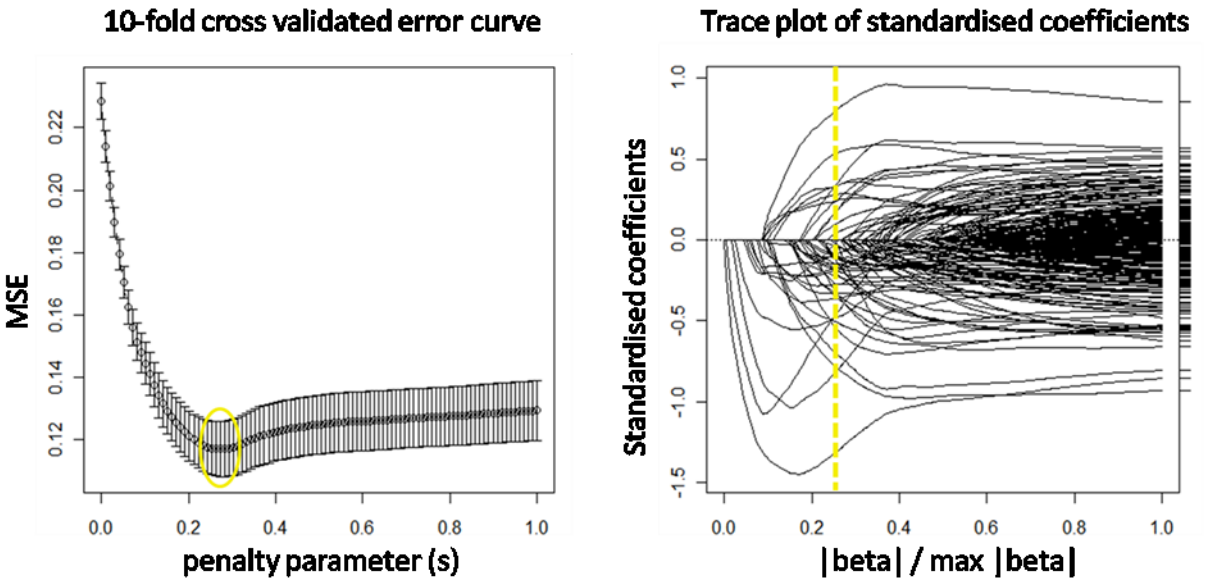


Figure 27: Mean Squared Prediction Error (MSE) plot and trace plot of the standardised coefficients for the TB vs. LTBI+OD comparison.

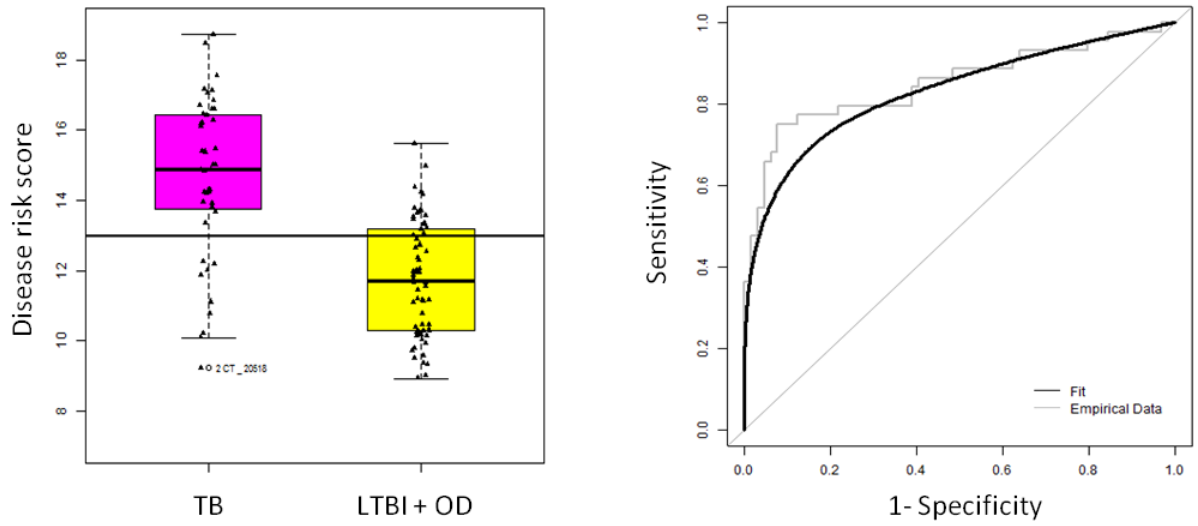


Figure 28: Classification of the TB vs. LTBI+OD test set using the Disease Risk Score based on the 53 TB/LTBI+OD transcript signature.

Disease risk score (left) and receiver operating characteristic curve (right) based on the 53 TB/LTBI+OD transcript signature applied to the combined HIV-infected and -uninfected South African/Malawi test cohort.

3.6 Adult TB Signatures

In this section we present the transcripts that comprise the transcriptomic signatures. We also present the direction of expression as well as the signatures' overlap and the overlap with the previously published signatures [60] (Table 19, Table 20). Additionally the biological role that the genes may play in the host response to TB is discussed.

Only three transcripts overlap between the TB/LTBI and TB/OD signatures, corresponding to genes: *DUSP3*, *GBP6* and *LHFPL2*. The TB/OD signature could be used to differentiate TB from OD in the clinical context of a sick patient, whereas the TB/LTBI signature may be used in contact screening where the problem is to distinguish active TB from latent infection. The combined TB vs. OD+LTBI signature which performed less favourably than the separate TB/OD signatures is also presented (Table 21).

Array ID	Gene Symbol	Probe ID	Direction*	Overlap with 393-transcript signature
130181	*ANKRD22*	ILMN_1799848	Up	√
5910019	*C1QB*	ILMN_1796409	Up	√
1440341	*C1QC*	ILMN_1785902	Up	
2650605	*C4ORF18*	ILMN_1672124	Up	
1340241	*C5*	ILMN_1746819	Up	√
5890470	*CCR6*	ILMN_1690907	Down	√
1780440	*CD79A*	ILMN_1659227	Down	
6450594	*CD79B*	ILMN_1710017	Down	
3890400	*CXCR5*	ILMN_2337928	Down	
6560156	*DUSP3*	ILMN_1797522	Up	√
2810373	*FAM20A*	ILMN_1812091	Up	
520086	*FCGR1A*	ILMN_2176063	Up	√
2710709	*FCGR1B*	ILMN_2261600	Up	√
6620209	*FCGR1B*	ILMN_2391051	Up	√
5570398	*FCGR1C*	ILMN_3247506	Up	
1510026	*FLVCR2*	ILMN_2204876	Up	√
70730	*GAS6*	ILMN_1779558	Up	√
4280632	*GAS6*	ILMN_1784749	Up	√
3780047	*GBP6*	ILMN_1756953	Up	√
1300139	*GNG7*	ILMN_1728107	Down	
360132	*LHFPL2*	ILMN_1747744	Up	√
5570039	*LOC728744*	ILMN_1654389	Up	√
3520601	*MPO*	ILMN_1705183	Up	
6060468	*S100A8*	ILMN_1729801	Up	
5910632	*SMARCD3*	ILMN_2309180	Up	√
2630195	*VAMP5*	ILMN_1809467	Up	√
2970397	*ZNF296*	ILMN_1693242	Down	

Table 19: The 27 transcript signature for distinguishing TB from latent TB infection
* in TB patients in relation to patients with latent TB infection.

Array ID	Gene Symbol	Probe ID	Direction*	Overlap with 86-transcript signature
1170332	*AAK1*	ILMN_1688755	Up	
1070477	*ALDH1A1*	ILMN_2096372	Up	
6510754	*ALDH1A1*	ILMN_1709348	Up	
5270753	*ARG1*	ILMN_1812281	Down	
3130600	*BTN3A1*	ILMN_1802708	Up	
6380681	*C19ORF12*	ILMN_1664920	Up	
1030433	*CALML4*	ILMN_1815707	Up	
540041	*CASC1*	ILMN_1708983	Up	
4560047	*CD74*	ILMN_1761464	Up	√
4070524	*CERKL*	ILMN_1801091	Up	
4290619	*CREB5*	ILMN_1728677	Up	
130086	*CYB561*	ILMN_1771179	Up	
840446	*CYB561*	ILMN_2378376	Up	
6560156	*DUSP3*	ILMN_1797522	Up	
1110592	*EBF1*	ILMN_1778681	Down	
3780047	*GBP6*	ILMN_1756953	Up	
2350114	*GJA9*	ILMN_1710161	Up	
1050360	*HLA-DPB1*	ILMN_1749070	Up	
270039	*HM13*	ILMN_1766269	Up	
7210110	*HM13*	ILMN_2236655	Up	
2000682	*HS.131087*	ILMN_1916292	Down	
5260161	*HS.162734*	ILMN_1893697	Down	
2340241	*IMPA2*	ILMN_2094061	Down	
360132	*LHFPL2*	ILMN_1747744	Up	
6760056	*LOC100133800*	ILMN_3287952	Up	
150224	*LOC196752*	ILMN_1803743	Up	
4570164	*LOC389386*	ILMN_3215715	Up	
5290100	*MAK*	ILMN_1803984	Down	
3120475	*MAP7*	ILMN_2216815	Down	
5820491	*MAP7*	ILMN_1712719	Down	
2260349	*MIR1974*	ILMN_3308961	Up	
2850315	*ORM1*	ILMN_1696584	Down	
3310504	*PDK4*	ILMN_1684982	Down	
1580437	*PGA5*	ILMN_1717572	Down	
380541	*PPPDE2*	ILMN_1737580	Up	
450379	*PRDM1*	ILMN_2294784	Up	
450132	*RBM12B*	ILMN_1805778	Up	
1690184	*RNF19A*	ILMN_1812327	Up	
3360553	*RP5-1022P6.2*	ILMN_1701111	Down	
4670458	*SEPT4*	ILMN_1776157	Up	
2030309	*SERPING1*	ILMN_1670305	Up	
6760471	*TMCC1*	ILMN_1677963	Down	
3840053	*UGP2*	ILMN_1671969	Up	
4640768	*VPREB3*	ILMN_1700147	Down	

Table 20: The 44 transcript signature for distinguishing TB from other diseases.
* in TB patients in relation to patients with OD

Array ID	Gene Symbol	Probe ID	Direction*
1070477	*ALDH1A1*	ILMN_2096372	Up
6510754	*ALDH1A1*	ILMN_1709348	Up
130181	*ANKRD22*	ILMN_1799848	Up
6330471	*BLK*	ILMN_1668277	Down
3130600	*BTN3A1*	ILMN_1802708	Up
1030433	*CALML4*	ILMN_1815707	Up
540041	*CASC1*	ILMN_1708983	Up
4560047	*CD74*	ILMN_1761464	Up
6380040	*COL9A2*	ILMN_1685122	Down
4290619	*CREB5*	ILMN_1728677	Up
1090497	*CREG1*	ILMN_1680624	Up
3890400	*CXCR5*	ILMN_2337928	Down
130086	*CYB561*	ILMN_1771179	Up
840446	*CYB561*	ILMN_2378376	Up
4540239	*DEFA1*	ILMN_2193213	Up
4860128	*DEFA1B*	ILMN_1725661	Up
7150170	*DEFA1B*	ILMN_2102721	Up
2970747	*DEFA3*	ILMN_2165289	Up
6560156	*DUSP3*	ILMN_1797522	Up
1110592	*EBF1*	ILMN_1778681	Down
6590646	*FAM26F*	ILMN_2066849	Up
520086	*FCGR1A*	ILMN_2176063	Up
6620209	*FCGR1B*	ILMN_2391051	Up
5570398	*FCGR1C*	ILMN_3247506	Up
5720180	*FZD2*	ILMN_1653711	Up
70730	*GAS6*	ILMN_1779558	Up
1510364	*GBP5*	ILMN_2114568	Up
3780047	*GBP6*	ILMN_1756953	Up
1300139	*GNG7*	ILMN_1728107	Down
3840753	*HEY1*	ILMN_1788203	Down
1940274	*IFI27L2*	ILMN_1740319	Up
870408	*IL15*	ILMN_1724181	Up
5260161	ILMN_1893697	ILMN_1893697	Down
2000682	ILMN_1916292	ILMN_1916292	Down
360132	*LHFPL2*	ILMN_1747744	Up
4570164	*LOC389386*	ILMN_3215715	Up
6400414	*LOC650546*	ILMN_1814812	Up
4670113	*LOC90925*	ILMN_1794927	Down
5820491	*MAP7*	ILMN_1712719	Down
3420259	*MIR21*	ILMN_3310840	Up
6760593	*OSBPL10*	ILMN_1669497	Down
1580437	*PGA5*	ILMN_1717572	Down
6380338	*POLB*	ILMN_1767894	Up
380541	*PPPDE2*	ILMN_1737580	Up
1660021	*RNU4ATAC*	ILMN_3240594	Up
4670458	*SEPT4*	ILMN_1776157	Up
2680136	*SIGLEC11*	ILMN_1674593	Up
6620161	*SPIB*	ILMN_2143314	Down
6760471	*TMCC1*	ILMN_1677963	Down
3840053	*UGP2*	ILMN_1671969	Up

2340682	*UHMK1*	ILMN_2096012	Up
4640768	*VPREB3*	ILMN_1700147	Down
3190113	*VPS13B*	ILMN_2268409	Up

Table 21: The 53 transcript signature for distinguishing TB from both latent TB and other diseases

* in TB patients in relation to patients with OD+LTBI

In the signatures, there are multiple transcripts for the same gene. This is a result of the "grouping effect" of the variables by elastic net. The inclusion of the transcripts from the same gene minimises further the MSE. However, lasso could be used as an alternative method that does not promote clusters of highly correlated variables in the final model.

The 27 transcripts for distinguishing active TB from latent TB infection correspond to 24 unique genes (18 up-regulated and 6 down-regulated in TB disease). Three of the up-regulated genes map to the complement pathway (complement component 1, q subcomponent B chain, *C1QB*; complement component 1, q subcomponent C chain *C1QC;* and complement component 5, *C5*) and have been reported elevated in the peripheral blood of active TB patients in comparison to LTBI and healthy controls [91, 137, 138]. The three *FCGR1* (Fc-gamma receptor 1) genes (that produce six different RNA transcripts by alternative splicing) encode for Fc receptors (FcR) that bind to immunoglobulin G (IgG) antibodies with high affinity. The FcRs are expressed by different immune effector cells and upon recognition of IgG, they trigger and regulate indicated inflammatory and cytotoxic cell responses, including phagocytosis, antigen presentation and T-cell activation that aid the elimination of pathogens [139]. *FCGR1s* are induced by interferon-γ and are mainly expressed by monocytes and macrophages. *FCGR1* genes have been previously identified as being among the most differentially expressed between patients with active TB and latently infected individuals in previous whole blood microarray studies. [60, 91] It has also been shown that *FCGR1A* has significantly higher levels of expression in participants with active TB than in those with LTBI regardless of HIV infection and ethnicity, using a dual-colour reverse transcriptase multiplex ligation-dependent probe amplification (RT-MLPA) assay [98]. *MPO* (myeloperoxidase) encodes for a protein abundantly expressed in neutrophils that is central to their microbicidal activity during the oxidative burst [140]. *GBP6* (guanylate-binding proteins 6) encodes for a member of a family of proteins that are involved in protection against intracellular pathogens and are induced by IFN-γ. Guanylate-

binding proteins contribute to interferon responses in a variety of organisms, including humans [141] and particularly in the IFNγ-mediated killing of intracellular pathogens. While it is known that *GBP6* has a role in IFNγ-mediated killing of intracellular bacteria, such as *Mycobacterium bovis* BCG, its mechanisms of action have not been fully determined [142]. *DUSP3* (dual specificity phosphatase 3) has been associated with murine susceptibility to bacterial infection and human sepsis [143].

As for the genes that are more highly expressed in LTBI than in active TB, *CCR6* (chemokine (C-C motif) receptor 6) is preferentially expressed by immature dendritic cells and memory T cells. It may also regulate the migration and recruitment of dendritic cells and T cells during inflammatory and immunological responses [144]. TB-specific CD4+ T cells with a characteristic chemokine expression signature (*CCR6*+ *CXCR3*+ *CCR4*-), are important for controlling TB infections [145] [146]. *CXCR5* (chemokine (C-X-C motif) receptor 5) has been identified as significantly differentially up-regulated in whole blood from individuals with LTBI in comparison to individuals with active TB regardless of HIV [98]. Also, *CXCR5*+ T helper cells play a protective role in the immune response against TB [147]. *CD79A* and *CD79B* encode for proteins that are required in co-operation for initiation of the signal transduction cascade activated by binding of antigen to the B-cell antigen receptor (BCR) complex which leads to internalisation of the complex and antigen presentation [144]. Out of the 24 unique genes, 15 were reported as being IFN-inducible (either type I or type II) in the *Interferome* database [148]. In order to explore the biological pathways associated with the signature genes, the signatures were analysed with QIAGEN's Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity). The top disease/function annotation reported was infectious disease, particularly bacterial infection, which included *C1QB*, *C5*, *CCR6*, *FCGR1A*, *FCGR1B*, *MPO*, *CD79*. As for canonical pathways analysis, the role of nuclear factor of activated T-cells (NFAT) in regulation of the immune response, the complement system, the B-cell development pathway and the role of pattern recognition receptors in recognition of bacteria and viruses pathway were among the top 5.

The 44 transcripts for distinguishing active TB from other diseases correspond to 36 unique genes (25 up-regulated and 11 down-regulated in TB disease). *CYB561* (cytochrome B561) that is up-regulated in active TB disease, encodes for a ferric reductase protein, harmonising with existing evidence that there is competition between the MTB and the host

for iron during infection [149, 150]. *HLA-DPB1* (HLA Class II Histocompatibility Antigen, DP Beta 1 Chain) is a protein coding gene that belongs to the HLA class II beta chain paralogues. The protein product binds to the HLA-DPA1 protein, forming a functional antigen-binding protein complex, the DPαβ heterodimer, that is expressed in antigen presenting cells (APCs: B lymphocytes, dendritic cells, macrophages). HLA plays a central role in the immune system by presenting foreign antigens to T-cells. Particularly, during *M.TB* infection APCs present *M.TB* antigens in association with MHC class II molecules to stimulate CD4+ T cells [151]. *ORM* (orosomucoid 1) gene encodes for an immune-related protein that has been identified as differentially expressed in human urine of patients with active TB disease [152]. Out of the 36 unique genes, 24 were reported as being IFN-inducible (either type I or type II) in the *Interferome* database [148]. In IPA, the top canonical pathway was antigen presentation, and the top diseases and function for the largest network containing 12 molecules, was cell-to-cell signalling and interaction and inflammatory response.

The 53 transcript signature for distinguishing TB from both latent TB and other diseases consists of 45 unique genes (33 up-regulated and 12 down-regulated in TB disease). It largely includes genes from the TB vs. LTBI signature and the TB vs. OD signature. It also includes a cluster of genes (*DEFA1* - defensin alpha 1, *DEFA1B* - defensin alpha 1B, *DEFA3* - defensin alpha 3) that encode for proteins that belong to the α-defensin family of antimicrobial peptides. Defensins are abundant in the granules of neutrophils and also found in the epithelia of mucosal surfaces and are involved in phagocyte-mediated host defence [144]. Out of the 45 unique genes, 25 were reported as being IFN-inducible (either type I or type II) in the *Interferome* database [148]. In IPA, the top canonical pathways were dendritic cell maturation, the role of macrophages, fibroblasts and endothelial cell in arthritis, and the role of nuclear factor of activated T-cells (NFAT) in regulation of the immune response. The top disease/function annotation was infectious disease where *DEFA1*, *DUSP3*, *FCGR1A*, *FCGR1B* and *IL15* were mapped, along with 6 other genes. The top upstream regulator of the genes in the signature was IFNγ, highlighting its central role in immune response during active TB disease [60].

In general, the blood transcriptional signatures are consistent with previous findings in the literature for pathway analysis of host TB signatures, showing up-regulation of IFN-inducible genes (both type I and type II) and inflammatory genes, and down-regulation of B

and T-cell genes [153]. The identification of TB specific genes and their good classificatory performance regardless of HIV infection suggests that they are not HIV status dependent.

3.7 Summary

The bioinformatics toolbox described in Chapter 2 was applied to a large prospective study to identify host biomarkers of TB from host transcriptional data. The study includes patients with active culture proven TB, latent TB infection and a broad range of infectious and malignant conditions that have similar clinical symptoms and findings with TB, regardless of HIV co-infection. Patients were recruited in South Africa and Malawi.

We investigated the hypothesis that host peripheral blood RNA expression would distinguish TB from other conditions prevalent in African populations and explored the use of a transcriptional signature as the basis for a diagnostic test. After pre-processing of the microarray data, the analysis identified a 27 transcript signature for discriminating TB from LTBI and a 44 transcript signature for discriminating TB from OD. The signatures were used to calculate a patient risk score able to classify the patients in the training and test sets with good performance. The signatures were also applied to a published dataset from South Africa which, unlike the discovery cohort, included only HIV-uninfected participants. The performance of the signature was further benchmarked in sub-groups of patients of particular interest and compared with the performance of signatures of smaller size.

The analysis highlights the ability of the signatures to discriminate active TB from latent TB infection and other diseases phenotypically similar to TB in adults regardless of HIV, location and smear status. Most previous attempts to identify biomarkers of TB have excluded HIV-infected individuals, have generally used control diseases from different populations, or have compared TB to other diseases not representative of the spectrum of diseases from which TB needs to be distinguished in African hospitals. In addition, the size of the signatures that had been reported would not allow potential translation into a test.

Although the signatures identified and DRS distinguished the majority of patients with TB from those with LTBI or OD, a proportion of patients were not correctly classified. Some false assignment by the test currently used as ‘gold standard’ is to be expected as

noted by post-mortem studies at which undiagnosed TB is confirmed. In addition, as all patients in the OD group presented in the clinics with symptoms similar to TB, it is possible that TB may have been misdiagnosed in a small proportion of OD patients despite the extensive clinical investigation used to assign each patient to each diagnostic group.

The 27/44 transcript signatures are derived and tested on an adult dataset. However, TB differs substantially in the pathophysiology, progression and clinical presentation in children which complicates the understanding as well as the diagnosis of the disease. In the next Chapter we aim to extend the findings of Chapter 3 by looking for biomarkers specific to paediatric TB disease.

Chapter 4

Identification of host gene expression biomarkers for TB in children

In Chapter 4 we present the analysis of a large, prospective study to identify host gene expression biomarkers of TB disease in children. The study included both HIV-infected and -uninfected children with LTBI, active TB and a wide spectrum of other diseases phenotypically similar to TB. The analysis was based on the pipeline described on Chapter 2. Minimal transcriptional signatures were identified and the patients' disease risk score was calculated. The performance of the signatures in the form of disease score were evaluated in an independent paediatric dataset. The accuracy of the disease risk score was also compared to a molecular diagnostic tool now establishing itself in clinical use: the Xpert® MTB/RIF. .

4.1 Paediatric Tuberculosis

Tuberculosis in children differs from adult disease in several aspects including clinical presentation, degree of infectiousness, diagnostic difficulty, progression of the disease, risk of disseminated disease and response to treatment. The majority of these can be attributed to the different immune responses in children compared to adults, and particularly to the immaturity of their immune system [154].

While most adult TB cases can be diagnosed by detection of AFB on sputum microscopy or culture, the majority of childhood TB cases are both smear and culture-negative. As children do not produce sputum or produce only small amounts, sample acquisition requires complex invasive procedures (i.e. hospital admission to obtain early morning gastric lavage fluids, or saline-induced sputum aspirates) [155]. Even then, microbiological confirmation is the exception rather than the rule due to the paucibacillary nature of the disease and frequent extra-pulmonary presentations [69, 70, 156]. Only 10-15% of children that are commenced on TB treatment are actually smear positive and 30% are culture positive [63]. Therefore, diagnosis and treatment initiation are often based solely on clinical suspicion.

The most common presentations of TB in symptomatic children are fever, respiratory symptoms, weight loss and anorexia. However, as these symptoms are common to a range of other conditions, such as pneumonia, viral and bacterial infections, malnutrition and HIV, clinical diagnosis is unreliable [138]. Clinical scoring systems that use evidence of exposure to an infectious adult, chest x-rays, TST and IGRA positivity are employed to aid the diagnostic procedure. However, chest x-rays findings are not specific and TST and IGRA cannot differentiate active disease from latent infection, which poses an impediment in high burden settings [157, 158]. Furthermore, children with TB, particularly if HIV-infected or malnourished, may be both TST and IGRA negative [159-162]. In addition to the above, NAAT have low sensitivity in diagnosing TB in smear negative samples [163].

The difficulty in microbiological and immunological diagnosis of childhood TB results in a high proportion of children with TB being diagnosed late, and in some cases the diagnosis is only made post-mortem [164]. Conversely, as definitions of latent infection and

disease for children are less definitive than for adults, many patients are treated unnecessarily due to the difficulty of excluding TB from the diagnosis.

Due to the lack of reliable diagnostics and underreporting, the true global burden of childhood TB remains unknown. Improved methods to diagnose childhood TB are therefore urgently needed, particularly in countries of sub-Saharan Africa where the burden of TB and HIV co-infection is highest [69, 156, 165-167]. Thus, we investigated the possibility that RNA expression signatures in whole blood might distinguish TB from latent TB infection and from other diseases, and explored the potential role of RNA expression signatures as a diagnostic tool.

4.2 Cohorts

A three site paediatric recruitment process was undertaken in Kenya, South Africa and Malawi. The study was conducted in these three sub-Saharan African countries as they have high burden of tuberculosis as well as HIV co-infection (Figure 29), while also having different patterns of malaria and other endemic diseases. The samples from children evaluated for suspected tuberculosis in hospitals in South Africa and Malawi were grouped together to form a discovery cohort. The performance of the signatures was assessed on the third cohort, comprising children evaluated for suspected tuberculosis in Kenya.

The incidence of new TB cases in Kenya is 268 per 100,000, higher than Malawi (156 per 100,000) but lower than South Africa (860 per 100,000). Out of the total number of TB cases, 38% tested positive for HIV [61]. Despite >98% infant BCG vaccination coverage in South Africa, 90%.5 in Malawi and 96%.7 in Kenya, there is a high incidence of TB in children in all three countries [168, 169]. In the areas where children were recruited for this study (Kilifi District and Mombasa in Coast Province), a mixed rural and urban population is based, malaria and malnutrition are endemic, while invasive bacterial and helminth infections are also common. Epidemiological data for TB in South Africa and Malawi were presented in Chapter 3.

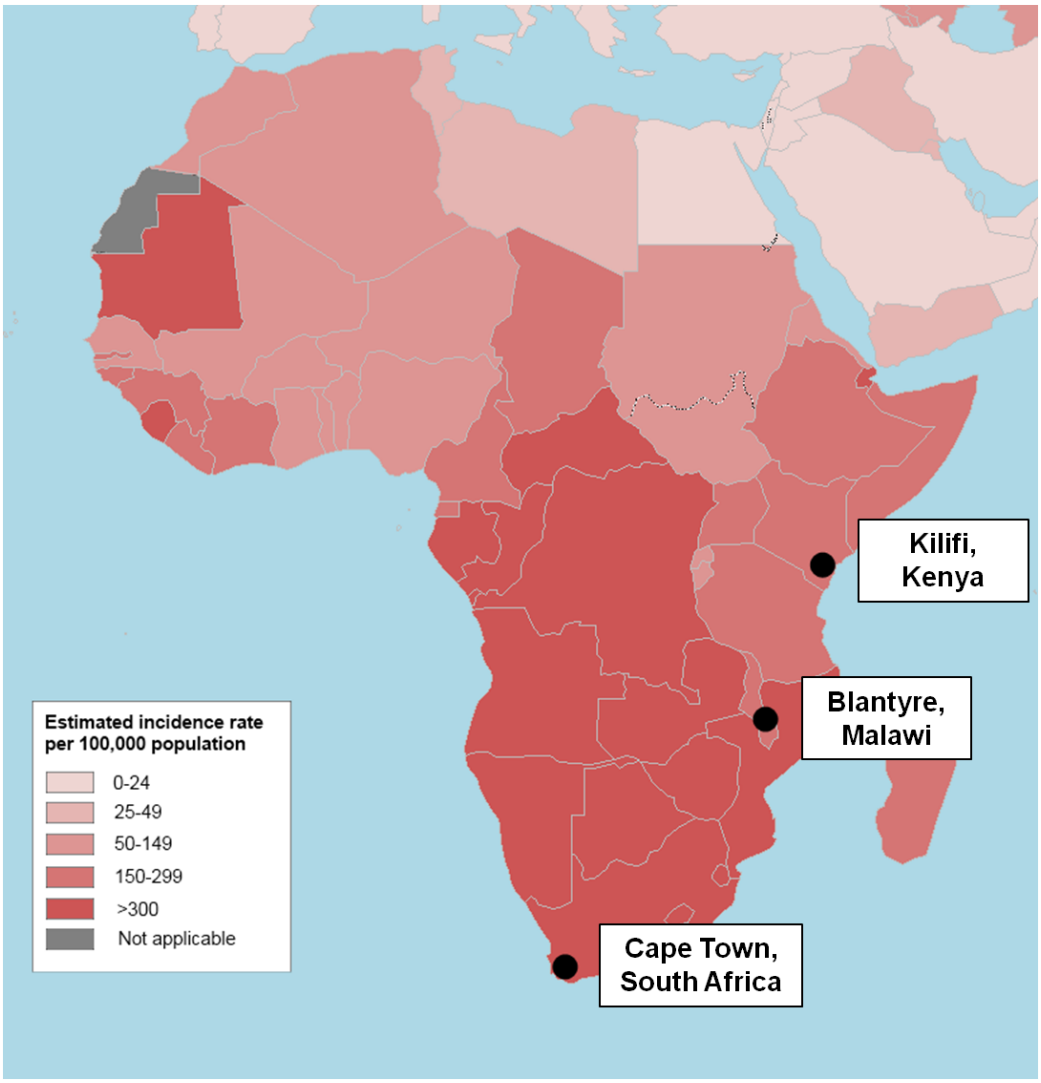


Figure 29: The incidence rate of TB in Africa and recruitment sites for the paediatric study. The map was plotted using WHO estimates of TB incidence by country imported in the R package *rworldmap* [170].

4.2.1 Recruitment

In South Africa and Malawi, children were referred to the study team for investigation if the clinician looking after the child suspected a diagnosis of TB. In Kenya, all children admitted to hospital were formally screened for features of TB disease (fever, weight loss and cough for more than two weeks). Children with one or more of these features plus those referred for outpatient investigation for TB and children <5 years old who were identified as household TB contacts of smear positive pulmonary TB, were eligible for inclusion in the study.

The investigations included chest radiography, measurement of the CRP levels, a serologic test or polymerase-chain-reaction (PCR) assay for HIV, and a TST, with or without an IGRA. Two spontaneous or induced sputum samples and a specimen of tissue or cerebrospinal fluid (if clinically indicated) were examined for AFB and cultured for mycobacteria. The Xpert® MTB/RIF real-time PCR assay (a test for presence of *M. tuberculosis* and resistance to the antibiotic rifampicin) was performed on respiratory samples in the Kenyan cohort. Bacterial cultures, histological examination of tissue-biopsy specimens, and analysis of blood films for the presence of malaria were performed as clinically indicated. Clinical follow-up was undertaken at 3 months to confirm that children with LTBI remained free of TB and other diseases and to determine whether there had been a response to treatment in children with confirmed or suspected tuberculosis.

After screening and evaluating 1356 children in South Africa and Malawi for symptoms of TB, 157 patients from South Africa and 189 patients from Malawi were included in the study. In Kenya, a total of 1599 children presenting to clinics in Kenya met the inclusion criteria for the study, and 1471 were evaluated for TB. Patients were assigned into study diagnostic groups based on examination results at enrolment and follow-up, as well as their HIV status.

4.2.2 Diagnostic Process

Following the diagnostic work-up, patients were assigned to groups (Figure 30),using the following definitions shown in Table 22 :

Group	Definition
Culture-confirmed TB	Clinical condition consistent with TB and microbiological evidence (*M. tuberculosis* was cultured from sputum or tissue samples).
Culture-negative TB	Negative mycobacterial culture in a child with clinical and radiologic features that prompted empirical treatment for tuberculosis. Culture-negative tuberculosis was further categorised as a case in which tuberculosis was highly probable, probable, or possible. (Figure 30)
Latent TB case (LTBI)	Clinically assessed as healthy on presentation and follow-up, contact with a person who had a positive smear for TB and had positive results on both the TST and the IGRA (discovery cohort) and had positive results on either the TST or the IGRA (validation cohort).
Other disease case (OD)	Children were classified as having diseases other than TB if they received a definitive alternative diagnosis or had no clinical deterioration on follow-up in the absence of TB therapy.

Table 22: Definitions for "gold standard" classification of patient samples.

Since a positive result on an IGRA in the group of patients with diseases other than TB might indicate either LTBI or primary tuberculosis that had resolved without treatment, we excluded patients in the discovery cohort who had a positive result on an IGRA. In the Kenyan validation cohort, patients who had diseases other than TB were included regardless of whether an IGRA result was positive or negative. Also, TST positivity was defined as ≥10mm in HIV-uninfected and ≥5mm in HIV-infected children.

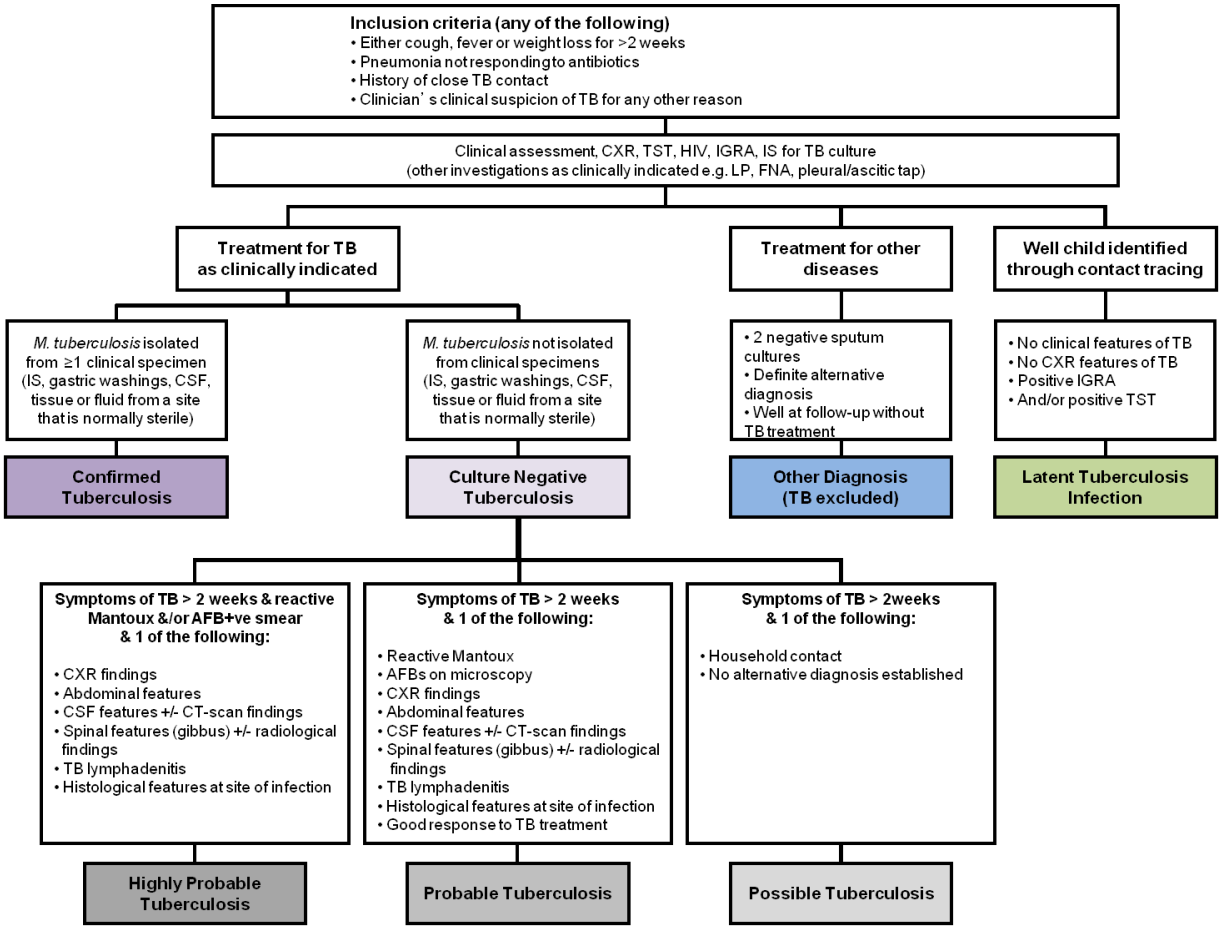


Figure 30: Diagnostic algorithm for patient classification into the culture confirmed TB, culture negative TB, other diseases and LTBI groups.
IGRA positive OD patients were excluded from Cape Town and Malawi but were included in the Kenyan cohort. IS: induced sputum, CXR: chest X-ray

Of the 346 children included in the study from South Africa and Malawi, 114 had culture-confirmed tuberculosis, 175 had diseases other than TB, and 57 had LTBI. The discovery cohort included only those children with TB that was confirmed using culture; children for whom the diagnosis of TB could not be confidently established or ruled out were excluded (Figure 31).

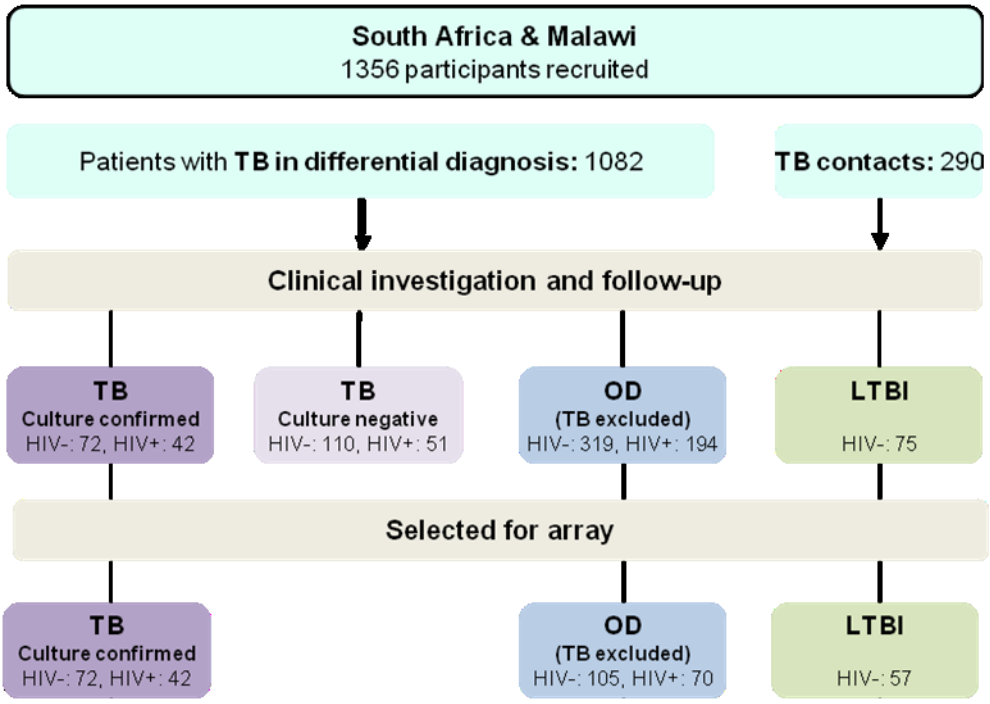


Figure 31: Samples randomly selected for microarrays from South Africa and Malawi.
Out of the 1356 participants recruited in the study from South Africa and Malawi, 144 patients with cultured confirmed TB (72 HIV-; 42 HIV+), 175 patients with OD (105 HIV-; 70 HIV+) and 57 LTBI individuals (all HIV-) were selected for array analysis.

From Kenya, 157 children were included in a nested case–control RNA expression study that contained the culture-confirmed TB cases and all subgroups in the group with culture-negative TB: children in whom TB was highly probable, probable, or possible. Microarray analysis was conducted for 35 patients with culture-confirmed TB and 14 patients with LTBI, 44 patients with culture-negative TB (8 patients in whom TB was highly probable, 19 in whom it was probable, and 17 in whom it was possible) and 64 randomly selected patients from the group with diseases other than TB (55 with negative IGRA results and 9 with positive IGRA results) (Figure 32).

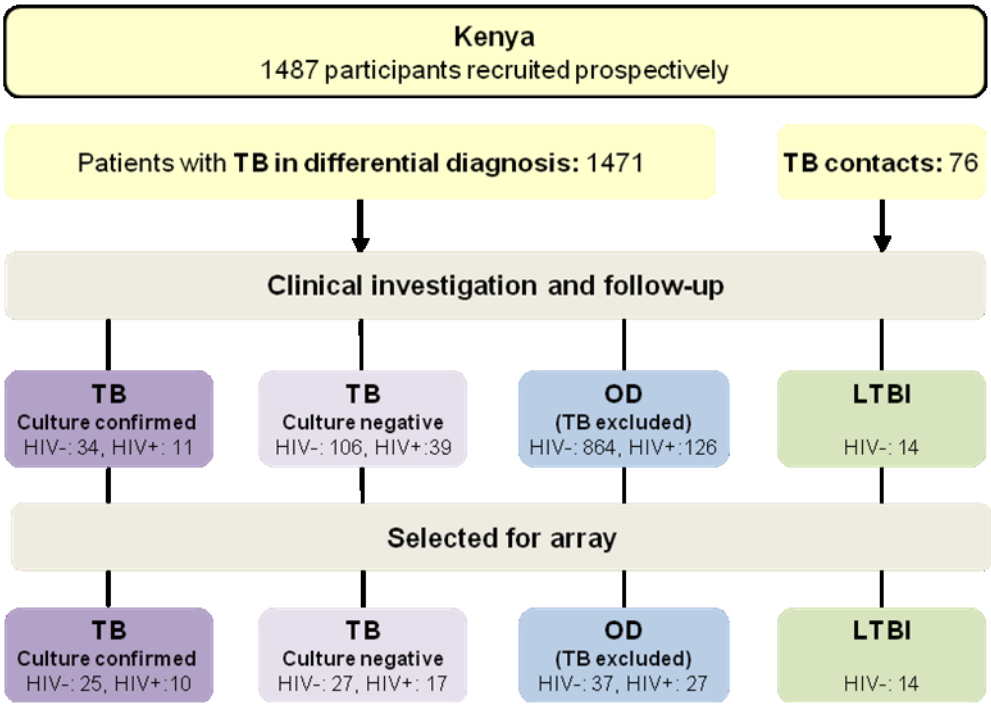


Figure 32: Samples selected for microarrays from the Kenyan cohort.
Out of the 1487 participants recruited in the study from Kenya, 35 patients with cultured confirmed TB (25 HIV-; 10 HIV+), 44 patients with cultured negative TB (27 HIV-; 17 HIV+), 54 patients with OD (37 HIV-; 27 HIV+) and 14 LTBI individuals (all HIV-) were selected for array analysis.

In order to assess if arrayed samples were representative of the population, clinical features of the patients included in the microarray study were compared to the features of patients who were not included, using Fisher's exact 2-sided test. The groups were very similar with two exceptions: in the group with probable tuberculosis, a history of close contact with a person who had tuberculosis was more common among patients included in the microarray study (Table 23), and in the group with diseases other than tuberculosis, weight loss and pleural effusion were more common among patients who were included (Table 24).

	Highly probable TB					Probable TB					Possible TB				
	Included (n=8)		Excluded (n=7)		p value	Included (n=19)		Excluded (n=45)		p value	Included (n=17)		Excluded (n=49)		p value
Tuberculosis exposure															
Close TB contact history	5	(63%)	2	(29%)	0.31	11	(58%)	10	(22%)	0.009	5	(29%)	8	(16%)	0.29
TST positive	8	(100%)	6	(86%)	0.47	6	(39%)	13	(29%)	1	1	(6%)	1	(2%)	0.45
Tuberculosis exposure	8	(100%)	6	(86%)	0.47	13	(68%)	20	(44%)	0.10	5	(29%)	8	(16%)	0.29
Clinical symptoms/signs of TB															
Persistent cough >2 weeks	5	(63%)	5	(71%)	1	11	(58%)	30	(67%)	0.57	12	(71%)	30	(61%)	0.57
Persistent fever >2 weeks	4	(50%)	4	(57%)	1	6	(32%)	27	(60%)	0.06	8	(47%)	33	(67%)	0.16
Night sweats >2 weeks	3	(38%)	3	(43%)	1	6	(32%)	7	(16%)	0.18	3	(18%)	10	(20%)	1
Weight loss/failure to thrive	7	(88%)	6	(86%)	1	12	(63%)	28	(62%)	1	12	(71%)	34	(69%)	1
CXR features of TB															
Airway compression	1	(13%)	0	(0%)	1	1	(5%)	0	(5%)	0.30	0	(0%)	0	(0%)	-
Lymphadenopathy	6	(75%)	6	(86%)	1	7	(37%)	19	(42%)	0.78	1	(6%)	4	(8%)	1
Airspace shadowing	3	(38%)	2	(29%)	1	9	(47%)	18	(40%)	0.59	3	(18%)	15	(31%)	0.36
Miliary/nodular shadowing	0	(0%)	0	(0%)	-	1	(5%)	1	(2%)	0.51	1	(6%)	1	(2%)	0.45
Pleural effusion	0	(0%)	1	(14%)	0.47	2	(11%)	2	(4%)	0.58	0	(0%)	1	(2%)	1
Cavities	0	(0%)	0	(0%)	-	2	(11%)	3	(7%)	0.63	0	(0%)	1	(2%)	1
Calcified Ghon focus	0	(0%)	0	(0%)	-	0	(0%)	0	(0%)	-	0	(0%)	0	(0%)	-
Vertebral spondylitis	0	(0%)	0	(0%)	-	0	(0%)	0	(0%)	-	0	(0%)	0	(0%)	-

Table 23: Comparison of culture negative TB cases included in & excluded from the microarray analysis of the cohort from Kenya.

	Included (n=64)		Excluded (n=935)		p value
Tuberculosis exposure					
Close TB contact history	6	(9%)	143	(15%)	0.28
TST positive	4	(6%)	118	(13%)	0.17
Tuberculosis exposure	16	(25%)	297	(32%)	0.27
Clinical symptoms/signs of TB					
Persistent cough >2 weeks	23	(36%)	433	(47%)	0.12
Persistent fever >2 weeks	21	(33%)	379	(41%)	0.24
Night sweats >2 weeks	10	(16%)	143	(15%)	1
Weight loss or failure to thrive	50	(78%)	514	(56%)	<0.001
CXR features of TB					
Airway compression	0	(0%)	2	(0.2%)	1
Lymphadenopathy	5	(8%)	67	(7%)	0.80
Airspace shadowing	17	(27%)	156	(17%)	0.06
Miliary/nodular shadowing	0	(0%)	0	(0%)	-
Pleural effusion	4	(6%)	17	(2%)	0.04
Cavities	0	(0%)	1	(0.1%)	1
Calcified Ghon focus	0	(0%)	0	(0%)	-
Vertebral spondylitis	0	(0%)	0	(0%)	-

Table 24: Comparison of OD cases included in & excluded from the microarray analysis of the cohort from Kenya.

Amongst the 175 patients of the group with diseases other than TB from South Africa and Malawi and the 64 patients from Kenya, pneumonia was the most common definite diagnosis. As shown in Table 25, some diseases were more prevalent in some locations or amongst HIV-infected individuals than others, therefore the groupings were taken into account in the partitioning of the discovery cohort (South Africa & Malawi) into the training and test set for the signatures identification, in order to control for overfitting.

Group	**HIV-infected**			**HIV-uninfected**			
Location	**SA**	**Malawi**	**Kenya**	**SA**	**Malawi**	**Kenya**	**Total**
Pneumonia	24	15	15	30	17	18	119
Bronchiectasis/chronic lung disease	2	7	-	-	1	-	10
Lymphocytic Interstitial Pneumonitis	-	2	-	-	-	-	2
Upper respiratory tract infection	-	-	-	11	-	-	11
Inflammatory bone&joint diseases	-	1	-	-	8	2	11
Bacterial soft tissue infection	-	5	-	-	16	-	21
Gastroenteritis	2	-	-	5	-	-	7
Infection at ≥2 sites	2	-	-	3	-	-	5
Sepsis without a focus	-	-	4	-	-	1	5
Kaposi Sarcoma	-	7	-	-	-	-	7
Other malignancy	-	-	-	-	5	1	6
Malaria + severe malnutrition	-	-	-	-	1	3	4
Primary diagnosis: sev. malnutrition[a]	-	1	4	-	1	3	9
Other[b]	-	2	3	1	6	1	13

Table 25: Major clinical diagnoses in the 'Other Diseases' groups from each study site.

[a] These are children who had a primary diagnosis of severe malnutrition; many other children in the OD group also had severe malnutrition in addition to the diagnoses listed.

[b] Includes cryptococcal meningitis (2); empyema (3); septicaemia (2); congenital spinal abnormalities (2); abscess + bacteremia (1); bacterial meningitis (1); severe anemia (1); and one child with severe malnutrition and a febrile illness of uncertain aetiology which resolved without TB treatment.

4.3 Power calculation

In order to evaluate the suitability of the sample cohorts to detect effects that would be appropriate for biomarker use in this study, a power calculation was used. As the study has two diagnostic arms (TB vs. LTBI and TB vs. OD) and two datasets (discovery and validation), different calculations were implemented. For an anticipated sensitivity of 90% and a specificity of 90% in the discovery set for the TB vs. LTBI comparison, a sample size of 72 TB patients and 57 LTBI patients will provide 95% confidence intervals spanning 7.2% from the point estimate of sensitivity and 8.5% from the point estimate of specificity. In an 80-20 split of the cohort, the 95% confidence intervals will span 17% for sensitivity and 20% for specificity for the test set. In the validation set, for the TB vs. LTBI comparison, a sample size of 35 TB patients and 14 LTBI patients will provide 95% confidence intervals spanning 10% from the point estimate of sensitivity and 17% from the point estimate of specificity.

For the TB vs. OD comparison, for an anticipated sensitivity of 90% and a specificity of 90% in the discovery set, a sample size of 72 TB patients and 175 OD patients will provide 95% confidence intervals spanning 7.2% from the point estimate of sensitivity and 4.5% from the point estimate of specificity. In an 80-20 split of the cohort the 95% confidence intervals will span 17% for sensitivity and 10% for specificity for the test set. In the validation set, a sample size of 35 TB patients and 64 LTBI patients will provide 95% confidence intervals spanning 10% from the point estimate of sensitivity and 7.6% from the point estimate of specificity. As for the sensitivity for the culture negative TB patients, a sample size of 44 patients will provide 95% confidence intervals spanning 9.3% from the point estimate of sensitivity. The adjusted Wald method was used for the calculations.

4.4 Results

4.4.1 Pre-processing

First we pre-processed and performed quality control analysis on the dataset from South Africa and Malawi by following the steps of the pipeline described in Chapter 2. First, the data underwent background adjustment and normalisation, and then quality control was performed in order to inspect for irrelevant groupings and to check for possible assignment errors. PCA analysis was employed and the first principal components were examined, based on all transcripts on all samples (Figure 33). Inspection revealed that the primary clustering was based on disease state (TB, LTBI, and OD) rather than geographical location or HIV status. Two samples indicated by the arrows (a TB/HIV+ and an OD/HIV+ case from Malawi) were removed from the analysis. Confidence ellipses calculated for the population mean are shown below (0.999 inner circle, 0.9999 outer circle).

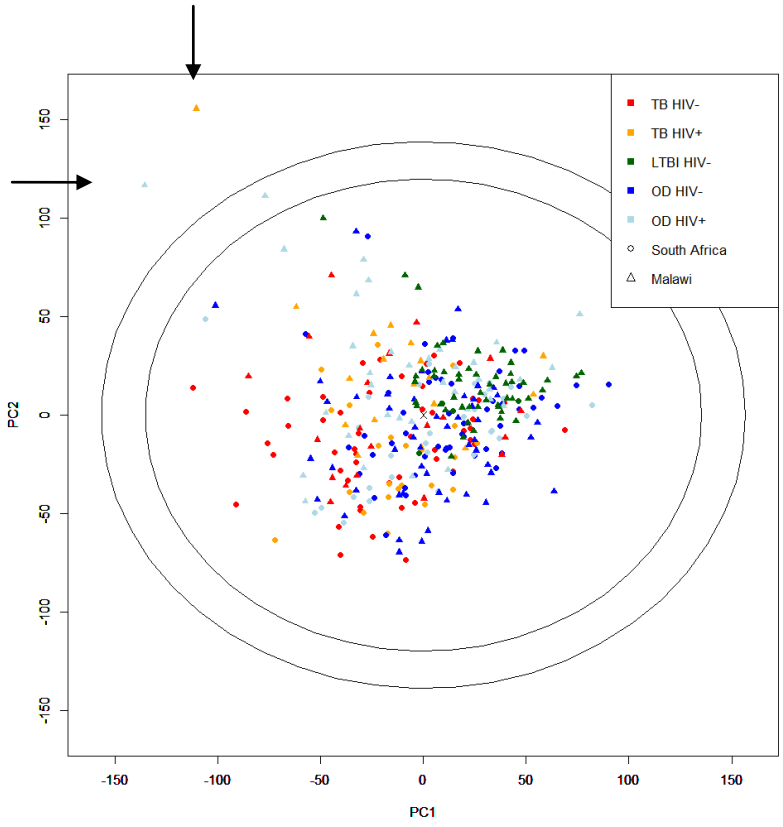


Figure 33: Principal components analysis (PCA) plot of the microarray samples from South Africa and Malawi.
The plot depicts PCA1 & PCA2 and is based on all transcripts and all samples after background adjustment and normalisation. The samples highlighted were removed from the analysis. Rings represent levels of confidence (0.9 inner circle; 0.9999 outer circle).

The R^2 values for the principal components, which reflect the importance of each component, are: R^2 (PC1) = 0.132; R^2 (PC2) = 0.103; R^2 (PC3) = 0.063; R^2 (PC4) = 0.041; R^2 (PC5) = 0.032; R^2 (PC6) = 0.024. Correlation analysis indicated that none of the variables is correlated with the components, with a correlation coefficient over 0.5. However, there is evidence of correlation for TB disease and TB infection with PC1 (-0.32 and 0.30) (Figure 33). Also, HIV status (positive) was correlated with PC4 (0.39) and location (South Africa) was correlated with PC5 (-0.48).

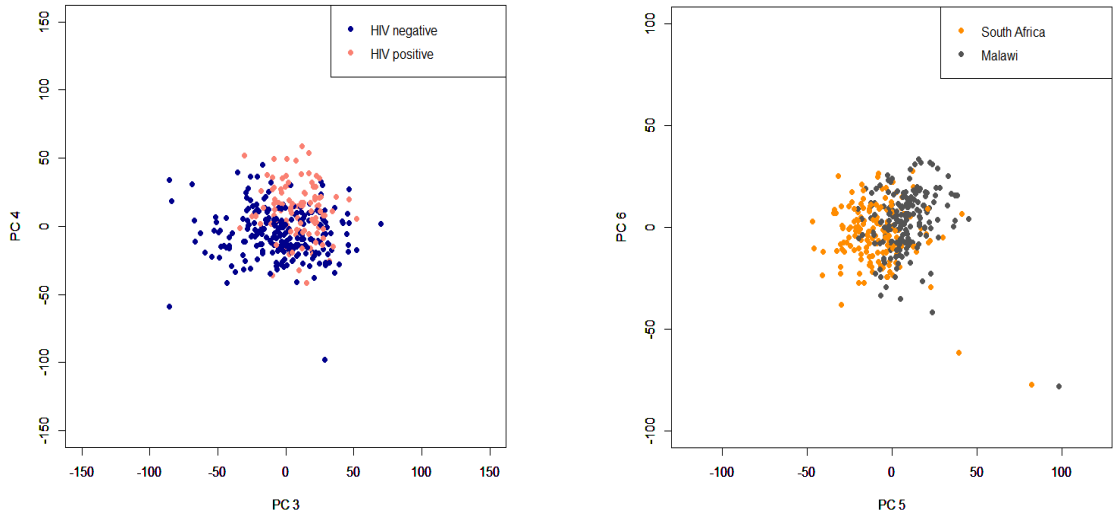


Figure 34: Principal components correlated with location and HIV status.
Principal components 3, 4 and 5, 6 are shown. On the left plot colour coding depicts the location (South Africa or Malawi), and on the right plot colour coding depicts the HIV status (negative or positive).

4.4.2 Application of Adult Signatures

As in Chapter 3 where we described the derivation of gene expression signatures for TB in adults from South Africa and Malawi, we first wanted to assess the performance of these signatures in the paediatric South African and Malawi cohort. We calculated the disease risk score of the patients in the paediatric cohort recruited at the same sites as the adults and calculated the classification metrics based on the 27 transcript signature for TB vs. LTBI (Table 26, Figure 35) and the 44 transcript signature for TB vs. OD (Table 26, Figure 36).

	Paediatric Dataset	
	TB vs. LTBI adult signature	**TB vs. OD adult signature**
AUC (95% CI)	**89%** (85-94%)	**81%** (76-86%)
Sensitivity (95% CI)	**80%** (72-87%)	**75%** (66-83%)
Specificity (95% CI)	**80%** (69-89%)	**74%** (67-81%)

Table 26: Classification of the South Africa/Malawi paediatric dataset using the Disease Risk Score based on the 27 TB/LTBI and 44 TB/OD transcript signatures.
The DRS based on the 27 TB/LTBI and 44 TB/OD transcript signatures was applied to the South African/Malawi paediatric dataset.

The adult TB vs. LTBI signature when applied to the paediatric cohort achieved a sensitivity and a specificity of 80% $CI_{95\%}$(72-87) and 80% $CI_{95\%}$(69-89) respectively. For the TB vs. OD comparison the DRS classified the paediatric samples with sensitivity of 75% $CI_{95\%}$(66-83) and specificity of 74% $CI_{95\%}$(67-81) (Table 26).

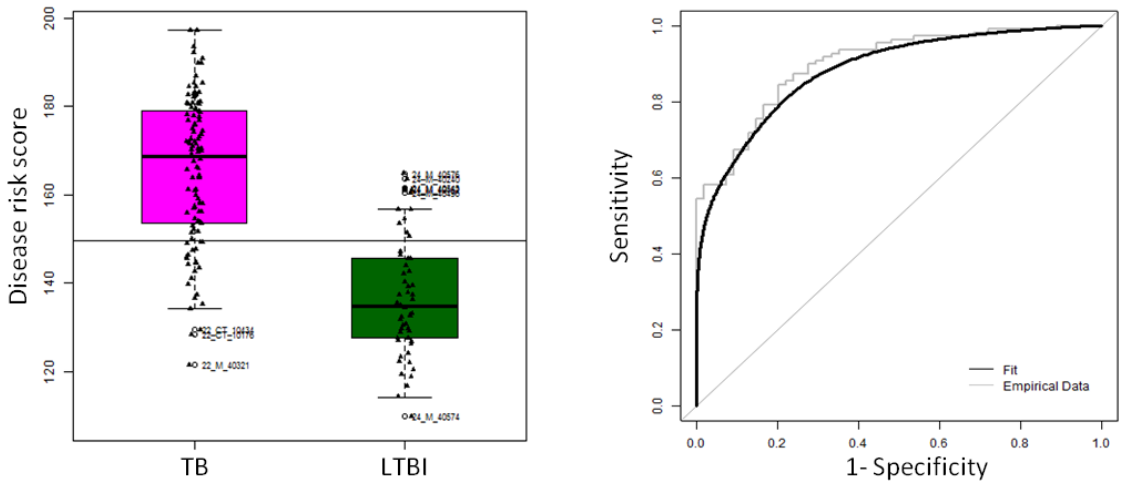


Figure 35: Classification of the South African/Malawi paediatric dataset using the Disease Risk Score based on the 27 TB/LTBI adult transcript signature.
Disease risk score (left) and receiver operating characteristic curve (right) based on the 27 TB/LTBI adult transcript signature applied to the patients of the South African/Malawi paediatric dataset. Sensitivity, specificity and AUC are reported in Table 26.

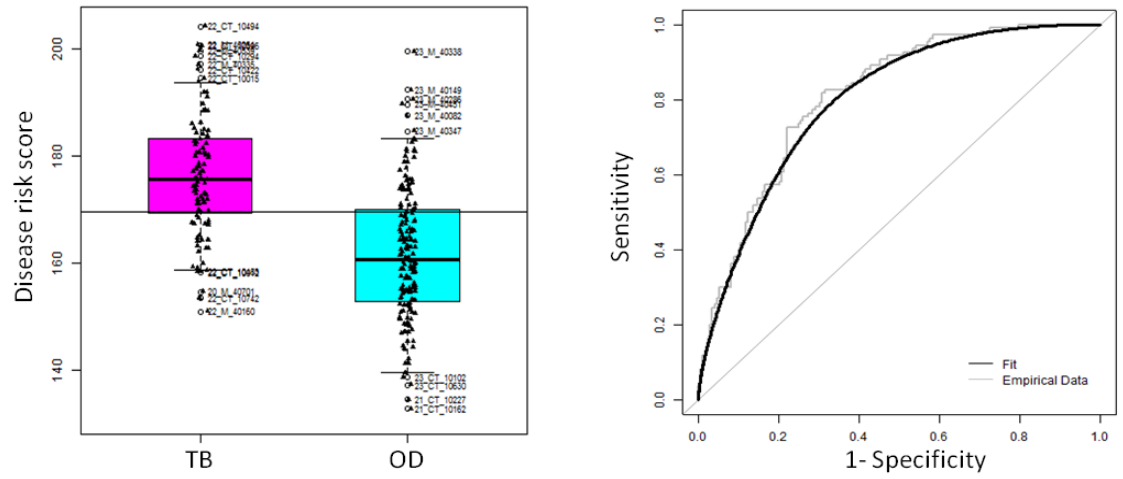


Figure 36: Classification of the South African/Malawi paediatric dataset using the Disease Risk Score based on the 44 TB/OD adult transcript signature.
Disease risk score (left) and receiver operating characteristic curve (right) based on the 44 TB/OD adult transcript signature applied to the patients of the South African/Malawi paediatric dataset. Sensitivity, specificity and AUC are reported in Table 26.

The classification achieved using the adult signatures has reduced performance in comparison to the performance in the adult test and validation datasets. This fact, in addition to the differences of the TB disease and infection between children and adults and

the effect that they are expected to have on gene expression in whole blood, led us to follow an analysis plan that would use the paediatric dataset to discover paediatric specific signatures and to validate them as well.

4.4.3 Discovery of Paediatric Signatures

In order to detect TB specific transcriptional signatures, accounting for the differences of the TB epidemic in the two countries, the microarray datasets from South Africa and Malawi were combined, pre-processed and analysed together. The Kenyan cohort was not included in the initial analysis and derivation of the signatures, but was subsequently used to benchmark the signatures (Figure 37).

The analysis was performed as previously described in Chapter 2. The combined dataset from South Africa and Malawi, the discovery dataset, was partitioned into training and test sets, comprising 80% and 20% of the samples respectively. As different prominent subclasses of different diseases were present in the other diseases group, we maintained a balanced representation in both sets. For the two comparisons of interest, TB vs. LTBI and TB vs. OD, we performed differential expression analysis for the transcripts on the training set. Then, the elastic net was run using the preselected sets of transcripts that were differentially expressed between the comparator groups.

Once the transcripts were selected, we calculated the disease risk score using the expression values of the transcripts selected by elastic net of the microarray samples from the Kenyan validation cohort.

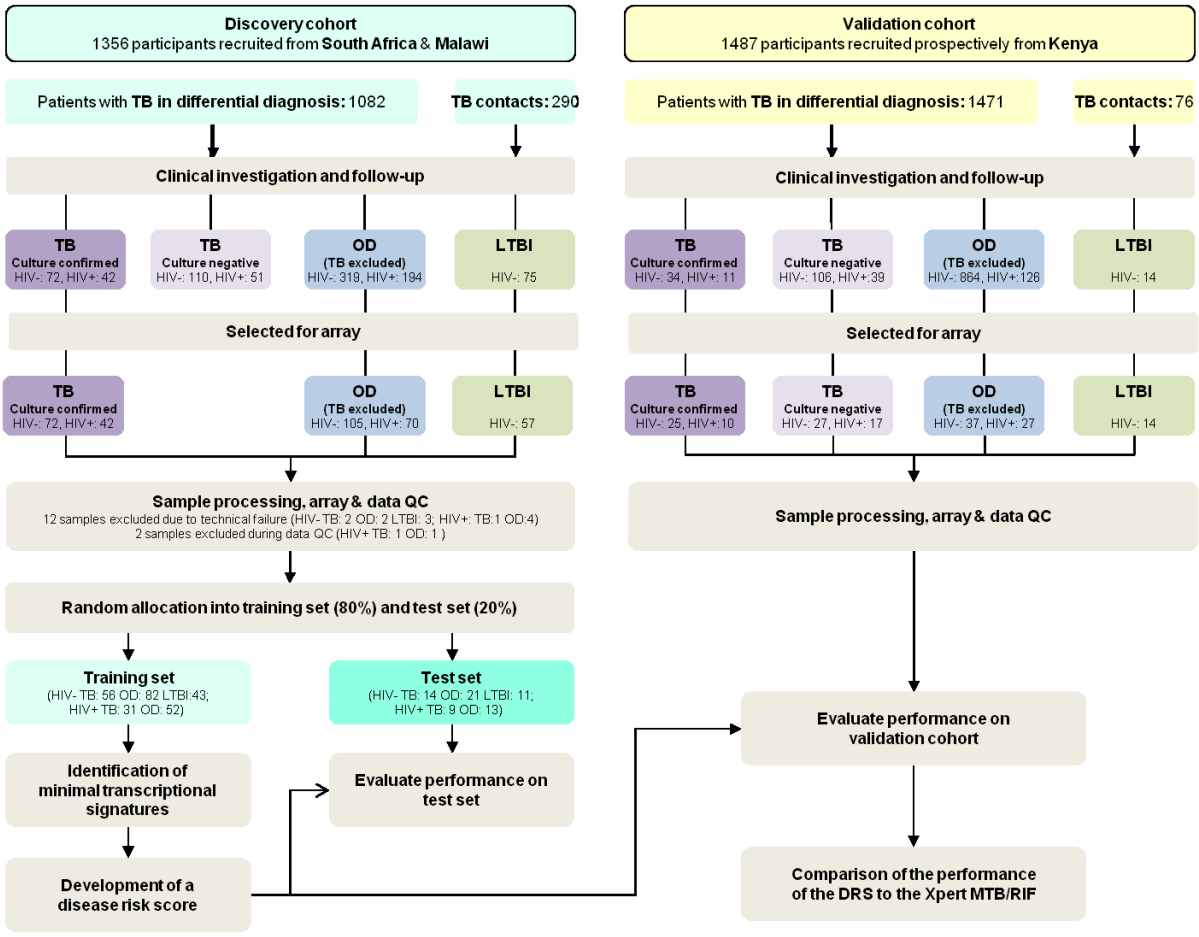


Figure 37: Microarray analysis plan for the paediatric microarray dataset.
The combined dataset from South Africa and Malawi was partitioned into training and test sets employed for signature discovery and in-sample testing. The signatures and Disease Risk Score were then validated using the microarray samples from the Kenyan cohort.

4.4.3.1 TB vs. LTBI

The training set for the TB vs. LTBI comparison included 86 culture confirmed TB cases and 43 LTBI individuals, comprising 80% of samples from the discovery cohort. After fitting a linear model accounting for HIV status and location we identified 3434 transcripts that were differentially expressed between TB and LTBI. These transcripts were used as input variables to run the elastic net. The 10-fold cross-validation for the elastic net parameters indicated that a model with λ_2=0.6 and s=0.27 would produce the minimum mean squared error, corresponding to a 42-transcript model (Figure 38).

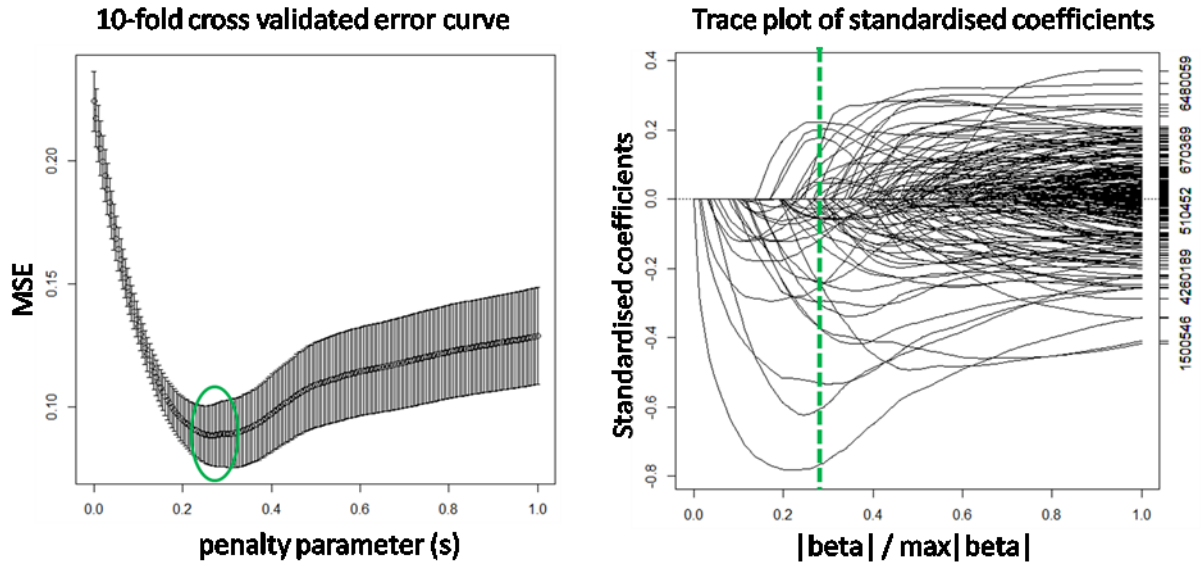


Figure 38: Mean Squared Prediction Error (MSE) plot and trace plot of the standardised coefficients for the TB vs. LTBI paediatric comparison (s=0.27).

After the signature identification by elastic net, the disease risk score using the 42 TB/LTBI transcript signature was calculated for the patients of the test (Figure 39), as well as the training set (Figure 40).

The test set included 23 culture confirmed TB cases (14 HIV-uninfected and 9 HIV-infected) and 11 LTBI individuals, while the training set included 86 culture confirmed TB cases (56 HIV-uninfected and 30 HIV-infected) and 43 LTBI individuals. As progression from latency to active disease in children is usually very quick, particularly for HIV infected children, no LTBI HIV-infected child was included in the study. AUC, sensitivity and specificity of the classifier were calculated to assess the DRS's classificatory performance (Table 27). The DRS based on the 42-transcript signature distinguished TB from LTBI in the patients of the test set with sensitivity and specificity of 96% $CI_{95\%}$ (87-100) and 91% $CI_{95\%}$ (73-100) respectively. Similar metrics were obtained for the classification of the training set samples.

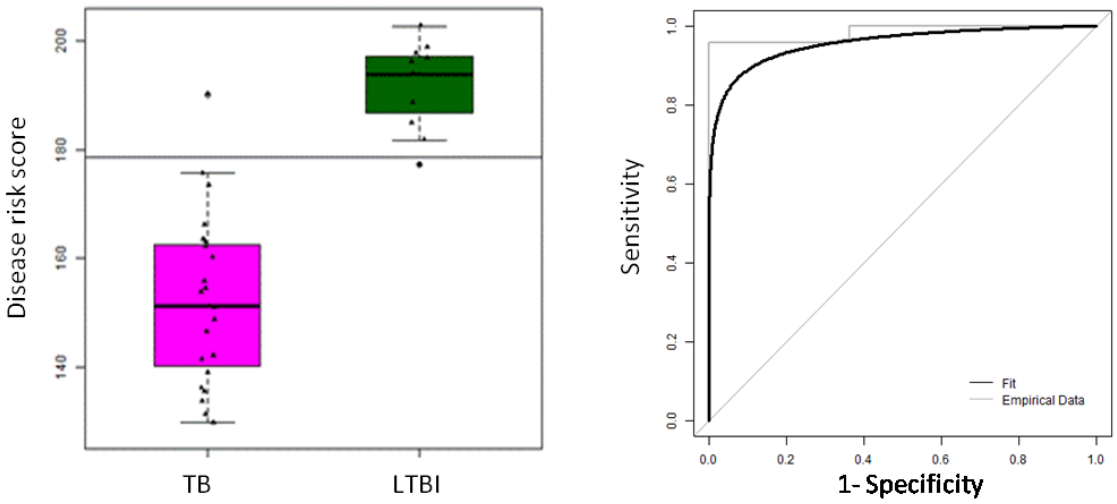


Figure 39: Classification of the South African/Malawi paediatric test set (20%) using the Disease Risk Score based on the 42 TB/LTBI transcript signature.

Disease risk score (left) and receiver operating characteristic curve (right) based on the 42 TB/LTBI transcript signature applied to the patients of the South African/Malawi paediatric test set (20%). Sensitivity, specificity and AUC are reported in Table 27 below.

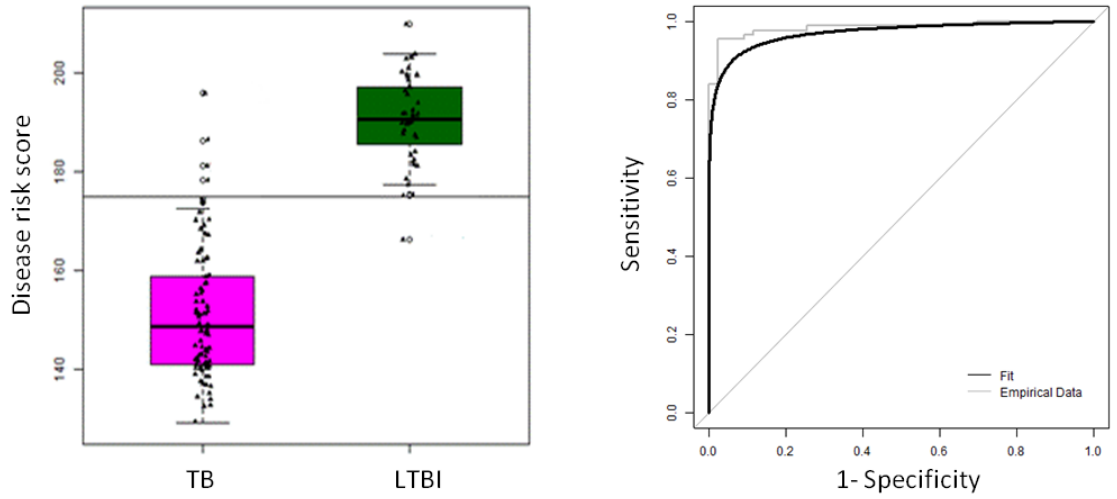


Figure 40: Classification of the South African/Malawi paediatric training set (80%) using the Disease Risk Score based on the 42 TB/LTBI transcript signature.

Disease risk score (left) and receiver operating characteristic curve (right) based on the 42 TB/LTBI transcript signature applied to the patients of the South African/Malawi paediatric training set (80%). Sensitivity, specificity and AUC are reported in Table 27 below.

	South Africa/Malawi cohort	
	Training set	**Test set**
AUC (95% CI)	**98.4%** (96.3 - 99.8%)	**98.4%** (94.5 - 100.0%)
Sensitivity (95% CI)	**95.4%** (90.8 - 98.9%)	**95.7%** (87.0 - 100.0%)
Specificity (95% CI)	**97.7%** (93.0- 100.0%)	**90.9%** (72.7 - 100.0%)

Table 27: Classification of the South Africa/Malawi paediatric training and test sets using the Disease Risk Score based on the 42 TB/LTBI transcript signature.
The DRS based on the 42 TB/LTBI transcript signature was applied to the South African/Malawi training and test sets.

4.4.3.2 TB vs. OD

In the training set, for the comparison of TB against other diseases, 86 culture-confirmed TB cases and 133 patients with other diseases were included, comprising 80% of samples from the discovery cohort. After fitting a linear model accounting for HIV status and location, we identified 409 transcripts that were differentially expressed and met the criterion for being candidate biomarkers. These transcripts were used as input variables to elastic net. The 10-fold cross-validation process in order to select the elastic net parameters suggested that a model with λ_2=0 and s=0.18 would minimise the mean squared error, corresponding to a 51-transcript model (Figure 41).

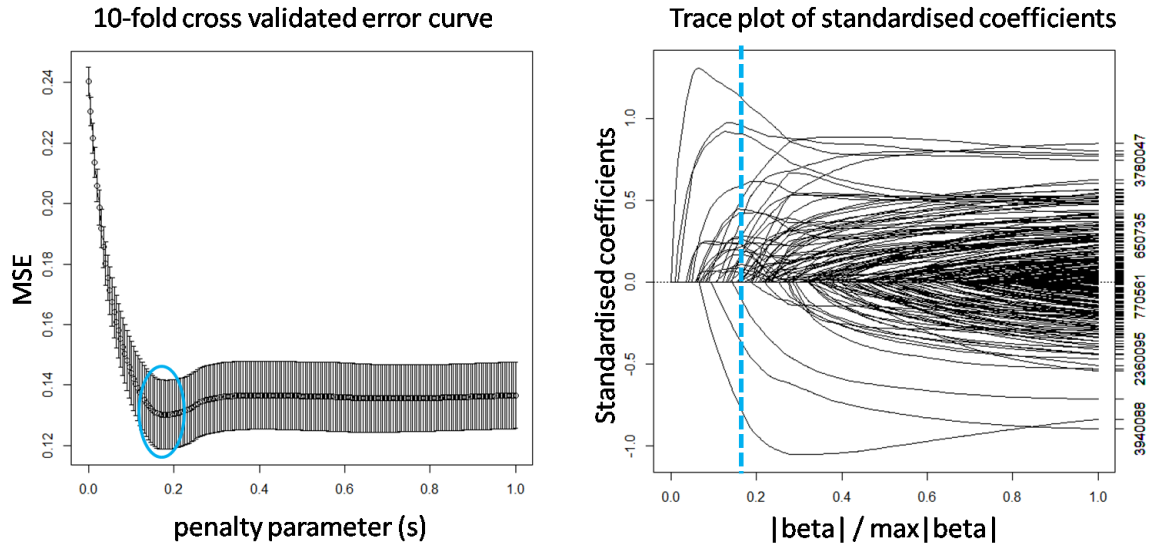


Figure 41: Mean Squared Prediction Error (MSE) plot and trace plot of the standardised coefficients for the TB vs. OD paediatric comparison (s=0.18).

The disease risk score was calculated for all the patients of the test set (Figure 42), as well as the training set using the 51 TB/OD transcript signature identified by the elastic net (Figure 43). The test set included 23 culture confirmed TB cases (14 HIV uninfected and 9 HIV infected) and 34 patients with other diseases (21 HIV uninfected and 13 HIV infected). AUC, sensitivity and specificity of the classifier were calculated to assess the DRS's classificatory performance (Table 28). The DRS based on the 51-transcript signature distinguished TB from other diseases of the patients in the test set with sensitivity and specificity of 78% $CI_{95\%}$(61-96) and 74% $CI_{95\%}$(56-88) respectively.

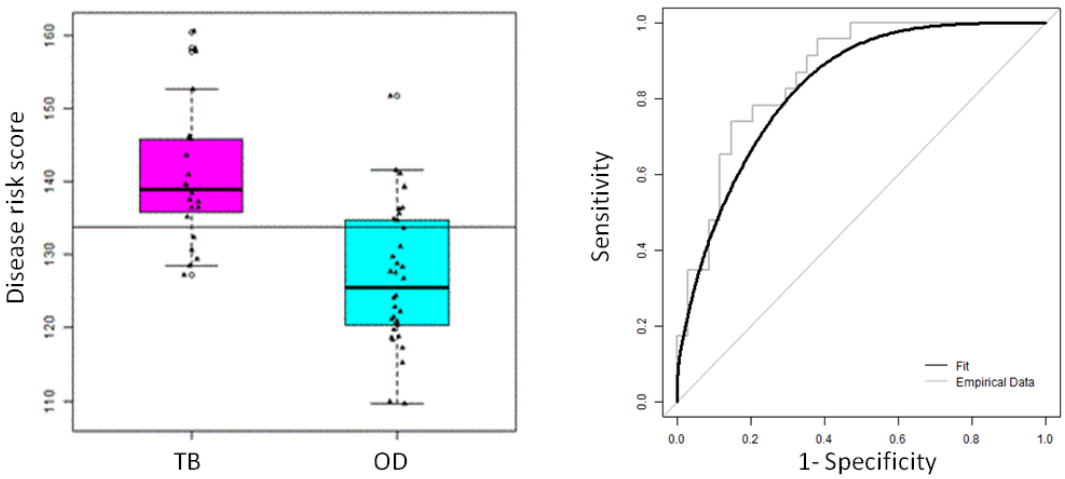


Figure 42: Classification of the South African/Malawi paediatric test set (20%) using the Disease Risk Score based on the 51 TB/OD transcript signature.

Disease risk score (left) and receiver operating characteristic curve (right) based on the 51 TB/OD transcript signature applied to the patients of the South African/Malawi paediatric test set (20%). Sensitivity, specificity and AUC are reported in Table 28 below.

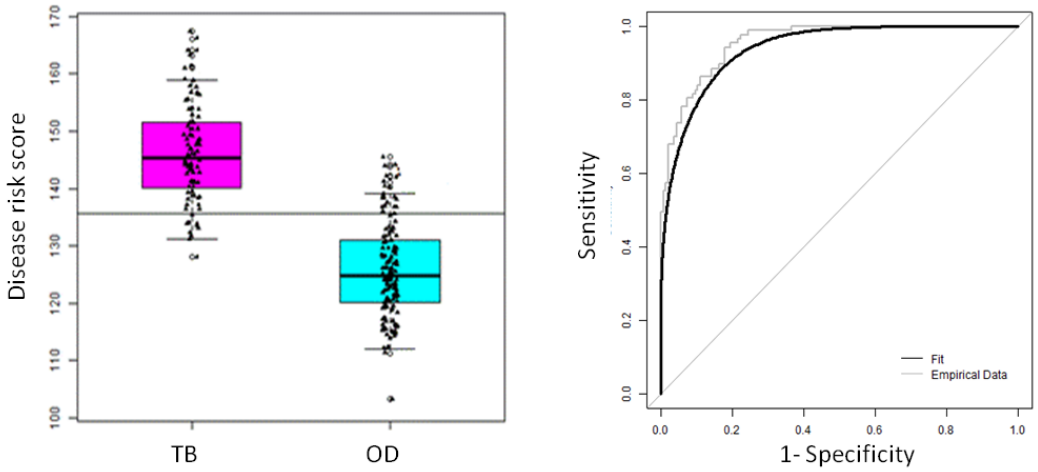


Figure 43: Classification of the South African/Malawi paediatric training set (80%) using the Disease Risk Score based on the 51 TB/OD transcript signature.

Disease risk score (left) and receiver operating characteristic curve (right) based on the 51 TB/OD transcript signature applied to the patients of the South African/Malawi paediatric training set (80%). Sensitivity, specificity and AUC are reported in Table 28 below.

	South Africa/Malawi cohort	
	Training set	**Test set**
AUC	**85.9%**	**86.2%**
(95% CI)	(96.3 - 99.8%)	(77.1 - 94.0%)
Sensitivity	**78.5%**	**78.3%**
(95% CI)	(61.0 - 94.5%)	(60.9 - 95.7%)
Specificity	**73.5%**	**73.9%**
(95% CI)	(58.8 - 88.2%)	(55.9 - 88.2%)

Table 28: Classification of the South Africa/Malawi paediatric training and test sets using the Disease Risk Score based on the 51 TB/OD transcript signature.
The DRS based on the 51 TB/OD transcript signature was applied to the South African/Malawi training and test sets.

The performance of the DRS on the HIV-infected and -uninfected patients of the test set separately was also assessed (Table 29). The DRS performed better among the HIV uninfected patients in comparison to the HIV-infected patients with sensitivity of 79% $CI_{95\%}$(57-100) and specificity of 81% $CI_{95\%}$(62-95). In the HIV-infected group the sensitivity was 78% $CI_{95\%}$(56-100) and the specificity 62% $CI_{95\%}$(31-85) respectively. This could be explained by the alterations in gene expression that HIV infection causes and the suppression of the immune system.

	South Africa/Malawi cohort (test set)	
	HIV-	**HIV+**
AUC	**88.4%**	**84.6%**
(95% CI)	(75.9 - 97.6%)	(64.0 - 96.6%)
Sensitivity	**78.6%**	**77.8%**
(95% CI)	(57.1 - 100.0%)	(55.6 - 100.0%)
Specificity	**81.0%**	**61.5%**
(95% CI)	(61.9 - 95.2%)	(30.8 - 84.6%)

Table 29: Classification of the South Africa/Malawi paediatric test set by HIV status using the Disease Risk Score based on the 51 TB/OD transcript signature.
The DRS based on the 51 TB/OD transcript signature was applied to the South African/Malawi test set, separately for the HIV-infected and -uninfected patients.

4.5 Benchmarking using an independent cohort

The performance of the disease risk score was assessed further using the samples from the paediatric Kenyan validation cohort. This cohort included patients that were TB culture positive, latently infected children, children with other diseases as well as children that were TB culture negative. Also, Xpert® MTB/RIF results and IGRA test results were available for the majority of the patients in the cohort which allowed a direct comparison of their results with the performance of the DRS. Additionally we explored the possibility of using housekeeping genes to establish cut-offs for the test.

4.5.1 Performance of the DRS on the Kenyan cohort

The Kenyan validation cohort was not included in the initial analysis and derivation of the signatures. The analysis of the microarrays from South Africa/Malawi was performed as previously described but the samples from Kenya as well as their microarrays were handled, pre-processed (background subtracted and normalised) and analysed separately. After transcript selection and signature identification using elastic net on the discovery cohort, we calculated the disease risk score by using the expression values of the preselected transcripts of the microarray samples from the Kenyan validation cohort.

First we explored the classificatory performance of the DRS based on the TB vs. LTBI 42 transcript signature on the culture-confirmed TB patients and the LTBI individuals (Figure 44). The dataset included 35 culture confirmed TB samples and 14 LTBI children. The DRS discriminated culture proven TB cases from the latently infected individuals with a sensitivity of 94% $CI_{95\%}$(86-100) and specificity of 100% $CI_{95\%}$(62-95) (Table 30). The performance of the DRS was slightly higher in the Kenyan cohort in comparison with the 20% test set from South Africa and Malawi.

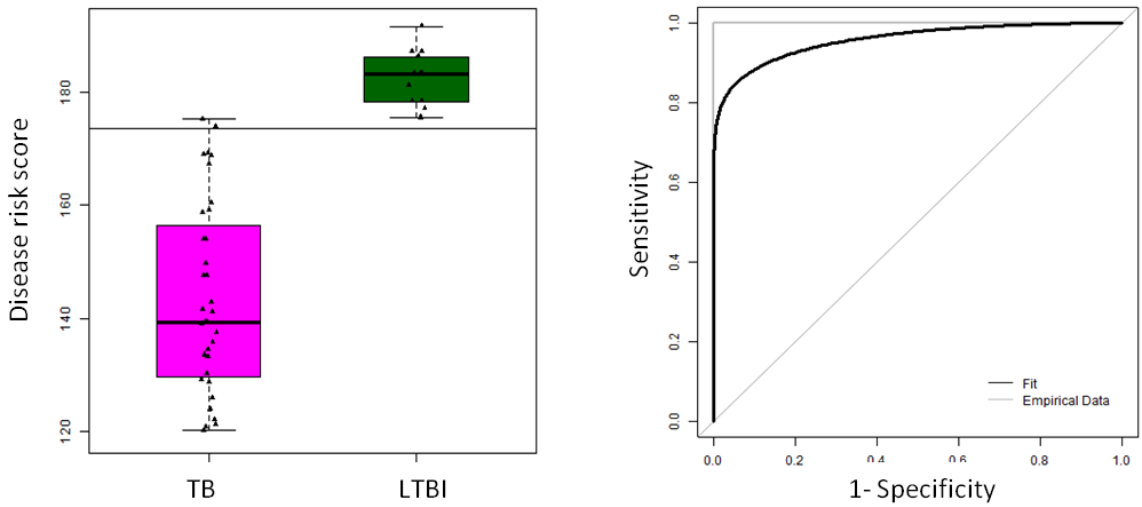


Figure 44: Classification of the paediatric validation cohort from Kenya using the Disease Risk Score based on the 42 TB/LTBI transcript signature.
Disease risk score (left) and receiver operating characteristic curve (right) based on the 42 TB/LTBI transcript signature applied to the patients of the paediatric validation cohort from Kenya. Sensitivity, specificity and AUC are reported in Table 30 below.

	Kenyan Validation Cohort	
	TB vs. LTBI	**TB vs. OD**
AUC	**100.0%**	**89.0%**
(95% CI)	(100.0 - 100.0%)	(82.3 - 94.9%)
Sensitivity	**94.3%**	**82.9%**
(95% CI)	(85.7 - 100.0%)	(68.6 - 94.3%)
Specificity	**100.0%**	**83.6%**
(95% CI)	(100.0 - 100.0%)	(74.6 - 92.7%)

Table 30: Classification of the Kenyan validation cohort using the Disease Risk Score based on the 42 TB/LTBI and 51 TB/OD transcript signatures.
The DRS based on the 42 TB/LTBI and 51 TB/OD transcript signature was applied to the validation cohort from Kenya.

Subsequently we assessed the classificatory performance of the DRS based on the TB/OD 51 transcript signature on the culture-confirmed TB patients and the patients with other diseases (Figure 45). The dataset included 35 culture confirmed TB samples and 55 children with a definite diagnosis other than TB that had a negative IGRA result. The DRS distinguished culture proven TB cases from patients with other diseases with a sensitivity of 83% $CI_{95\%}$(67-94) and specificity of 84% $CI_{95\%}$(75-93) (Table 30). The performance of the DRS

for the TB vs. OD comparison was slightly higher in the Kenyan cohort compared to the 20% test set from South Africa and Malawi.

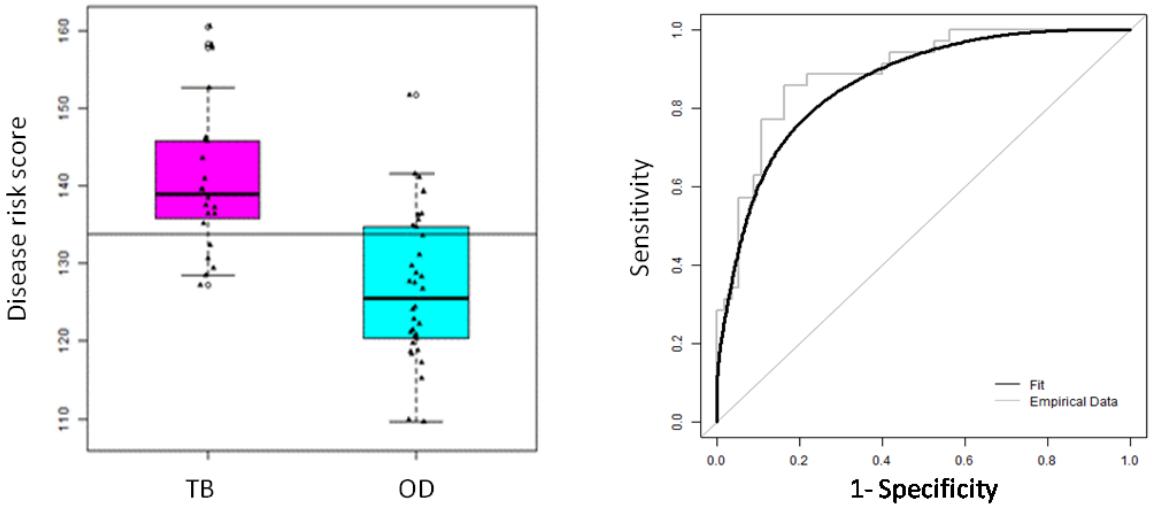


Figure 45: Classification of the paediatric validation cohort from Kenya using the Disease Risk Score based on the 51 TB/OD transcript signature.
Disease risk score (left) and receiver operating characteristic curve (right) based on the 51 TB/OD transcript signature applied to the patients of the paediatric validation cohort from Kenya. Sensitivity, specificity and AUC are reported in Table 27 above.

We further evaluated the classificatory performance of the DRS on the HIV-infected and -uninfected children separately (Figure 46, Table 31). The HIV-uninfected dataset consists of 25 culture positive TB cases and 29 OD patients. The HIV-infected dataset comprises 10 patients with culture confirmed TB and 29 patients with other diseases. In HIV uninfected group DRS distinguished TB cases from patients with other diseases with a sensitivity of 80% $CI_{95\%}$(64-92) and specificity of 81% $CI_{95\%}$(65-93). The performance of DRS for the HIV-infected children was slightly higher, with sensitivity of 90% $CI_{95\%}$(70-100) and specificity of 92% $CI_{95\%}$(81-100).

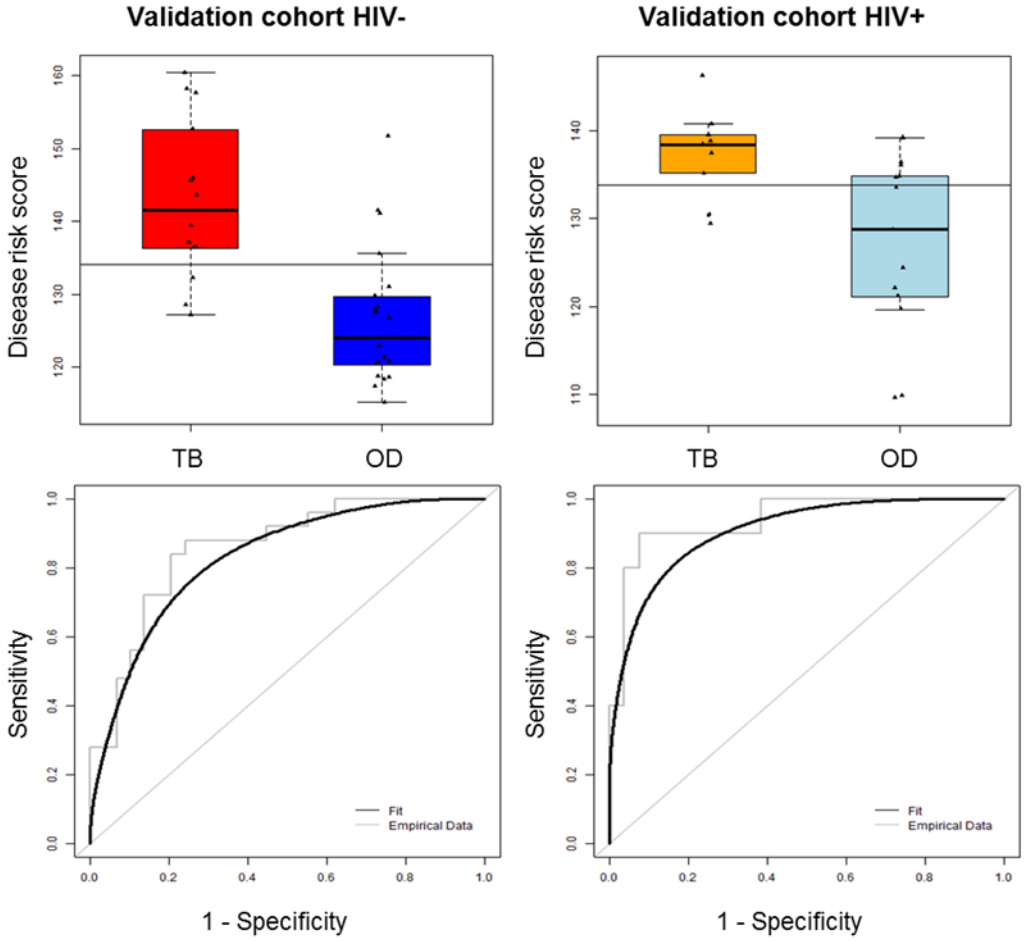


Figure 46: Classification of the paediatric validation cohort from Kenya by HIV status using the Disease Risk Score based on the 51 TB/OD transcript signature.
Disease risk score (top panels) & receiver operating characteristic curves (bottom panels) based on the 51 TB/OD transcript signature applied to the validation cohort from Kenya, calculated for the HIV-infected and -uninfected patients separately. Sensitivity, specificity and AUC are reported in Table 31 below.

	Kenyan Validation Cohort	
	HIV-	**HIV+**
AUC	**85.7%**	**93.9%**
(95% CI)	(75.0 - 94.4%)	(83.9 - 100.0%)
Sensitivity	**80.0%**	**90.0%**
(95% CI)	(64.0 - 92.0%)	(70.0 - 100.0%)
Specificity	**81.0%**	**92.3%**
(95% CI)	(65.4 - 93.1%)	(80.8 - 100.0%)

Table 31: Classification of the Kenyan validation cohort using the Disease Risk Score based on the 51 TB/OD transcript signature by HIV status.
The DRS based on the 51 TB/OD transcript signature was applied to the validation cohort from Kenya for the HIV-infected and -uninfected patients separately.

4.5.2 Comparison with Xpert® MTB/RIF

The Xpert® MTB/RIF assay was performed on respiratory samples from the patients included for gene expression analysis in the Kenyan cohort. Therefore we compared the performance of the DRS with that of the Xpert® MTB/RIF in the culture-confirmed TB group, while the OD group was used to calculate test specificity. The Xpert® MTB/RIF test had a positive result for 19 out of 35 culture-confirmed TB cases and none out of 55 other diseases cases. Overall, 25 culture positive TB and 29 OD patients were HIV negative while the rest were HIV positive.

	Kenyan Validation Cohort		
	HIV-/+	**HIV-**	**HIV+**
AUC	**77.1%**	**74.0%**	**85.0**
(95% CI)	(69.9 - 85.7%)	(64.0 - 84.0%)	(70.0 - 95.1)
Sensitivity	**54.3%**	**48.0%**	**70.0**
(95% CI)	(37.1 - 68.6%)	(28.0 - 64.1%)	(40.0 - 100.0)
Specificity	**100.0%**	**100.0%**	**100.0%**
(95% CI)	(100.0 - 100.0%)	(100.0 - 100.0%)	(100.0 - 100.0%)

Table 32: Diagnostic performance of the Xpert® MTB/RIF in the Kenyan validation cohort. The Xpert® MTB/RIF test assay was used for classification of the patients in the validation cohort from Kenya. Results are presented for the HIV-infected and -uninfected patients combined as well as separately.

Xpert® MTB/RIF is highly specific (specificity of 100%) but its sensitivity is limited (sensitivity of 54.3 $CI_{95\%}$(37.1-68.6)) (Table 32). These results agree with previously published data on the performance of Xpert® MTB/RIF in children showing that it detects only a portion of the culture confirmed TB cases, even when two tests are undertaken but it is highly specific [171]. Comparison with the results acquired using DRS from Table 30 showed that the DRS identified a higher proportion of culture-confirmed TB cases but was less specific than Xpert® MTB/RIF.

4.5.3 Culture negative TB cases

Many children are commenced on TB treatment without any bacteriological confirmation and it is widely accepted that clinical diagnostic scores over-diagnose TB but the extent of over-diagnosis is unknown [172]. As culture negative TB represents the majority of paediatric TB cases worldwide, we further explored the performance of the DRS in this group, for which no 'gold standard' is available. The DRS was calculated for the 44 patients with culture-negative tuberculosis (8 patients in whom tuberculosis was highly probable, 19 in whom it was probable, and 17 in whom it was possible) (Figure 47).

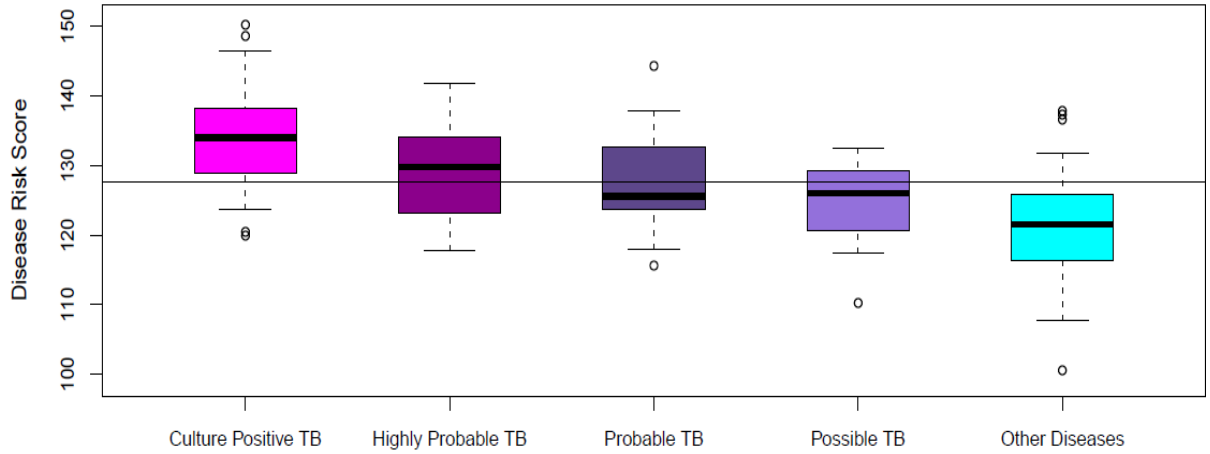


Figure 47: Disease risk score based on the TB/OD 51-transcript signature applied to the independent Kenyan validation cohort by clinical subgroup
Disease risk score based on the 51 TB/OD transcript signature was applied to the validation cohort from Kenya (culture confirmed TB, culture negative TB and other diseases), calculated for the HIV-infected and -uninfected patients combined.

Xpert® MTB/RIF results on the culture negative group of children were also available, so we compared the performance of both DRS and Xpert® MTB/RIF in the culture-confirmed, "highly probable", "probable" and "possible" culture-negative TB groups separately. The Xpert® MTB/RIF test had a positive outcome for 3 out of 44 culture-negative TB cases and 0 out of 55 other diseases cases. In each case we used the same OD group as the TB-negative comparator group to calculate test specificity. Among patients with negative cultures who were treated for tuberculosis, the DRS identified 63% $CI_{95\%}$(25 - 100) of those in whom tuberculosis was highly probable, 42% $CI_{95\%}$(21 - 63) of those in whom it

was probable, and 35% $CI_{95\%}$(12 - 59) of those in whom it was possible. For the same patients Xpert® MTB/RIF identified 25% $CI_{95\%}$(25 - 50) of those in whom TB was highly probable, 5% $CI_{95\%}$(0 - 18) of those in whom it was probable, and 0% of those in whom it was possible. Its specificity was 100% as the same OD group was used (Figure 48).

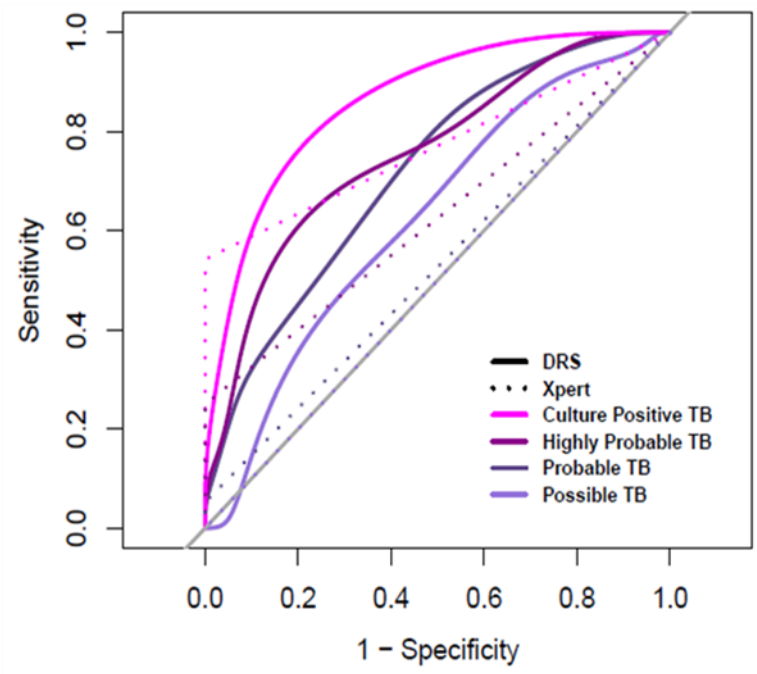


Figure 48: Receiver operating characteristic curves for the Kenyan cohort using the Disease Risk Score based on the 51 TB/OD transcript signature for the different subgroups.
Receiver operating characteristic curves based on the 51 TB/OD transcript signature applied to the different subgroups of the validation cohort from Kenya. Sensitivity, specificity and AUC are reported in Table 33.

However, each category of the culture negative TB group is a mixture of “actual” TB cases and OD cases clinically confused with TB. Therefore, and in order to obtain more realistic estimates of the test sensitivity across the culture-negative groups, this was accounted for in the following way:

The observed true-positive rate (TPR) was modelled as a function of the unknown actual TPR, the false-positive rate (FPR) estimated from the OD group, and the prevalence of TB (Equation 4.2), from which a corrected Receiver Operator Characteristic (ROC) curve and estimates of 'effective' sensitivity were calculated in each category. As the prevalence of TB

in each category is unknown, we investigated a range of prevalence of 70%-90%, 40%-60% and 30%-50% for "highly probable", "probable" and "possible" TB respectively. The already calculated unadjusted sensitivity is equivalent to assuming a TB prevalence of 100% in each category. As the application of a classifier, such as DRS, to the culture-negative TB group results in an observed estimate of the true-positive rate (**TPR_{obs}**) which is the proportion of all observed culture–negative TB cases (**P_{obs}**) scored as 'positive' by the classifier. However, these observed positives are in fact a mixture of actual true TB and false TB (i.e. OD), hence:

$$TPR_{obs} = \frac{TP_{obs}}{P_{obs}} = \frac{TP_{actual} + FP_{actual}}{P_{actual} + F_{actual}}$$

$$TPR_{obs} = \frac{TPR_{effective} * P_{actual} + FPR_{effective} * F_{actual}}{P_{actual} + F_{actual}}$$

$$TPR_{obs} = TPR_{effective} * \frac{P_{actual}}{P_{actual} + F_{actual}} + FPR_{effective} * \frac{F_{actual}}{P_{actual} + F_{actual}}$$

$$TPR_{obs} = TPR_{effective} * \Pr(TB) + FPR_{effective} * (1 - \Pr(TB)) \quad (4.1)$$

where:

F_{actual} is the number of OD and Pr(TB) is the prevalence of true TB and in the group under consideration.

$FPR_{effective}$ is the false-positive rate, estimated as the proportion of OD cases that are falsely called TB by the classifier. We can re-arrange equation (4.1) to obtain a formula for the effective TPR in terms of the group prevalence and the FPR estimated from the OD group:

$$TPR_{effective} = \frac{TPR_{obs} - FPR_{actual} * (1 - \Pr(TB))}{\Pr(TB)} \quad (4.2)$$

		AUC % (95% CI)	Sensitivity % (95% CI)	Effective Sensitivity % (95% CI)		
	Estimated "actual" TB prevalence in group	**100%**	**100%**	**70%**	**80%**	**90%**
Highly Probable TB vs. OD (n_{TB}=8 n_{OD}=55)	**DRS 51 TB vs. OD signature**	77.5 (58.2 - 94.3)	62.5 (25.0 - 100.0)	82.3 (41.9 - 100.0)	74.1 (37.6 - 100.0)	67.6 (35.1 - 100.0)
	Xpert® MTB/RIF	62.5 (50.0 - 81.3)	25.0 (0.0 - 50.0)	35.7 (1.1 - 65.7)	31.3 (1.0 - 57.6)	27.8 (1.0 - 51.3)
	Estimated "actual" TB prevalence in group	**100%**	**100%**	**40%**	**50%**	**60%**
Probable TB vs. OD (n_{TB}=19 n_{OD}=55)	**DRS 51 TB vs. OD signature**	72.3 (59.6 - 84.2)	42.1 (21.1 - 63.2)	80.8 (36.4 - 100.0)	67.9 (32.7 - 100.0)	59.3 (30.2 - 90.6)
	Xpert® MTB/RIF	52.6 (50.0 - 57.9)	5.3 (0.0 - 17.8)	13.3 (0.0 - 36.5)	10.6 (0.0 - 29.3)	8.8 (0.0 - 24.5)
	Estimated "actual" TB prevalence in group	**100%**	**100%**	**30%**	**40%**	**50%**
Possible TB vs. OD (n_{TB}=17 n_{OD}=55)	**DRS 51 TB vs. OD signature**	64.5 (48.4 - 77.7)	35.3 (11.8 - 58.8)	79.6 (7.2 - 100.0)	63.8 (9.2 - 100.0)	54.3 (10.2 - 91.0)
	Xpert® MTB/RIF	50.0 (50.0 - 50.0)	0.0 (0.0 - 0.0)	0.0 (0.0 - 0.0)	0.0 (0.0 - 0.0)	0.0 (0.0 - 0.0)

Table 33: Classification of the culture-negative TB samples from the Kenyan validation cohort using the Disease Risk Score based on the 51 TB/OD transcript signature and the Xpert® MTB/RIF results using a range of prevalence for each subgroup.

The effective sensitivity of the disease risk score for highly probable, probable, and possible cases of tuberculosis was 67.6 to 82.3%, 59.3 to 80.8%, and 54.3 to 79.6%, respectively (Table 33, Figure 49). The sensitivity of the disease risk score was higher than that of the Xpert® MTB/RIF assay in all tuberculosis categories, with the Xpert® MTB/RIF assay having sensitivities of 27.8 to 35.7% for the subgroup in which tuberculosis was highly probable, 8.8 to 13.3% for that in which it was probable, and 0% for that in which it was

possible. However, the Xpert® MTB/RIF assay was highly specific (100%). And the specificity of DRS remains the same as in Table 30.

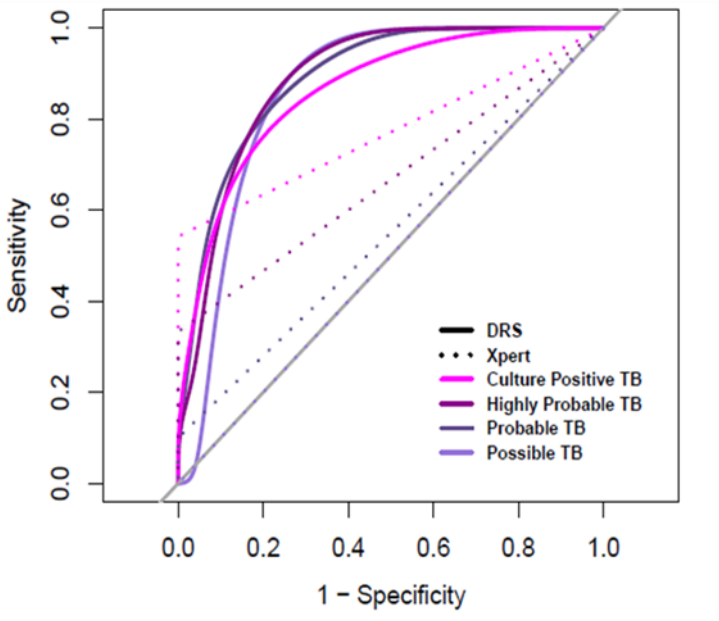


Figure 49: Receiver operating characteristic curves for the Kenyan cohort based on the Disease Risk Score based on the 51 TB/OD transcript signature for the different subgroups, using an adjusted sensitivity of 'actual' TB of 80% for the highly probable, 50% for the probable and 40% for the possible TB cases.
Sensitivity, specificity and AUC are reported in Table 33.

To explore how the DRS for TB/OD may contribute to TB diagnosis in clinical practice, we evaluated its positive and negative predictive values, calculated as functions of specificity, sensitivity and prevalence according to the formulae (2.19 – 2.20). Calculation of PPV and NPV require prior knowledge of the proportion of TB cases in the population under investigation. However, as prevalence of TB has a strong dependence on the recruitment strategy, as evidenced by the different prevalence in the three cohorts, we evaluated the performance of the DRS using a range of estimates representative of the three populations undergoing screening (10%, 30% and 50%), and probably representative of the prevalence among a future populations undergoing TB screening.

- 10%: reflects the prevalence of TB in the Kenyan cohort
- 30%: reflects the prevalence of TB in the South Africa and Malawi recruitment
- 50%: reflects prior filtering of the patients or combination with another test

Also, given the dependency of NPV and PPV on test sensitivity and specificity, we calculated estimates for these values reflective of the different scenarios in which a diagnostic test for TB would be applied. We employed the specificity of the DRS for the HIV-infected and -uninfected other disease group from the Kenyan validation set. Regarding the sensitivity, we used an estimate derived from the combined culture-positive and culture-negative TB groups: a weighted average of the 'effective' sensitivity in the culture-confirmed, “highly probable” (HP), “probable” (Pr) and “possible” (Pos) TB. The weights are given by the proportion of samples in the Kenyan prospective study which were assigned to each of these groups. We considered three scenarios, with proportions listed below:

- **A:** TB prevalence in: culture confirmed = 100%; HP TB = 70%; Pr TB =40%; Pos TB = 30%
- **B:** TB prevalence in: culture confirmed = 100%; HP TB = 80%; Pr TB =50%; Pos TB = 40%
- **C:** TB prevalence in: culture confirmed = 100%; HP TB = 90%; Pr TB =60%; Pos TB = 50%

These scenarios allowed us to calculate a combined sensitivity: the average sensitivity across all culture-negative and -positive TB groups.

Combined sensitivity	Measure	Prevalence		
		10%	**30%**	**50%**
A: 70%	**PPV %**	38.3	70.5	84.8
	(95% CI)	(23.4 - 53.3)	(57.4 - 83.7)	(76.6 - 943.0)
	NPV %	93.6	87.1	74.4
	(95% CI)	(94.9 - 97.7)	(82.8 - 91.5)	(67.0 - 81.8)
B: 75%	**PPV %**	41.0	72.9	86.2
	(95% CI)	(25.8 - 56.3)	(60.4 - 85.3)	(78.7 - 93.7)
	NPV %	96.9	89	77.6
	(95% CI)	(95.5 - 98.2)	(84.6 - 93.4)	(69.8 - 85.4)
C: 82%	**PPV %**	44.3	75.4	87.8
	(95% CI)	(28.8 - 59.8)	(63.8 - 87.1)	(81.0 - 94.5)
	NPV %	97.8	91.9	82.9
	(95% CI)	(96.5 - 99.0)	(87.6 - 96.1)	(74.9 - 90.9)

Table 34: PPV and NPV for the Kenyan validation cohort in different prevalence scenarios & based on the sensitivity in both culture-negative and culture-positive TB groups.

PPV was highest when the prevalence of TB in the screened population was high, and decreased as prevalence decreased, whereas NPV increased with the decreasing prevalence (Table 34).

4.5.4 Comparison with IGRA and CRP

To assess further the potential diagnostic role of transcriptional signatures and the DRS, we compared it against measurements of the levels of C-reactive protein (CRP) and the IGRA. In the Kenyan cohort 35 culture-confirmed TB patients and 54 patients from the OD group were IGRA tested. IGRA distinguished culture confirmed TB cases from other diseases with a sensitivity of 60% $CI_{95\%}$(33-87) and a specificity of 83% $CI_{95\%}$(72-93). The disease score was more sensitive than the IGRA while its specificity was the same.

Previous studies have suggested that C-reactive protein levels may be useful for TB diagnosis in adult smear negative patients [173]. Therefore CRP was measured in patients' serum collected at the same time as blood for RNA expression by ELISA. No significant differences were identified between the 44 culture negative TB, 52 other diseases and 34 culture confirmed TB patients that were tested (Figure 45).

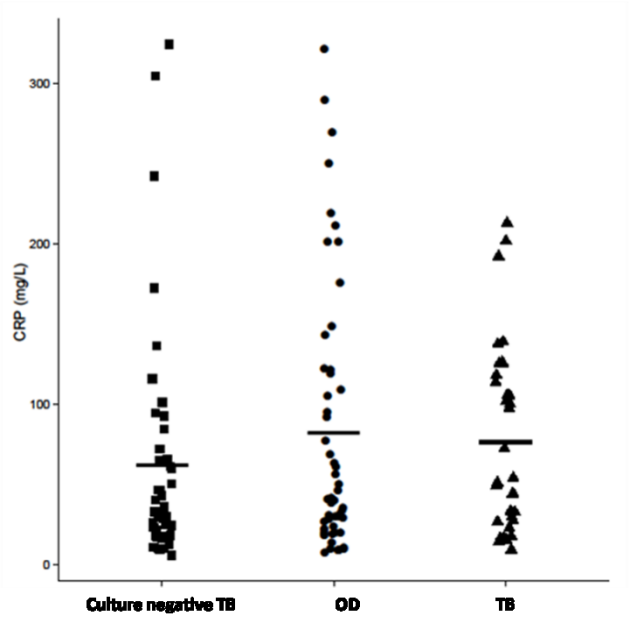


Figure 50: Serum CRP measurements from Kenyan validation cohort patients.

Each dot represents one patient. Bars indicate median value.

4.5.5 Including IGRA+ patients in the OD group

We have deliberately excluded from the OD category children with a positive IGRA to minimise this potentially important source of misclassification bias. However, in order to evaluate the OD patients that had a positive IGRA result, we performed the TB vs. OD comparison with inclusion of the IGRA positive patients. The IGRA positive patients in the OD group may have either self-resolving primary TB or latent TB infection. There were 9 IGRA+ patients in the OD group randomly selected for array. The sensitivity of DRS for TB vs. OD remained unchanged with or without the IGRA+ patients, while specificity was 1% lower for when the IGRA+ patients were included (Figure 51, Table 35).

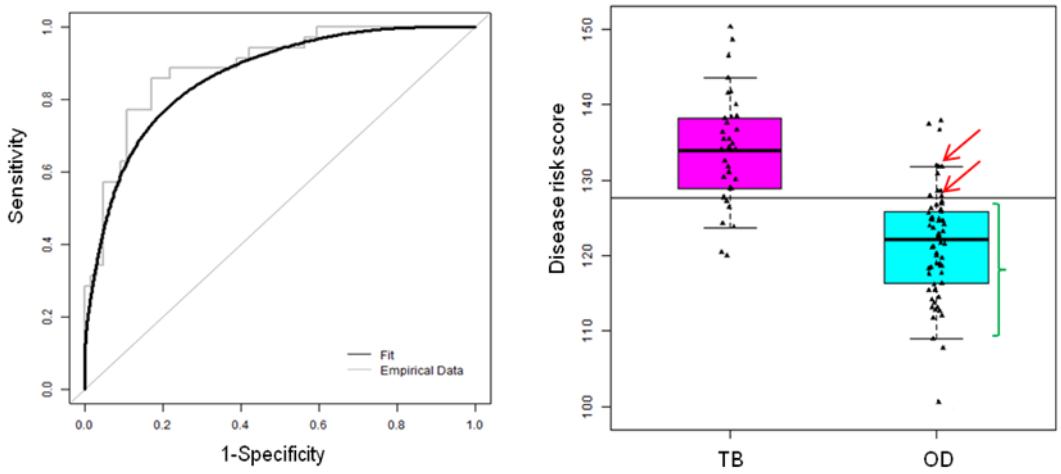


Figure 51: Classification of the Kenyan paediatric dataset using the Disease Risk Score based on the 51 TB/OD transcript signature, including the IGRA+ OD patients.
Sensitivity, specificity and AUC are reported in Table 35 below.

	Kenya Validation Cohort TB vs. OD (including IGRA+)
AUC	**89.0%**
(95% CI)	(81.9 - 95.3%)
Sensitivity	**82.9%**
(95% CI)	(68.6 - 94.3%)
Specificity	**82.8%**
(95% CI)	(73.4 - 92.2%)

Table 35: Classification of the Kenyan validation cohort using the Disease Risk Score based on the 51 TB/OD transcript signature, including the IGRA+ OD patients.

Disease risk score based on the 51 TB/OD transcript signature was applied to the TB/OD validation cohort from Kenya. The OD control group included the IGRA+ patients.

Seven out of the nine patients of the other disease group that were IGRA positive were classified as "not TB" by the DRS. The two misclassified OD IGRA+ patients are indicated by red arrows in Figure 51. As sensitivity of the DRS was unchanged when IGRA positive patients were included or excluded, and to exclude possibility of including self-resolving primary TB in the OD group, we used only IGRA negative OD cases for calculation of performance of DRS in the culture negative group.

4.5.4 Threshold computation by housekeeping genes

As the threshold values for the DRS are acquired from the population samples and are slightly different between training, test and validation cohorts, we explored the possibility of having a global DRS threshold based on information extrapolated from housekeeping genes. The housekeeping genes are genes that are crucial for fundamental functions in the cells and are expressed under all conditions.

First, we identified the housekeeping transcripts within the training cohort. As a measure for variability for every transcript we employed the coefficient of variation in the training cohort calculated by dividing the standard deviation by the mean. Low values of the coefficient of variation indicate more homogeneity across samples regardless of location, HIV status and different disease group (Figure 52). We identified the 20 transcripts with lowest coefficient of variation values for use as housekeeping transcripts.

Currently the thresholds for the training, test and Kenyan validation cohort are: 135.6, 133.8 and 127.6 respectively and their ratios relative to the training set are 1, 0.99 and 0.94. These ratios are very similar to the ratios of the medians of all the medians of the housekeeping transcripts which calculated in relation to the training set are: 1, 0.99 and 0.95. The ratios of the housekeeping transcripts seem to reflect the scaling of the thresholds. This scaling could be attributed to the different array intensities, RNA handling bias or other batch effects.

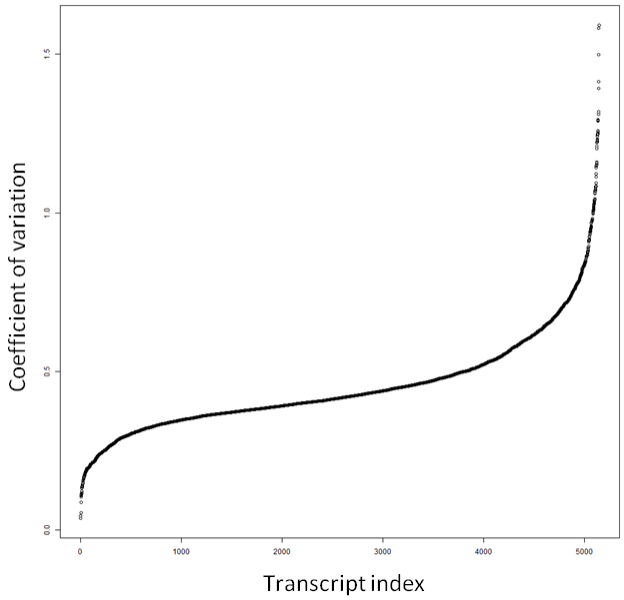


Figure 52: Coefficient of variation calculated for every transcript across samples of the training cohort.

Using the medians of the housekeeping genes and the threshold of the training set, we calculated the thresholds in the testing and validation cohorts for TB vs. OD comparison, which were 134.9 and 128.1 respectively. In the validation cohort from Kenya specificity improves from 83.6% to 87.2% while sensitivity decreases from 82.8% to 77.1%. Further work needs to be undertaken to finalise this analysis.

4.6 Application of the TB vs. LTBI signature to the culture negative TB cases

The culture negative TB cases presented with TB symptoms and were categorised as having highly probable, probable and possible TB according to the algorithm presented in Figure 30. The inclusion criteria in the study were any of: cough, fever or weight loss for more than 2 weeks, pneumonia not responding to antibiotics, history of close TB contact or clinician's clinical suspicion of TB for any other reason. Although *M. Tuberculosis* was not isolated from clinical specimens, the majority of the patients - particularly in the highly probable and the probable group - were smear positive or had a reactive Mantoux. We

further explored the performance of the TB vs. LTBI 42-transcript signature in the highly probable, probable and possible groups (Figure 53).

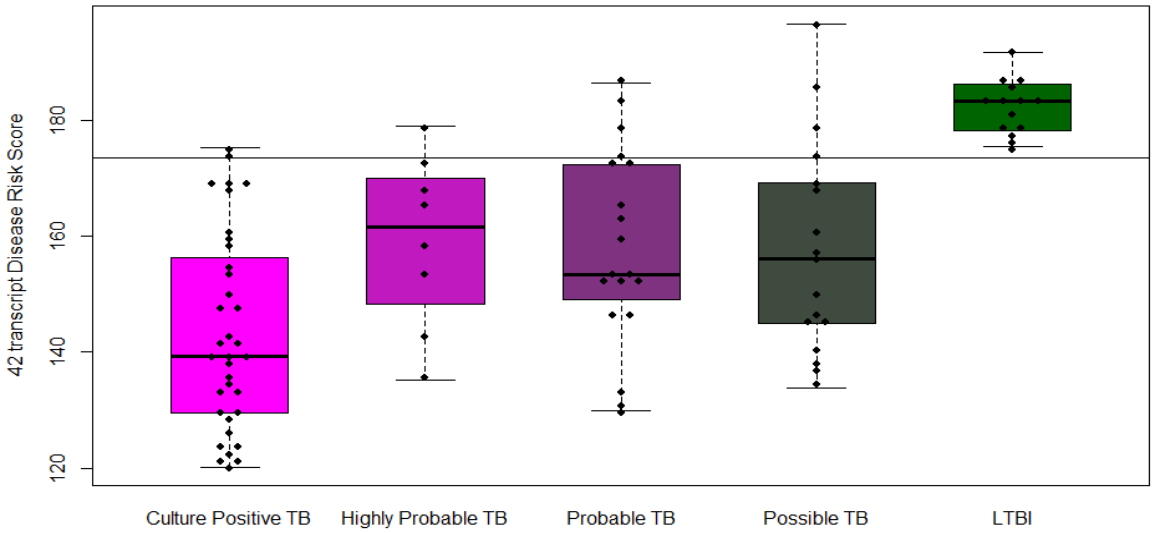


Figure 53: Disease risk score based on the TB/LTBI 42-transcript signature applied to the independent Kenyan validation cohort by clinical subgroup
Disease risk score based on the 42 TB/LTBI transcript signature was applied to the validation cohort from Kenya (culture confirmed TB, culture negative TB and LTBI).

The majority of culture negative TB cases were classified as TB. Only one case in the highly probably TB group, three cases in the probable TB group and four in the possible TB group were classified as LTBIs. However, as all these patients show symptoms of disease they may be classified as TB, because the TB vs. LTBI signature is also capturing a "disease vs. healthy" signature. It also is expected that clinical diagnosis would over estimate true TB prevalence in the "highly probable", "probable" and "possible" TB groups.

4.7 Signatures

In this section we present the transcripts that are included in the paediatric transcriptomic signatures, as well as their direction of regulation (Table 36, Table 37). Eight transcripts of the TB/LTBI signature were also present in the previously published 393-transcript signature [60], while three were also included in the TB/OD adult signature. The overlap with the adult TB/LTBI signature was with only one transcript corresponding to *GBP6*.

Array ID	Gene Symbol	Probe ID	Direction*	Overlap with 393-transcript signature
6480059	*ACTA2*	ILMN_6588	up	√
3310324	*ALKBH7*	ILMN_7229	down	
5550397	*APOL6*	ILMN_38312	up	√
7400341	*C11ORF2*	ILMN_10940	down	
1500546	*C20ORF201*	ILMN_25727	down	
6380187	*C21ORF57*	ILMN_21121	down	
1470706	*C8ORF55*	ILMN_25304	down	
2030170	*CARD16*	ILMN_21555	up	√
6110427	*CLIP1*	ILMN_15054	up	
5340246	*CRIP2*	ILMN_29728	down	
4540239	***DEFA1***	**ILMN_29692**	**up**	
4860128	*DEFA1B*	ILMN_176067	up	
2970747	*DEFA3*	ILMN_11220	up	
7200274	*DGCR6*	ILMN_138781	down	
3440647	*DNAJC30*	ILMN_30295	down	
3390068	*E4F1*	ILMN_23848	down	
4670441	*FBLN5*	ILMN_29187	down	
1510364	***GBP5***	**ILMN_24462**	**up**	√
3780047	***GBP6***	**ILMN_1956**	**up**	√
450632	*GNG3*	ILMN_7558	down	
1500575	*HS.538100*	ILMN_103699	down	
4590026	*IMPDH2*	ILMN_3439	down	
7330575	*KLHL28*	ILMN_22112	down	
2810669	*LCMT1*	ILMN_16696	down	
5340414	*LGTN*	ILMN_4831	down	
2140541	*LOC389816*	ILMN_182870	down	
620403	*LOC400759*	ILMN_181219	up	
2230538	*LRRN3*	ILMN_306943	down	
5560075	*MFGE8*	ILMN_11368	down	
4210411	*NDRG2*	ILMN_19545	down	√
6450424	*NME3*	ILMN_23571	down	
6770603	*NOG*	ILMN_7080	down	
4150017	*PAQR7*	ILMN_3765	down	
2140382	*PASK*	ILMN_19873	down	√
4150100	*PASK*	ILMN_19873	down	√
7150189	*PHF17*	ILMN_1535	down	
3400468	*RAP1A*	ILMN_20446	up	

4670487	*SIVA*	ILMN_6846	down
6280433	*SNHG7*	ILMN_371358	down
4260189	*TGIF1*	ILMN_162784	down
4050059	*U2AF1L4*	ILMN_8757	down
6550358	*UBA52*	ILMN_27795	down

Table 36: The 42 transcript signature for distinguishing TB from LTBI in children
* in TB patients in relation to patients with LTBI. 3 transcripts were also in the TB/OD signature (in bold).

In the TB/OD paediatric signature, no overlap was found with the previously published TB/OD 86 transcript signature [60]. However, seven transcripts were common between the TB/OD adult and paediatric signatures corresponding to five genes *ALDH1A1*, *CYB561* (2), *LOC389386* (2), *MIR1974*.

Array ID	Gene Symbol	Probe ID	Direction*	Overlap with 86-transcript signature
4180768	*ALAS2*	ILMN_13644	up	
1070477	*ALDH1A1*	ILMN_177898	up	
5910019	*C1QB*	ILMN_36274	up	
4290026	*C20ORF103*	ILMN_165304	down	
2600634	*C3HC4*	ILMN_6980	down	
1580048	*CAST*	ILMN_163108	up	
3390564	*CCDC52*	ILMN_23129	up	
3940754	*CD226*	ILMN_3877	up	
1780440	*CD79A*	ILMN_37614	up	
5890653	*CDKN1C*	ILMN_20689	down	
5340767	*CEACAM1*	ILMN_21651	down	
130086	*CYB561*	ILMN_8373	up	
840446	*CYB561*	ILMN_20474	up	
4540239	***DEFA1***	**ILMN_29692**	up	
1050068	*F2RL1*	ILMN_176188	up	
6510707	*FER1L3*	ILMN_18562	up	
6840767	*FRMD3*	ILMN_11826	down	
2350189	*GBP3*	ILMN_3653	up	
1510364	***GBP5***	**ILMN_24462**	up	
3780047	***GBP6***	**ILMN_1956**	up	
6220739	*GRAMD1B*	ILMN_308544	down	
5260484	*HLA-DRB1*	ILMN_20550	up	
6370315	*HLA-DRB5*	ILMN_3178	up	
620544	*HLA-DRB6*	ILMN_5312	up	
630619	*HPSE*	ILMN_165418	down	
5340762	*HS.106234*	ILMN_74965	up	
7320678	*HS.171481*	ILMN_80341	up	
4880370	*JUP*	ILMN_3789	down	
1050215	*KCNJ15*	ILMN_164363	down	
2570438	*KIFC3*	ILMN_4695	up	
7560114	*KLHDC8B*	ILMN_6513	up	
5310445	*KREMEN1*	ILMN_41914	down	

4570164	*LOC389386*	ILMN_165610	up
4780044	*LOC389386*	ILMN_352098	up
2350121	*LOC642678*	ILMN_38908	up
6480364	*LOC647460*	ILMN_38026	down
6900291	*LOC649210*	ILMN_33006	down
830639	*LOC653778*	ILMN_32201	down
2260349	*MIR1974*	ILMN_388657	down
830750	*NCF1B*	ILMN_168368	up
6760593	*OSBPL10*	ILMN_11112	up
3170246	*PDCD1LG2*	ILMN_3561	up
2000292	*SCGB3A1*	ILMN_23096	down
160368	*SEMA6B*	ILMN_21277	down
1400593	*SIGLEC14*	ILMN_309673	up
460463	*SMARCD3*	ILMN_19301	up
540520	*SNORD8*	ILMN_366693	up
1240554	*TNFRSF17*	ILMN_17574	up
4760747	*TPST1*	ILMN_174128	up
2630195	*VAMP5*	ILMN_20179	down
3940088	*ZBED2*	ILMN_4927	down

Table 37: The 51 transcript signature for distinguishing TB from OD in children
* in TB patients in relation to patients with OD. 3 transcripts were also in the TB/LTBI signature (in bold).

The 42 transcripts for distinguishing active TB from latent TB infection in children correspond to 34 unique genes (10 up-regulated and 24 down-regulated in TB disease). When IPA was used to map the transcripts to diseases and functions, inflammatory disease and antimicrobial response were the top two. Due to the nature of the disease in children and the recruitment process, the LTBI group includes only HIV uninfected children, while the active TB group includes both HIV infected and uninfected children. Although, the effect of HIV infection was accounted for the derivation of the signature, a more rigorous biological approach that would take into account all the differentially expressed genes in patients without HIV infection is needed, to gain biological insight into the tuberculosis disease pathogenesis.

The 51 transcripts for distinguishing active TB from other diseases in children correspond to 41 unique genes (28 up-regulated and 13 down-regulated in TB disease). Some of the transcripts overlap with the adult signatures and have been discussed in Chapter 3. Another gene encoding for the HLA class II beta chain (*HLA-DRB1*) is included in the TB vs. OD signature paediatric signature (*HLA-DPB1* is included in the adult signature), being up-regulated in the active TB patients. A paralogue (*HLA-DRB5*) and a related

pseudogene (*HLA-DRB6*) are also included [144]. A recent meta-analysis showed that *HLA-DRB1* polymorphisms were associated with susceptibility to pulmonary tuberculosis especially in Asian populations [174]. Additionally, *CD226* (Cluster of Differentiation 226) that is also up-regulated is a protein coding gene, encoding for a glycoprotein expressed on the surface of NK cells, platelets, monocytes and a subset of T cells that is a member of the Ig-superfamily [144]. *F2RL1* (coagulation factor II (thrombin) receptor-like 1) is involved in regulation of innate and adaptive immunity, and promotes inflammation [144]. *KCNJ15* (potassium inwardly-rectifying channel, subfamily J, member 15) has been found significantly differentially expressed in a recent whole blood TB biomarker study [175].

The top canonical pathways in IPA included B cell development, altered T-cell and B-cell signalling, communication between innate and adaptive immune cells, crosstalk between dendritic cells and natural killer cells and the antigen presentation pathway. The top diseases and disorders are immunological disease, inflammatory disease and inflammatory response. The top upstream regulator is *IRF1* that encodes interferon regulatory factor 1, a member of the interferon regulatory transcription factor (IRF) family. Out of the 41 unique genes, 15 were reported as being IFN-inducible (either type I or type II) in the *Interferome* database [148].

In order to explore the overlap of the transcriptomic signatures between adults and children, a 4-way Venn diagram of the genes corresponding to the transcripts of the TB signatures was created (Figure 54). Although this is not a systematic manner to deduce findings about the underlying biology of TB disease, we observe that the majority of genes are unique to each signature. There are a few genes in common between pairs of signatures, while only one gene, *GBP6* (guanylate-binding protein 6), is common to all 4 signatures. Its biological role is described in Chapter 3.

A more rigorous biological approach that would take into account all the differentially expressed genes, and not only the members of the signatures, is needed for the exploration of the human host response to TB disease and infection in terms of pathways.

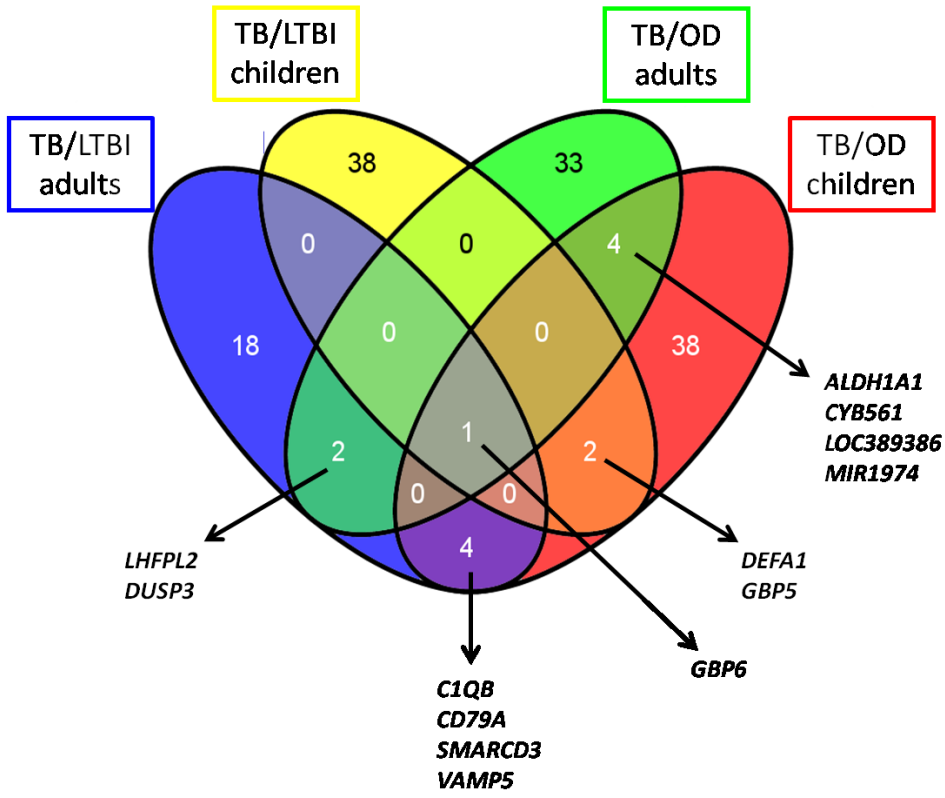


Figure 54: Venn diagram of the corresponding genes for the TB/LTBI and TB/OD adult and paediatric transcriptomic signatures.

4.8 Summary

In Chapter 4 I describe the results obtained from the analysis of a paediatric whole blood gene expression study including patients recruited in three African sites, to detect TB biomarkers. Microarrays from South Africa and Malawi were used as the discovery set for the biomarkers, while microarrays from patients recruited in a prospective independent study in Kenya were used for validation and benchmarking.

The bioinformatics pipeline described in Chapter 2 was applied to the samples of the discovery cohort to identify minimal transcriptomics signatures. The signatures in the form of the DRS were subsequently used to classify patients in the validation cohort. The particular challenges of this analysis were posed by the TB culture negative group: children that receive TB treatment without bacteriological confirmation, so the actual burden of TB within this category is unknown.

The results showed that a small number of RNA transcripts detected in whole blood were able to distinguish TB from latent TB infection and from other diseases with clinical features similar to TB disease, regardless of HIV status. These findings extend the results acquired by the analysis of the adult study in Chapter 3, despite the profound differences in pathophysiology and clinical presentation of the disease between adults and children.

The signatures in the form of the DRS identified the majority of TB culture confirmed cases and were found to be more sensitive than Xpert® MTB/RIF in both the culture confirmed and culture negative TB groups. The DRS identified correctly the majority of other disease cases regardless of their HIV and IGRA status. The high accuracy of the test provides evidence that diagnosis based on a host transcriptomic signature for paediatric TB is feasible. The transformation of multi-transcript signatures into a single value risk score will aid towards translation into clinical practise.

Chapter 5

Conclusions - Discussion

In Chapter 5, I present briefly the achievements and conclusions of my thesis as a whole, by bringing together the results and conclusions from the previous chapters. I also discuss some future research routes along with limitations and lessons that were learnt in this PhD thesis. Potential implications in the field and a personal perspective are also described.

5.1 What this thesis has achieved

During this PhD I reviewed the literature on identification of host biomarkers for infectious diseases and particularly TB disease, using whole blood gene expression data (Chapter 1). The need of deriving gene expression signatures from microarray data in a systematic way was identified, while addressing the particular challenges that underlying co-infections, and multiple-site studies pose. Previous studies have not been focusing on the diagnostic potential that small transcriptomic signatures of disease derived from representative populations may hold. Therefore, I reviewed methods for the statistical analysis of microarray data as well as methods for variable selection (Chapter 2). Then, a pipeline in R was implemented with the capacity to analyse gene expression microarray data and identify small sets of gene expression biomarkers regardless of country of origin and HIV infection, based on elastic net. The models defined by elastic net were subsequently incorporated into a single value disease risk score for every patient, without substantial loss of classificatory performance.

The pipeline was first applied on a whole blood gene expression adult case-control study including patients with TB, LTBI and diseases phenotypically similar to TB and a paediatric cohort study with a similar set-up (Chapters 3, 4). After refining the pipeline to match the different needs for analysis of the two datasets, we found that patients with TB can be distinguished from LTBI individuals by measuring the expression values of 27 transcripts in adults and 42 transcripts in children. Most importantly, the expression values of 44 transcripts in adults and 54 in children were able to discriminate TB from the range of conditions that present with similar clinical features. The signatures were validated using independent datasets and benchmarked against tests that are currently being used in clinical practice. The good classificatory performance of signatures, which was reproducible in both HIV-infected and -uninfected cohorts, in different geographic locations and in independent TB patient datasets suggests that these transcript sets are promising biomarkers of TB. In addition, the small number of transcripts in the signatures and the summarisation of the elastic net model into a single value disease risk score may increase the potential for using transcriptional profiling as a clinical diagnostic tool from a single peripheral blood sample.

To my knowledge, this was the first analysis of host gene expression data to identify small sets of TB biomarkers in adults and children from different locations, while including HIV-infected individuals and controls representing the 'real world' disease spectra from which TB should be clinically differentiated.

5.2 Future Directions

The field of diagnosis of infectious diseases has recently started moving from detecting the infectious agent to discovery of unique pathogen specific identifying host gene expression signatures. This analysis provides proof of principle of the feasibility of developing TB diagnostics based on host gene expression data, using a pipeline for both adult and paediatric populations [176]. In clinical practice, the application of the TB signatures may be distinct as the TB/LTBI signature would be of value in contact screening, where the task is distinguishing active disease from previous exposure in minimally symptomatic individuals. The TB/OD signature would be of most value in evaluating symptomatic patients presenting to medical services with symptoms of TB. The use of these biomarkers in a clinical decision process either as an initial screening tool, or in conjunction with other diagnostic tools should be refined. Studies on prospective populations would be required in which the decisions about whether and when to initiate TB treatment are evaluated against the new biomarkers.

Although a major concern in using transcriptional signatures as a clinical diagnostic tool in resource poor settings is the complexity and cost of the current methodologies, the concept of the DRS could address this. Binding of RNA from a patient's blood to probes complementary to the signatures' transcripts could be detected as a signal from all the up- or down- regulated transcripts combined. Expression of up- or down-regulated transcripts in an individual patient could be compared with that of housekeeping genes which have minimal variation, to allow for normalisation. There are methods that may be suitable for direct analysis of multiple transcript signatures in blood and at a relatively low cost, including lateral flow RT-PCR based systems, nano-pore technology [177], nano-particle enzyme linked detection and detection using nano-wires [178, 179] and electrical

impedance [180]. If signatures are to be translated into simple cheap assays, further evaluation of cost effectiveness and how a test based on host gene expression might be used operationally is needed.

However, the first step is to examine whether the classification is reproducible using RT-PCR analysis. Some improvement in sensitivity and specificity of our DRS may also be achieved by weighting the signal from the most discriminatory transcripts, and this could be explored in subsequent refinements of the model. In terms of diagnostics it would be advantageous if the number of transcripts in the signatures could be reduced further, even with some small loss of sensitivity. Currently, in our group we are testing in-house methods for detecting biomarkers and they may be of use in this dataset.

In addition to the diagnostics perspective, the two gene expression studies can act as an unprecedented resource to explore the biological processes of the response to the infection with *M. tuberculosis* that will enhance our understating about TB disease. They also offer the opportunity to study, compare and contrast paediatric and adult specific responses as well as the differences in the response between HIV-infected and -uninfected individuals.

5.3 Limitations

While this analysis provides proof of principle that relatively small numbers of RNA transcripts can be used to discriminate active TB from latent TB infection and other diseases in Africa, some limitations remain. The blood cell count for the patients was not available and therefore not incorporated in the analysis. This may contain additional complexity for the biological interpretation of the signatures and particularly for the HIV-infected individuals.

Although microarrays have been successfully addressing the identification of differentially expressed genes, they measure gene expression in a targeted way. NGS technologies could have probably allowed for detection of novel sequences and alternative splicing events associated with TB disease and infection. In addition, regardless of the

overlap between the adult and paediatric signatures that has been observed, the gene expression TB signatures are quite different between children and adults. However, a combined analysis of the two datasets and identification of one TB signature independent of age would make the translation of the signatures into a more feasible test.

Apart from the analysis of the microarray data and the required laboratory work, these studies involved large clinical endeavours over several years. The scale of the studies and the differences between the cohorts necessitate a rigorous "gold standard" classification from the clinical teams. A small proportion of false assignment by the current "gold standard" tests is to be expected as noted by post-mortem studies at which undiagnosed TB cases are confirmed [181, 182].

5.4 Closing Remark

Host gene expression signatures hold great potential for diagnosis of infectious diseases and the field is currently moving towards their use as clinical diagnostic tools. Further minimisation of the numbers of the transcripts in the diagnostic signatures and validation using other gene expression measuring techniques will prepare the grounds for diagnostic devices that can transfer the research findings from the bench to the bedside. What also constitutes a real challenge is the identification of systematic ways that can allow us to further our understanding about disease pathogenesis and progression in the host by extrapolating information from the multi-dimensional whole blood gene expression datasets. The approach presented in this thesis, can be extended and applied to other infections, diseases or conditions, which are currently seeking for new diagnostic and tests.

In clinical management of TB patients, a simple transcriptome-based test which reliably diagnoses or excludes TB in the majority of patients undergoing investigation for suspected TB, using a single blood sample, would be of great value. Host gene expression signatures in conjunction with reliable results of existing tests, new vaccines and new drugs will hopefully revolutionise patient management and control of TB disease in high burden countries and worldwide. The challenge for the academic research community and for industry is to develop innovative methods to translate minimal transcript signatures with

maximal diagnostic performance into simple, cheap test suitable for use in African health facilities.

References

1. John Playfair, G.B., *Infection & Immunity*. Fourth Edition ed. 2013: Oxford University Press.
2. Dranoff, G., *Cytokines in cancer pathogenesis and cancer therapy.* Nature reviews Cancer, 2004. **4**(1): p. 11-22.
3. Dulin, D., et al., *Studying genomic processes at the single-molecule level: introducing the tools and applications.* Nat Rev Genet, 2013. **14**(1): p. 9-22.
4. Moss, L., *The question of questions: What is a gene? Comments on Rolston and Griffths & Stotz.* Theoretical Medicine and Bioethics, 2006. **27**(6): p. 523-534.
5. Alberts B, J.A., Lewis J, et al. , *Molecular Biology of the Cell*, G. Science, Editor. 2002: New York.
6. Adams, C.M.O.C.J.U., *Essentials of Cell Biology*. 2010, Cambridge, MA: NPG Education.
7. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
8. Tian, Q., et al., *Integrated genomic and proteomic analyses of gene expression in Mammalian cells.* Mol Cell Proteomics, 2004. **3**(10): p. 960-9.
9. Gygi, S.P., et al., *Correlation between protein and mRNA abundance in yeast.* Molecular and cellular biology, 1999. **19**(3): p. 1720-30.
10. Greenbaum, D., et al., *Comparing protein abundance and mRNA expression levels on a genomic scale.* Genome biology, 2003. **4**(9): p. 117.
11. Lodish, H.F., *Molecular cell biology*. 7th ed. 2013, New York: W.H. Freeman and Co. xxxiii, 1154, [58] p.
12. Stanton, L.W., *Methods to profile gene expression.* Trends Cardiovasc Med, 2001. **11**(2): p. 49-54.
13. Gibson, U.E., C.A. Heid, and P.M. Williams, *A novel method for real time quantitative RT-PCR.* Genome Res, 1996. **6**(10): p. 995-1001.
14. Heid, C.A., et al., *Real time quantitative PCR.* Genome Res, 1996. **6**(10): p. 986-94.
15. Bustin, S.A., *AZ of quantitative PCR*. 2004: International University Line La Jolla.
16. Ekins, R. and F.W. Chu, *Microarrays: their origins and applications.* Trends in Biotechnology, 1999. **17**(6): p. 217-218.
17. Southern, E.M., *Detection of Specific Sequences among DNA Fragments Separated by Gel-Electrophoresis.* Journal of Molecular Biology, 1975. **98**(3): p. 503-&.
18. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
19. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-+.
20. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
21. Schulze, A. and J. Downward, *Navigating gene expression using microarrays - a technology review.* Nature Cell Biology, 2001. **3**(8): p. E190-E195.
22. Conzone, S.D. and C.G. Pantano, *Glass slides to DNA microarrays.* Materials Today, 2004. **7**(3): p. 20-26.
23. Duggan, D.J., et al., *Expression profiling using cDNA microarrays.* Nature Genetics, 1999. **21**: p. 10-14.
24. Trevino, V., F. Falciani, and H.A. Barrera-Saldana, *DNA microarrays: a powerful genomic tool for biomedical and clinical research.* Molecular Medicine, 2007. **13**(9-10): p. 527-541.

25. Altman, N.S. and J. Hua, *Extending the loop design for two-channel microarray experiments.* Genetical Research, 2006. **88**(3): p. 153-163.
26. Marguerat, S. and J. Bahler, *RNA-seq: from technology to biology.* Cellular and Molecular Life Sciences, 2010. **67**(4): p. 569-579.
27. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
28. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nature biotechnology, 2011. **29**(7): p. 644-52.
29. Gilbert, J.A. and M. Hughes, *Gene expression profiling: metatranscriptomics.* Methods in molecular biology (Clifton, N J), 2011. **733**: p. 195-205.
30. Gosalbes, M.J., et al., *Metatranscriptomic approach to analyze the functional human gut microbiota.* PloS one, 2011. **6**(3): p. e17447.
31. Garber, M., et al., *Computational methods for transcriptome annotation and quantification using RNA-seq.* Nature methods, 2011. **8**(6): p. 469-77.
32. Marioni, J.C., et al., *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.* Genome research, 2008. **18**(9): p. 1509-17.
33. Kogenaru, S., et al., *RNA-seq and microarray complement each other in transcriptome profiling.* BMC genomics, 2012. **13**: p. 629.
34. Lipson, D., et al., *Quantification of the yeast transcriptome by single-molecule sequencing.* Nature biotechnology, 2009. **27**(7): p. 652-8.
35. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Human molecular genetics, 2010. **19**(R2): p. R227-40.
36. Murray, D., et al., *In silico gene expression analysis - an overview.* Molecular Cancer, 2007. **6**.
37. Biomarkers Definitions Working, G., *Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.* Clinical pharmacology and therapeutics, 2001. **69**(3): p. 89-95.
38. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
39. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.
40. Kikuchi, T., et al., *Expression profiles of non-small cell lung cancers on cDNA microarrays: identification of genes for prediction of lymph-node metastasis and sensitivity to anti-cancer drugs.* Oncogene, 2003. **22**(14): p. 2192-205.
41. Baehner, F.L., et al., *Genomic Signatures of Cancer: Basis for Individualized Risk Assessment, Selective Staging and Therapy.* Journal of Surgical Oncology, 2011. **103**(6): p. 563-573.
42. Jenner, R.G. and R.A. Young, *Insights into host responses against pathogens from transcriptional profiling.* Nature reviews Microbiology, 2005. **3**(4): p. 281-94.
43. Waddell, S.J., P.D. Butcher, and N.G. Stoker, *RNA profiling in host-pathogen interactions.* Current Opinion in Microbiology, 2007. **10**(3): p. 297-302.
44. Boldrick, J.C., et al., *Stereotyped and specific gene expression programs in human innate immune responses to bacteria.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(2): p. 972-977.
45. Tran, P., et al., *Host's innate immune response to fungal and bacterial agents in vitro: up-regulation of interleukin-15 gene expression resulting in enhanced natural killer cell activity.* Immunology, 2003. **109**(2): p. 263-270.
46. Pathan, N., et al., *Role of interleukin 6 in myocardial dysfunction of meningococcal septic shock.* Lancet, 2004. **363**(9404): p. 203-9.
47. Kim, H.S., et al., *Expression of genes encoding innate host defense molecules in normal human monocytes in response to Candida albicans.* Infection and Immunity, 2005. **73**(6): p. 3714-3724.

48. Li, Q.S., et al., *Microarray Analysis of Lymphatic Tissue Reveals Stage-Specific, Gene Expression Signatures in HIV-1 Infection.* Journal of Immunology, 2009. **183**(3): p. 1975-1982.
49. Liu, M., et al., *Early days: genomics and human responses to infection.* Curr Opin Microbiol, 2006. **9**(3): p. 312-9.
50. Liu, M., et al., *Early days: genomics and human responses to infection.* Current Opinion in Microbiology, 2006. **9**(3): p. 312-319.
51. Zaas, A.K., et al., *Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans.* Cell Host & Microbe, 2009. **6**(3): p. 207-217.
52. Whitney, A.R., et al., *Individuality and variation in gene expression patterns in human blood.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(4): p. 1896-1901.
53. Liew, C.C., et al., *The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool.* Journal of Laboratory and Clinical Medicine, 2006. **147**(3): p. 126-132.
54. Hu, X.R., et al., *Gene expression profiles in febrile children with defined viral and bacterial infection.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(31): p. 12792-12797.
55. Ramilo, O., et al., *Gene expression patterns in blood leukocytes discriminate patients with acute infections.* Blood, 2007. **109**(5): p. 2066-77.
56. Griffiths, M.J., et al., *Genomewide analysis of the host response to malaria in Kenyan children.* J Infect Dis, 2005. **191**(10): p. 1599-611.
57. Loke, P., et al., *Gene Expression Patterns of Dengue Virus-Infected Children from Nicaragua Reveal a Distinct Signature of Increased Metabolism.* Plos Neglected Tropical Diseases, 2010. **4**(6).
58. Ardura, M.I., et al., *Enhanced Monocyte Response and Decreased Central Memory T Cells in Children with Invasive Staphylococcus aureus Infections.* Plos One, 2009. **4**(5).
59. Andersson, H., et al., *Transcriptional profiling of the peripheral blood response during tularemia.* Genes and Immunity, 2006. **7**(6): p. 503-513.
60. Berry, M.P.R., et al., *An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis.* Nature, 2010. **466**(7309): p. 973-U98.
61. Organisation, W.H. *Global Tuberculosis report 2013*. 2013.
62. Smith, I., *Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence.* Clinical Microbiology Reviews, 2003. **16**(3): p. 463-+.
63. Newton, S.M., et al., *Paediatric tuberculosis.* Lancet Infect Dis, 2008. **8**(8): p. 498-510.
64. O'Garra, A., et al., *The immune response in tuberculosis.* Annu Rev Immunol, 2013. **31**: p. 475-527.
65. Kaufmann, S.H. and A.J. McMichael, *Annulling a dangerous liaison: vaccination strategies against AIDS and tuberculosis.* Nat Med, 2005. **11**(4 Suppl): p. S33-44.
66. Gandhi, N.R., et al., *Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa.* Lancet, 2006. **368**(9547): p. 1575-80.
67. Selwyn, P.A., et al., *A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection.* N Engl J Med, 1989. **320**(9): p. 545-50.
68. Organization, W.H., *Roadmap for childhood tuberculosis: towards zero deaths.* Geneva, Switzerland, 2013.
69. Perez-Velez, C.M. and B.J. Marais, *Tuberculosis in children.* N Engl J Med, 2012. **367**(4): p. 348-61.
70. Zar, H.J., T.G. Connell, and M. Nicol, *Diagnosis of pulmonary tuberculosis in children: new advances.* Expert Rev Anti Infect Ther, 2010. **8**(3): p. 277-88.

71. Albert Calmette, A.B., Leopold Négre, *La vaccination préventive contre la tuberculose par le "BCG" (in French).* 1927.
72. Roy, A., et al., *Effect of BCG vaccination against* Mycobacterium tuberculosis *infection in children: systematic review and meta-analysis.* Bmj-British Medical Journal, 2014. **349**.
73. *Diagnostic Standards and Classification of Tuberculosis in Adults and Children. This official statement of the American Thoracic Society and the Centers for Disease Control and Prevention was adopted by the ATS Board of Directors, July 1999. This statement was endorsed by the Council of the Infectious Disease Society of America, September 1999.* American journal of respiratory and critical care medicine, 2000. **161**(4 Pt 1): p. 1376-95.
74. Thye, T., et al., *Common variants at 11p13 are associated with susceptibility to tuberculosis.* Nature genetics, 2012. **44**(3): p. 257-9.
75. Farhat, M., et al., *False-positive tuberculin skin tests: what is the absolute effect of BCG and non-tuberculous mycobacteria?* International Journal of Tuberculosis and Lung Disease, 2006. **10**(11): p. 1192-1204.
76. Pai, M., L.W. Riley, and J.M. Colford, *Interferonn assays-gamma in the immunodiagnosis of tuberculosis: a systematic review.* Lancet Infectious Diseases, 2004. **4**(12): p. 761-776.
77. Flynn, J.L., et al., *An Essential Role for Interferon-Gamma in Resistance to Mycobacterium-Tuberculosis Infection.* Journal of Experimental Medicine, 1993. **178**(6): p. 2249-2254.
78. Pottumarthy, S., et al., *Evaluation of the tuberculin gamma interferon assay: Potential to replace the Mantoux skin test.* Journal of Clinical Microbiology, 1999. **37**(10): p. 3229-3232.
79. Lalvani, A., et al., *Rapid detection of* Mycobacterium tuberculosis *infection by enumeration of antigen-specific T cells.* American Journal of Respiratory and Critical Care Medicine, 2001. **163**(4): p. 824-828.
80. Cole, S.T., et al., *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence (vol 393, pg 537, 1998).* Nature, 1998. **396**(6707): p. 190-198.
81. Boehme, C.C., et al., *Rapid Molecular Detection of Tuberculosis and Rifampin Resistance.* New England Journal of Medicine, 2010. **363**(11): p. 1005-1015.
82. Boehme, C.C., et al., *Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study.* Lancet, 2011. **377**(9776): p. 1495-1505.
83. Lange, C., et al., *Interferon-gamma release assays for the diagnosis of active tuberculosis: sensible or silly?* European Respiratory Journal, 2009. **33**(6): p. 1250-1253.
84. Rangaka, M.X., et al., *Interferon release does not add discriminatory value to smear-negative HIV-tuberculosis algorithms.* European Respiratory Journal, 2012. **39**(1): p. 163-171.
85. Barth, R.E., T. Mudrikova, and A.I.M. Hoepelman, *Interferon-gamma release assays (IGRAs) in high-endemic settings: could they play a role in optimizing global TB diagnostics? Evaluating the possibilities of using IGRAs to diagnose active TB in a rural African setting.* International Journal of Infectious Diseases, 2008. **12**(6): p. E1-E6.
86. Jacobsen, M., et al., *Candidate biomarkers for discrimination between infection and disease caused by Mycobacterium tuberculosis.* Journal of Molecular Medicine-Jmm, 2007. **85**(6): p. 613-621.
87. Mistry, R., et al., *Gene-expression patterns in whole blood identify subjects at risk for recurrent tuberculosis.* Journal of Infectious Diseases, 2007. **195**(3): p. 357-365.
88. Lu, C., et al., *Novel biomarkers distinguishing active tuberculosis from latent infection identified by gene expression profile of peripheral blood mononuclear cells.* PLoS One, 2011. **6**(8): p. e24290.
89. Lesho, E., et al., *Transcriptional responses of host peripheral blood cells to tuberculosis infection.* Tuberculosis (Edinb), 2011. **91**(5): p. 390-9.
90. Maertzdorf, J., et al., *Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis.* PLoS One, 2011. **6**(10): p. e26938.

91. Maertzdorf, J., et al., *Human gene expression profiles of susceptibility and resistance in tuberculosis.* Genes Immun, 2011. **12**(1): p. 15-22.
92. Bloom, C.I., et al., *Detectable Changes in The Blood Transcriptome Are Present after Two Weeks of Antituberculosis Therapy.* Plos One, 2012. **7**(10).
93. Ottenhoff, T.H.M., et al., *Genome-Wide Expression Profiling Identifies Type 1 Interferon Response Pathways in Active Tuberculosis.* Plos One, 2012. **7**(9).
94. Cliff, J.M., et al., *Distinct Phases of Blood Gene Expression Pattern Through Tuberculosis Treatment Reflect Modulation of the Humoral Immune Response.* Journal of Infectious Diseases, 2013. **207**(1): p. 18-29.
95. Koh, G.C.K.W., et al., *Host Responses to Melioidosis and Tuberculosis Are Both Dominated by Interferon-Mediated Signaling.* Plos One, 2013. **8**(1).
96. Maertzdorf, J., et al., *Common patterns and disease-related signatures in tuberculosis and sarcoidosis.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(20): p. 7853-7858.
97. Bloom, C.I., et al., *Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers.* PLoS One, 2013. **8**(8): p. e70630.
98. Sutherland, J.S., et al., *Differential gene expression of activating Fcgamma receptor classifies active tuberculosis regardless of human immunodeficiency virus status or ethnicity.* Clin Microbiol Infect, 2014. **20**(4): p. O230-8.
99. Joosten, S.A., et al., *Identification of biomarkers for tuberculosis disease using a novel dual-color RT-MLPA assay.* Genes and Immunity, 2012. **13**(1): p. 71-82.
100. Dhanasekaran, S., et al., *Identification of biomarkers for Mycobacterium tuberculosis infection and disease in BCG-vaccinated young children in Southern India.* Genes and Immunity, 2013. **14**(6): p. 356-364.
101. Hladik, F. and M.J. McElrath, *Setting the stage: host invasion by HIV.* Nat Rev Immunol, 2008. **8**(6): p. 447-57.
102. 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).
103. Verhagen, L.M., et al., *A predictive signature gene set for discriminating active from latent tuberculosis in Warao Amerindian children.* Bmc Genomics, 2013. **14**.
104. Berry, M.P., et al., *Systems approaches to studying the immune response in tuberculosis.* Curr Opin Immunol, 2013. **25**(5): p. 579-87.
105. King, H.C. and A.A. Sinha, *Gene expression profile analysis by DNA microarrays: promise and pitfalls.* JAMA, 2001. **286**(18): p. 2280-8.
106. Simon, R.M., *Design and Analysis of DNA Microarray Investigations*. 2003: Springer Science & Business Media, 2003.
107. Golub, T.R., et al., *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-537.
108. Schmid, R., et al., *Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3.* Bmc Genomics, 2010. **11**.
109. Du, P., W.A. Kibbe, and S.M. Lin, *lumi: a pipeline for processing Illumina microarray.* Bioinformatics, 2008. **24**(13): p. 1547-1548.
110. Bolstad, B., *Probe Level Quantile Normalization of High Density Oligonucleotide Array Data.* Division of Biostatistics, University of California, Berkeley.
111. Cleveland, W.S., *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association, 1979. **74**(368): p. 829-836.
112. Jolliffe, I.T., *Principal component analysis*. Springer series in statistics. 1986, New York: Springer-Verlag. xiii, 271 p.

113. Koepf, W. and M. Masjed-Jamei, *A generalization of Student's t-distribution from the viewpoint of special functions.* Integral Transforms and Special Functions, 2006. **17**(12): p. 863-875.
114. Hotelling, H., *The generalization of Student's ratio.* Annals of Mathematical Statistics, 1931. **2**(3): p. 360-378.
115. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3(1)**: p. Article 3.
116. Bonferroni, C.E., *Teoria statistica delle classi e calcolo delle probabilita.* Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8, 3--62, 1936.
117. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
118. Saeys, Y., I. Inza, and P. Larranaga, *A review of feature selection techniques in bioinformatics.* Bioinformatics, 2007. **23**(19): p. 2507-17.
119. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 2005. **67**: p. 301-320.
120. Hoerl, A.E. and R.W. Kennard, *Ridge Regression - Biased Estimation for Nonorthogonal Problems.* Technometrics, 1970. **12**(1): p. 55-&.
121. Tibshirani, R., *Regression shrinkage and selection via the Lasso.* Journal of the Royal Statistical Society Series B-Methodological, 1996. **58**(1): p. 267-288.
122. Hui Zou, T.H., *Elasticnet: elastic-net for sparse estimation and sparse PCA.* R package version, 2008.
123. Levin M.; Kaforou M.;Herberg, J.V., VJ.; Coin, LJM.;, *Method for Calculating a Disease Risk Score*, I. Innovations, Editor. 2013.
124. Altman, D.G. and J.M. Bland, *Statistics Notes - Diagnostic-Tests-1 - Sensitivity and Specificity .3.* British Medical Journal, 1994. **308**(6943): p. 1552-1552.
125. Altman, D.G. and J.M. Bland, *Diagnostic-Tests-2 - Predictive Values .4.* British Medical Journal, 1994. **309**(6947): p. 102-102.
126. Altman, D.G. and J.M. Bland, *Diagnostic-Tests-3 - Receiver Operating Characteristic Plots .7.* British Medical Journal, 1994. **309**(6948): p. 188-188.
127. Deeks, J.J. and D.G. Altman, *Statistics notes - Diagnostic tests 4: likelihood ratios.* British Medical Journal, 2004. **329**(7458): p. 168-169.
128. Altman, D.G. and J.M. Bland, *Diagnostic tests 2: Predictive values.* BMJ, 1994. **309**(6947): p. 102.
129. Simel, D.L., G.P. Samsa, and D.B. Matchar, *Likelihood ratios with confidence: sample size estimation for diagnostic test studies.* J Clin Epidemiol, 1991. **44**(8): p. 763-70.
130. Carpenter, J. and J. Bithell, *Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians.* Stat Med, 2000. **19**(9): p. 1141-64.
131. Team, R.D.C., *R: A language and environment for statistical computing*. 2010, R Foundation for Statistical Computing.
132. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biology, 2004. **5**(10).
133. Smyth, G.K., *Limma: linear models for microarray data*, in *Bioinformatics and computational biology solutions using R and Bioconductor*. 2005, Springer. p. 397-420.
134. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010. **33**(1): p. 1-22.
135. Robin, X., et al., *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinformatics, 2011. **12**: p. 77.
136. Ripley, B.D., *Stochastic simulation*. 1987, Wiley & Sons.

137. Cai, Y., et al., *Increased complement C1q level marks active disease in human tuberculosis.* PLoS One, 2014. **9**(3): p. e92340.
138. Cliff, J.M., et al., *Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response.* J Infect Dis, 2013. **207**(1): p. 18-29.
139. Nimmerjahn, F. and J.V. Ravetch, *Fcgamma receptors as regulators of immune responses.* Nat Rev Immunol, 2008. **8**(1): p. 34-47.
140. Borelli, V., et al., *Myeloperoxidase exerts microbicidal activity against Mycobacterium tuberculosis.* Infect Immun, 1999. **67**(8): p. 4149-52.
141. Vestal, D.J. and J.A. Jeyaratnam, *The guanylate-binding proteins: emerging insights into the biochemical properties and functions of this family of large interferon-induced guanosine triphosphatase.* J Interferon Cytokine Res, 2011. **31**(1): p. 89-97.
142. Kim, B.H., et al., *A family of IFN-gamma-inducible 65-kD GTPases protects against bacterial infection.* Science, 2011. **332**(6030): p. 717-21.
143. Yan, Q., et al., *Dusp3 and Psme3 are associated with murine susceptibility to Staphylococcus aureus infection and human sepsis.* PLoS Pathog, 2014. **10**(6): p. e1004149.
144. Pruitt, K.D., et al., *RefSeq: an update on mammalian reference sequences.* Nucleic Acids Res, 2014. **42**(Database issue): p. D756-63.
145. Arlehamn, C.L., et al., *Transcriptional profile of tuberculosis antigen-specific T cells reveals novel multifunctional features.* J Immunol, 2014. **193**(6): p. 2931-40.
146. Lindestam Arlehamn, C.S., et al., *Memory T cells in latent Mycobacterium tuberculosis infection are directed against three antigenic islands and largely contained in a CXCR3+CCR6+ Th1 subset.* PLoS Pathog, 2013. **9**(1): p. e1003130.
147. Slight, S.R., et al., *CXCR5(+) T helper cells mediate protective immunity against tuberculosis.* J Clin Invest, 2013. **123**(2): p. 712-26.
148. Rusinova, I., et al., *Interferome v2.0: an updated database of annotated interferon-regulated genes.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1040-6.
149. Boelaert, J.R., et al., *The effect of the host's iron status on tuberculosis.* J Infect Dis, 2007. **195**(12): p. 1745-53.
150. Ratledge, C., *Iron, mycobacteria and tuberculosis.* Tuberculosis (Edinb), 2004. **84**(1-2): p. 110-30.
151. Harding, C.V. and W.H. Boom, *Regulation of antigen presentation by Mycobacterium tuberculosis: a role for Toll-like receptors.* Nat Rev Microbiol, 2010. **8**(4): p. 296-307.
152. Young, B.L., et al., *The identification of tuberculosis biomarkers in human urine samples.* Eur Respir J, 2014. **43**(6): p. 1719-29.
153. Cliff, J.M., et al., *The human immune response to tuberculosis and its treatment: a view from the blood.* Immunol Rev, 2015. **264**(1): p. 88-102.
154. Hoskyns, W., *Paediatric tuberculosis.* Postgrad Med J, 2003. **79**(931): p. 272-8.
155. Zar, H.J., et al., *Induced sputum versus gastric lavage for microbiological confirmation of pulmonary tuberculosis in infants and young children: a prospective study.* Lancet, 2005. **365**(9454): p. 130-4.
156. World Health Organisation, *Global Tuberculosis Report* 2012.
157. Swingler, G.H., et al., *Diagnostic accuracy of chest radiography in detecting mediastinal lymphadenopathy in suspected pulmonary tuberculosis.* Arch Dis Child, 2005. **90**(11): p. 1153-6.
158. Machingaidze, S., et al., *The utility of an interferon gamma release assay for diagnosis of latent tuberculosis infection and disease in children: a systematic review and meta-analysis.* Pediatr Infect Dis J, 2011. **30**(8): p. 694-700.
159. Mandalakas, A.M., et al., *Interferon-gamma release assays and childhood tuberculosis: systematic review and meta-analysis.* Int J Tuberc Lung Dis, 2011. **15**(8): p. 1018-32.

160. Eamranond, P. and E. Jaramillo, *Tuberculosis in children: reassessing the need for improved diagnosis in global control strategies.* Int J Tuberc Lung Dis, 2001. **5**(7): p. 594-603.
161. Graham, S.M., B.J. Marais, and R.P. Gie, *Clinical features and index of suspicion of tuberculosis in children*, in *Tuberculosis a comprehensive reference*, Z.A. Schaaf HS, Editor. 2009, Saunders Elsevier. p. 154-163.
162. Kampmann, B., et al., *Interferon-gamma release assays do not identify more children with active tuberculosis than the tuberculin skin test.* Eur Respir J, 2009. **33**(6): p. 1374-82.
163. Greco, S., et al., *Current evidence on diagnostic accuracy of commercially based nucleic acid amplification tests for the diagnosis of pulmonary tuberculosis.* Thorax, 2006. **61**(9): p. 783-90.
164. Chintu, C., et al., *Lung diseases at necropsy in African children dying from respiratory illnesses: a descriptive necropsy study.* Lancet, 2002. **360**(9338): p. 985-90.
165. Brent, A.J., *Childhood TB Surveillance: Bridging the Knowledge Gap to Inform Policy.* J Trop Med, 2012. **2012**: p. 865436.
166. Hatherill, M., S. Verver, and H. Mahomed, *Consensus statement on diagnostic end points for infant tuberculosis vaccine trials.* Clin Infect Dis, 2012. **54**(4): p. 493-501.
167. Whittaker, E. and H.J. Zar, *Promising directions in the diagnosis of childhood tuberculosis.* Expert Rev Respir Med, 2012. **6**(4): p. 385-95.
168. *Kenya 2003: results from the Demographic and Health Survey.* Stud Fam Plann, 2005. **36**(2): p. 163-7.
169. *Malawi 2010: results from the demographic and health survey.* Stud Fam Plann, 2012. **43**(2): p. 147-52.
170. South, A., *rworldmap: A New R package for Mapping Global Data.* R Journal, 2011. **3**(1): p. 35-43.
171. Nicol, M.P., et al., *Accuracy of the Xpert MTB/RIF test for the diagnosis of pulmonary tuberculosis in children admitted to hospital in Cape Town, South Africa: a descriptive study.* Lancet Infect Dis, 2011. **11**(11): p. 819-24.
172. Cuevas, L.E., et al., *Evaluation of tuberculosis diagnostics in children: 2. Methodological issues for conducting and reporting research evaluations of tuberculosis diagnostics for intrathoracic tuberculosis in children. Consensus from an expert panel.* J Infect Dis, 2012. **205 Suppl 2**: p. S209-15.
173. Wilson, D., M. Badri, and G. Maartens, *Performance of serum C-reactive protein as a screening test for smear-negative tuberculosis in an ambulatory high HIV prevalence population.* PLoS One, 2011. **6**(1): p. e15248.
174. Li, C.P., et al., *Relationship of HLA-DRB1 gene polymorphism with susceptibility to pulmonary tuberculosis: updated meta-analysis.* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 2015. **19**(7): p. 841-9.
175. Satproedprai, N., et al., *Diagnostic value of blood gene expression signatures in active tuberculosis in Thais: a pilot study.* Genes and immunity, 2015. **16**(4): p. 253-60.
176. Peeling, R.W., P.G. Smith, and P.M. Bossuyt, *A guide for diagnostic evaluations.* Nat Rev Microbiol, 2010. **8**(12 Suppl): p. S2-6.
177. Wang, Y., et al., *Nanopore-based detection of circulating microRNAs in lung cancer patients.* Nat Nanotechnol, 2011. **6**(10): p. 668-74.
178. Laromaine, A., et al., *Protease-triggered dispersion of nanoparticle assemblies.* J Am Chem Soc, 2007. **129**(14): p. 4156-7.
179. Lowe, S.B., et al., *Multiplex sensing of protease and kinase enzyme activity via orthogonal coupling of quantum dot-peptide conjugates.* ACS Nano, 2012. **6**(1): p. 851-7.
180. Morrow, T.J., et al., *Programmed assembly of DNA-coated nanowire devices.* Science, 2009. **323**(5912): p. 352.

181. Cox, J.A., et al., *Autopsy causes of death in HIV-positive individuals in sub-Saharan Africa and correlation with clinical diagnoses.* AIDS Rev, 2010. **12**(4): p. 183-94.
182. Ansari, N.A., et al., *Pathology and causes of death in a group of 128 predominantly HIV-positive patients in Botswana, 1997-1998.* Int J Tuberc Lung Dis, 2002. **6**(1): p. 55-63.

Appendix

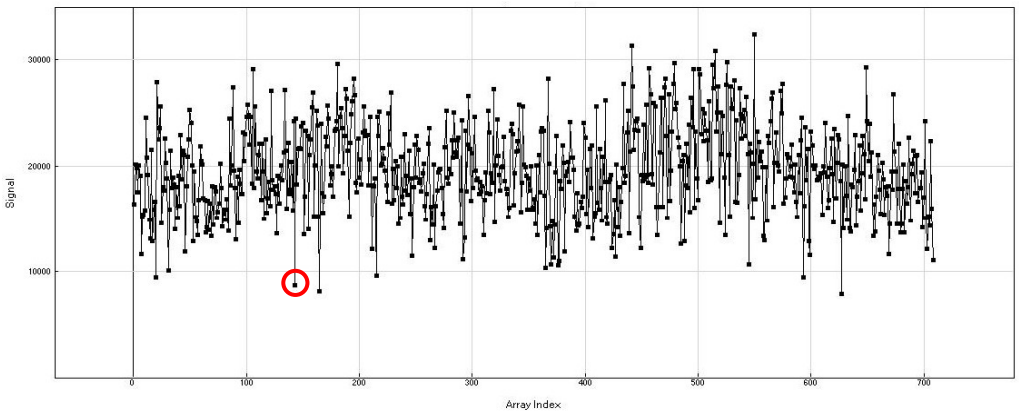
Quality Control of the gene expression microarrays

Using Genome Studio software the microarray images were inspected for artefacts and QC parameters were assessed. Different metrics were used for ascertaining data quality. Illumina BeadChips have internal control probes to monitor data quality at different stages of the experimental process, which are either sample-dependent or sample-independent. Data from the sample-dependent probes are based on metrics from each sample, while the sample-independent probes employ oligonucleotides spiked into the hybridisation buffer that bind to the array during the hybridisation process and allow for a quality overview of this experimental stage. I present here the control summary plots for both the adult and the paediatric study, along with their descriptions.

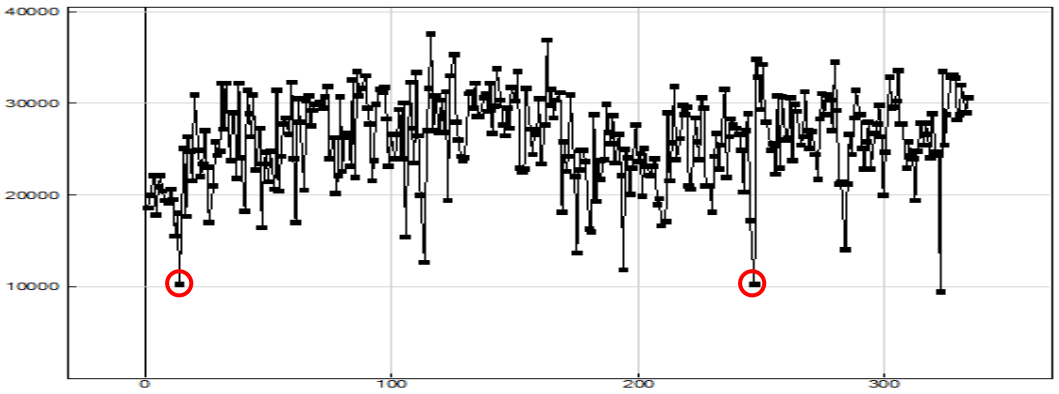
Housekeeping genes

'Housekeeping genes' expression is a sample-dependent measure, and this will vary from sample to sample due to biological variation. The fluctuation in signal intensity observed between the samples in our TB study showed no extreme outliers; however a few arrays had lower intensity values for the HK probes (i.e. <10,000) which indicates there may be a problem with the sample or labelling, so these were further assessed in the PCA data analysis to see if they were also outliers for subsequent removal from the analysis.

Adult study*:



Paediatric study*:

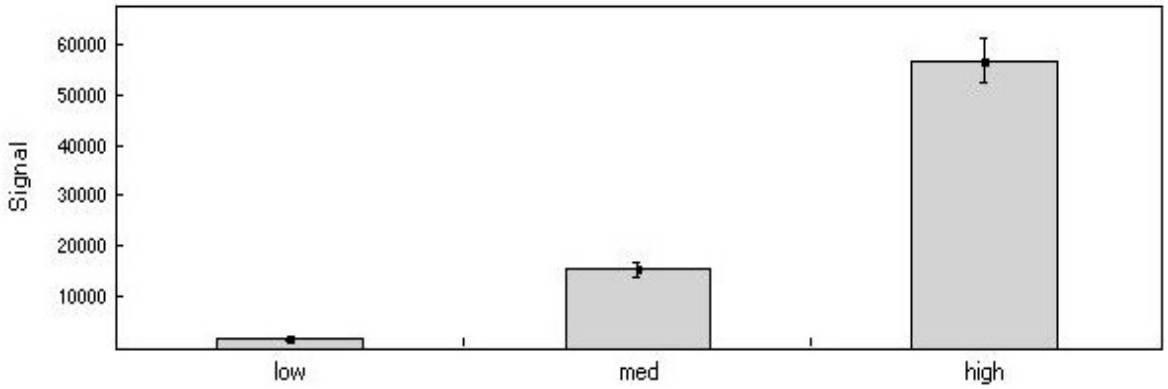


*: samples highlighted in red are the ones excluded in the PCA.

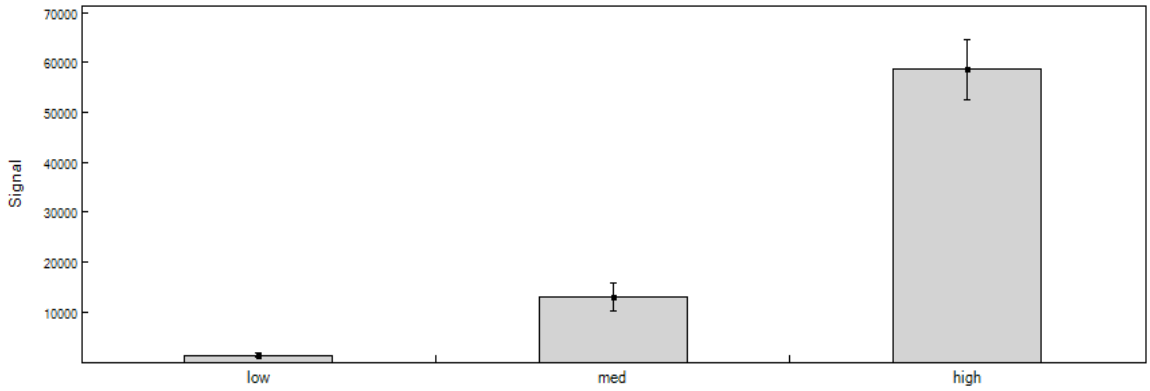
Hybridisation Controls

Hybridisation controls are a sample-independent measure based on spiked oligonucleotides in the hybridisation buffer. These spiked probes are designed to give a range of signal intensities once bound to the array surface and are expected to follow a "low" < "medium" < "high" signal intensity. This was seen in both sets of samples indicating that the hybridisation of the sample cRNA to the probes worked well.

Adult study:



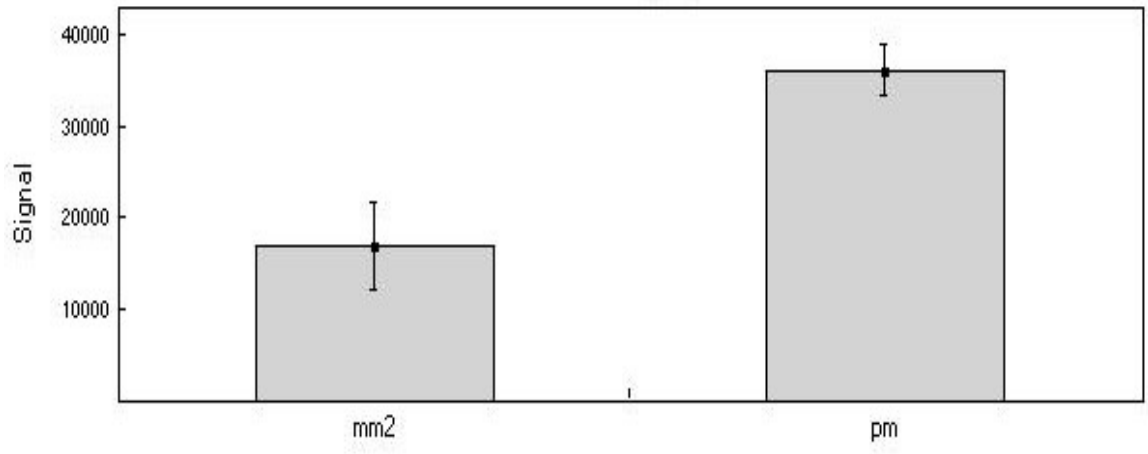
Paediatric study:



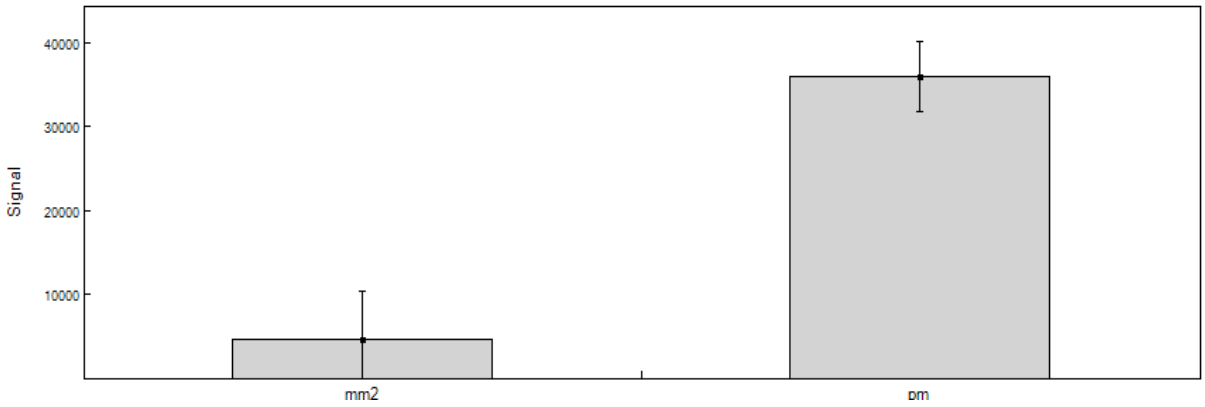
Low Stringency

Low Stringency is a sample-independent measure based on spiked oligonucleotides in the hybridisation buffer. Pairs of probes on the array are present so that one of the pairs will bind to the oligonucleotides in the hybridisation buffer perfectly and the other contains two mis-match bases. During hybridisation, the oligonucleotides will anneal to both the MM2 and PM probes. The oligonucleotides that have not bound as well to the MM2 probe should be washed off during the experiment and this will be reflected in the intensity signal. The Perfect Match (PM) probe signal is expected to be higher than the Mismatch (MM2) probe signal. This was observed indicating that the High Temp wash buffer stage was successful.

Adult study:



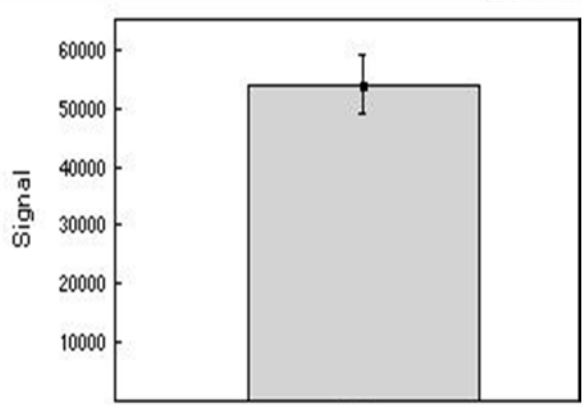
Paediatric study:



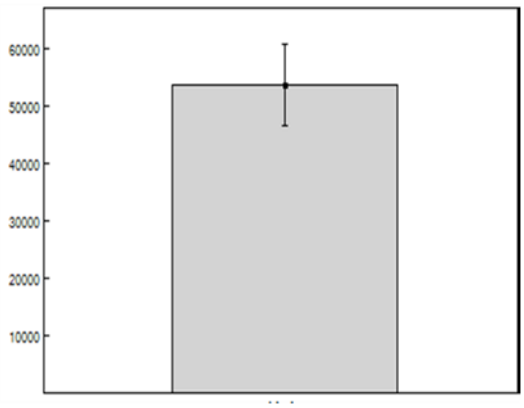
Biotin

An amplified pool of biotin-labelled cRNA is generated from mRNA during the amplification stage of the microarray experimental process. During the hybridisation stage, the incorporated biotin in this cRNA is subsequently bound by streptavidin which is labelled with a fluorescent dye (Cy3). The biotin metric assess how well the biotin-streptavidin/Cy3 binding has worked using control probes that have been labelled with biotin. For both studies, the signal intensity was high, indicating successful labelling during the hybridization stage.

Adult study:



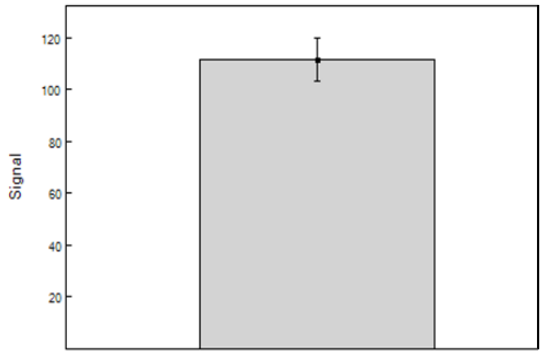
Paediatric study:



Negative controls

These are probes of random sequence selected to have no corresponding targets in the genome. The mean of these signals defines the system background and are expected to have very low signal intensity values. This was seen in both studies.

Adult study:



Paediatric study: