
Discovery and validation of host transcriptomic and proteomic biomarkers for paediatric tuberculosis

Ortensia Vito, MSc., BSc.

Department of Infectious Disease

Faculty of Medicine

Imperial College London

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

April 2023

*To my parents and my sister,
who have always been my wonderful shelter*

ABSTRACT

Although being a curable disease, Tuberculosis (TB) is still one of the leading causes of mortality and morbidity worldwide. Most of the global cases reside in Low- and Middle-Income Countries, where the resources are limited and a high number of patients go undiagnosed or untreated. The existing diagnostic tests for TB are mainly sputum-based, with lengthy turnaround time, low sensitivity, and high cost. Good-quality sputum samples are difficult to collect in children, where invasive procedures, such as induced sputum or gastric aspiration, are needed. Moreover, sputum-based diagnostic tests are suboptimal in patients with Extra-Pulmonary Tuberculosis (EPTB).

The presence of pathogens (i.e., bacteria or viruses) causes changes in gene expression and/or protein abundance detectable in the host's blood. Sparse combinations of biological molecules, which are known as host diagnostic signatures, can be used to distinguish between disease groups of interest and developed into a Point-of-care Test (POCT) for rapid and accurate disease diagnosis. A host signature for TB, especially childhood TB, might be the way towards a low-cost, easy-to-use, fast and blood-based POCT.

In this thesis, I have used host transcriptomic and proteomic profiles to identify sparse diagnostic signatures for paediatric TB. Firstly, I have described a 6-gene signature that can accurately discriminate between children with microbiologically confirmed TB and unlikely TB, with or without Human Immunodeficiency Virus (HIV) infection, with an Area Under the ROC Curve (AUC) of 89.3% [CI_{95%} 82.5%-96.1%], sensitivity of 88.5% [CI_{95%} 78.9%-96.2%] and specificity of 86.0% [CI_{95%} 75.4%-94.7%] in an independent validation set. Secondly, I have described a 4-protein signature that achieved an AUC of 88.7% [CI_{95%} 85.0%-92.4%], sensitivity of 78.5% [CI_{95%} 71.1%-85.6%] and specificity of 86.1% [CI_{95%} 81.3%-90.4%] in an independent validation set. Lastly, I have described a novel multi-omics approach where I have integrated transcriptomic and proteomic datasets to identify a multi-omics signature. The 5-feature signature achieved an AUC of 97.5% [CI_{95%} 93.6%-100.0%], sensitivity of 92.9% [CI_{95%} 78.6%-100.0%] and specificity of 92.3% [CI_{95%} 80.8%-100.0%] in the test set.

In summary, this thesis has demonstrated that statistical and machine learning tools can identify host single-omics and multi-omics signatures, which are accurate enough to offer low-cost, easy-to-use and fast alternative diagnostic approaches for paediatric TB.

TABLE OF CONTENTS

LIST OF FIGURES	9
LIST OF TABLES	13
LIST OF ACRONYMS	15
STATEMENT OF ORIGINALITY	19
COPYRIGHT STATEMENT	20
ACKNOWLEDGMENTS	21
LIST OF PUBLICATIONS	22
1 INTRODUCTION	23
1.1 Overview and chapter aims	23
1.2 Tuberculosis disease	23
1.3 Host immune response to pathogens	25
1.4 Host response to <i>Mycobacterium tuberculosis</i>	26
1.5 From latent infection to disease	28
1.6 Current diagnostic tests for tuberculosis	29
1.7 Paediatric tuberculosis	32
1.8 The central dogma of molecular biology	35
1.9 Biomarkers as diagnostic test	39
1.10 Measuring gene expression	40
1.10.1 Reverse Transcription - quantitative Polymerase Chain Reaction (RT-qPCR)	41
1.10.2 Microarray	43
1.10.3 RNA Sequencing	43
1.10.4 Transcriptomic signatures for paediatric tuberculosis	46
1.11 Measuring protein abundance	52
1.11.1 The Enzyme-Linked Immunosorbent Assay (ELISA)	53
1.11.2 The Luminex® assay	53
1.11.3 The SomaScan® assay	54
1.11.4 Proteomic signatures for paediatric tuberculosis	55
1.12 Multi-omics data integration	59
1.12.1 Overview	59
1.12.2 Multi-omics data	60
1.12.3 Networks in molecular biology	60
1.12.4 Multi-omics data integration tools	61
1.13 Summary and thesis contribution	63
2 MATERIALS AND METHODS	65
2.1 Overview and chapter aims	65
2.2 Clinical REDCap database	65
2.3 Case definitions	66
2.4 Description of the clinical cohorts	67
2.5 Data analysis of biological data	70
2.5.1 RT-qPCR data generation and pre-processing	71
2.5.2 RNA-Sequencing data generation and pre-processing	72
2.5.3 ELISA data generation and pre-processing	75
2.5.4 Luminex® data generation and pre-processing	75
2.5.5 SomaScan® data generation and pre-processing	76
2.6 Quality control and normalisation	76
2.6.1 RT-qPCR data	78
2.6.2 RNA-Seq data	79
2.6.3 ELISA and Luminex® data	80
2.6.4 SomaScan® data	81
2.7 Downstream data analysis	82
2.7.1 Differential expression and differential abundance analysis	82
2.7.2 Clustering	85

2.7.3	Pathway enrichment analysis	85
2.7.4	Feature selection algorithm	85
2.7.5	Signature performance evaluation	86
2.7.6	Cross validation	89
2.8	Multi-omic data integration	89
2.8.1	DIABLO pipeline	91
3	DISCOVERY AND VALIDATION OF A SIX-MARKER TRANSCRIPTOMIC SIGNATURE FOR DIAGNOSIS OF CHILDHOOD TUBERCULOSIS IN AN AFRICAN MULTI-COUNTRY COHORT	92
3.1	Overview and chapter aims	92
3.2	Cohorts and study design	93
3.2.1	Case definitions	94
3.2.2	Data generation	95
3.3	Results	96
3.3.1	Discovery phase	96
3.3.2	Exon analysis and RT-qPCR assay optimisation	102
3.3.3	Cross-platform validation phase	103
3.3.4	Independent validation phase	108
3.3.5	Biology	113
3.4	Summary	117
4	CHANGE IN GENE EXPRESSION PROFILES DURING ANTI-TUBERCULOSIS TREATMENT IN CHILDREN	119
4.1	Overview and chapter aims	119
4.2	Cohorts and study design	120
4.2.1	Case definitions	121
4.2.2	Data generation	121
4.3	Results	122
4.3.1	Data pre-processing	122
4.3.2	Comparison between confirmed TB versus unlikely TB	125
4.3.3	Time course analysis: Comparison between confirmed TB pre-TB treatment versus confirmed TB post-TB treatment	126
4.3.4	Performance of the 6-gene diagnostic signature pre- and post-TB treatment	129
4.4	Summary	132
5	A NOVEL AND PARSIMONIOUS PROTEOMIC SIGNATURE FOR DISTINGUISHING BETWEEN TUBERCULOSIS DISEASE AND OTHER RESPIRATORY INFECTIONS IN CHILDREN	134
5.1	Overview and chapter aims	134
5.2	Cohorts and study design	135
5.2.1	Case definitions	136
5.2.2	Data generation	137
5.3	Results	140
5.3.1	Discovery phase	140
5.3.2	Validation phase	151
5.3.3	Biological functions of the 4 proteins	155
5.4	Summary	156
6	INTEGRATING TRANSCRIPTOMIC AND PROTEOMIC DATASETS TO IDENTIFY MULTI-OMICS DIAGNOSTIC BIOMARKERS FOR PAEDIATRIC TUBERCULOSIS	158
6.1	Overview and chapter aims	158
6.2	Cohorts and study design	159
6.2.1	Case definitions	159
6.2.2	Data generation	161
6.3	Results	161
6.3.1	Transcriptomic data	161
6.3.2	Proteomic data	166
6.3.3	Multi-omics integration	169
6.4	Biological function of the 5 features	172
6.5	Summary	174
7	DISCUSSION AND FUTURE DIRECTIONS	176
7.1	Limitations and future directions	177
7.2	Conclusion	178
	BIBLIOGRAPHY	179
A	CASE DEFINITIONS FOR EXTRAPULMONARY TUBERCULOSIS IN CHILDREN	195

B	SDE GENES IN THE DISCOVERY RNA-SEQ DATA WHEN COMPARING MICROBIOLOGICALLY CONFIRMED TB VERSUS UNLIKELY TB	197
C	SDE GENES IN THE RNA-SEQ DATA WHEN COMPARING MICROBIOLOGICALLY CONFIRMED TB VERSUS UNLIKELY TB AT BASELINE	201
D	SDE GENES IN THE RNA-SEQ DATA WHEN COMPARING MICROBIOLOGICALLY CONFIRMED TB VERSUS UNLIKELY TB AT FOLLOW-UP VISIT	205
E	SDE GENES IN THE RNA-SEQ DATA WHEN COMPARING MICROBIOLOGICALLY CONFIRMED TB AFTER ANTI-TB TREATMENT VERSUS MICROBIOLOGICALLY CONFIRMED TB BEFORE ANTI-TB TREATMENT	206
F	SDE GENES IN THE RNA-SEQ DATA WHEN COMPARING MICROBIOLOGICALLY CONFIRMED TB VERSUS UNLIKELY TB IN THE MULTI-OMICS COHORT	217
G	BEST AVAILABLE TREATMENT STUDY FOR MIS-C (BATS)	226
H	COPYRIGHT PERMISSION INFORMATION	228

LIST OF FIGURES

1.1	Estimated TB incidence in 2021 for countries with at least 100,000 incident cases [2].	24
1.2	Timeline of TB: from <i>Mtb</i> discovery to a global emergency.	25
1.3	Time from infection to T cell response: a comparison between other pathogens and <i>Mtb</i> [28].	27
1.4	The host immune response to <i>Mtb</i> in the lungs [32].	28
1.5	The dynamic spectrum from <i>Mtb</i> exposure to tuberculosis disease in children [83]. . .	33
1.6	The central dogma of molecular biology.	35
1.7	Molecular structure of the Deoxyribonucleic Acid (DNA).	36
1.8	Transcription from Deoxyribonucleic Acid (DNA) to Ribonucleic Acid (RNA).	38
1.9	Translation from messenger RNA (mRNA) to proteins.	38
1.10	Multiple omics data types and approaches for disease diagnostic and characterisation [109].	39
1.11	Polymerase Chain Reaction (PCR).	42
1.12	Overview of the steps involved in DNA microarray experiments [129].	44
1.13	Overview of the steps involved in a RNA-Seq experiment [131].	45
1.14	Overview of the Luminex® assay [170].	54
1.15	Conceptual model of multi-omics and human disease [196].	61
1.16	Example of a multilayer network used in the context of biology [200].	62
1.17	Overview of multi-omics data integration tools [204].	63
2.1	Diagnostic algorithm for case definitions.	67
2.2	Maps of the study population.	68
2.3	Z-score of the children in the study population broken down by country.	69
2.4	HIV status (exposed, infected or uninfected) in the study population broken down by country and disease group.	70
2.5	Data analysis pipeline for biological data.	71
2.6	Theoretical plot of a PCR output curve [218].	72
2.7	RNA-Seq data analysis pipeline.	73
2.8	Theoretical plot of a Luminex® output curve.	76
2.9	Example of a ROC curve.	88
2.10	Cross-validation algorithm.	89
3.1	PCA biplots from PCA performed on the discovery RNA-Seq dataset coloured according to sequencing batch.	97
3.2	PCA biplots from PCA performed on the discovery RNA-Seq dataset coloured according to clinical variables.	98
3.3	Volcano plot showing log ₂ FC and BH-adjusted <i>p</i> -values from DE analysis comparing microbiologically confirmed TB vs. unlikely TB in the discovery RNA-Seq data.	99
3.4	Results of FS-PLS with cross-validation in the discovery RNA-Seq dataset.	100

3.5	Logged expression values of the 6 candidate biomarker genes identified in the discovery RNA-Seq dataset coloured according to disease group.	101
3.6	Boxplot of the wDRS of the 6-gene signature and its ROC curve in the discovery RNA-Seq cohort.	102
3.7	PCA biplots from PCA performed on the cross-platform validation RT-qPCR dataset coloured according to sample plate.	104
3.8	PCA biplots from PCA performed on the cross-platform validation RT-qPCR dataset coloured according to clinical variables.	106
3.9	C_t values of the 6 candidate biomarker genes measured in the cross-platform RT-qPCR dataset coloured according to disease group.	107
3.10	Boxplots of the original wDRS (i.e., statistical model trained on RNA-Seq data), retrained wDRS (i.e., statistical model retrained on RT-qPCR data) and DRS of the 6-gene signature and the correspondent ROC curves in the cross-platform RT-qPCR dataset.	108
3.11	PCA biplots from PCA performed on the independent validation RT-qPCR dataset coloured according to sample plate.	109
3.12	PCA biplots from PCA performed on the independent validation RT-qPCR dataset coloured according to clinical variables.	110
3.13	C_t values of the 6 candidate biomarker genes measured in the independent RT-qPCR dataset coloured according to disease group.	111
3.14	Boxplots of the wDRS and DRS of the 6-gene signature and the correspondent ROC curves in the independent RT-qPCR dataset.	112
3.15	Pathways identified as significantly enriched from the list of SDE genes between microbiologically confirmed TB and unlikely TB in the discovery RNA-Seq dataset when querying the KEGG and Reactome databases.	114
3.16	Pathways identified as significantly enriched from the list of SDE genes between microbiologically confirmed TB and unlikely TB in the discovery RNA-Seq dataset when querying the GO database (subontology = Biological Process; GO:BP).	116
3.17	Heatmap showing clustering of genes (rows) and samples (columns) of the SDE genes between microbiologically confirmed TB and unlikely TB in the discovery RNA-Seq dataset.	117
4.1	PCA biplots from PCA performed on the RNA-Seq dataset coloured according to sample batch.	123
4.2	PCA biplots from PCA performed on the RNA-Seq dataset coloured according to clinical variables.	124
4.3	Volcano plots showing $\log_2FC$ and BH-adjusted p -values from Differential Expression analysis comparing microbiologically confirmed TB vs. unlikely TB in the RNA-Seq data before and during treatment.	125
4.4	Concordance and discordance of the $\log_2FC$ of genes with BH-adjusted p -value <0.05 in microbiologically confirmed TB vs. unlikely TB before treatment against the $\log_2FC$ of genes with BH-adjusted p -value <0.05 in microbiologically confirmed TB vs. unlikely TB during treatment.	126
4.5	Volcano plot showing $\log_2FC$ and BH-adjusted p -values from DE analysis comparing microbiologically confirmed TB during treatment vs. microbiologically confirmed TB before treatment.	127

4.6	Heatmap showing clustering of genes (rows) and samples (columns) of the SDE genes between microbiologically confirmed TB during treatment and microbiologically confirmed TB before treatment.	128
4.7	Pathways over-expressed and under-expressed between microbiologically confirmed TB patients after and before anti-TB treatment.	129
4.8	Logged expression values before treatment of the 6 candidate biomarker genes identified in Chapter 3 coloured according to disease group or time point.	131
4.9	Boxplots of the wDRS of the 6-gene signature identified in Chapter 3 and the correspondent ROC curves when contrasting microbiological confirmed TB vs. unlikely TB before and during treatment.	132
5.1	Timeline of discovery and validation phases of the diagnostic proteomic signature for childhood TB.	135
5.2	Overview of the biomarker discovery and validation phases carried out and described in Chapter 5.	138
5.3	PCA biplots and density plots from PCA performed on the 4 proteins identified in the SELDI dataset and measured by immunoassay coloured accordingly to batch.	142
5.4	Logged abundance values of the 4 protein biomarkers identified in the SomaScan® dataset coloured according to disease group.	143
5.5	Boxplot of the wDRS of the 4-protein SomaScan® signature and its ROC curve.	144
5.6	Overview of the indefinite values (under-the-limit-of-detection, over-the-limit-of-detection and NULL values) in the Luminex® experiment.	146
5.7	PCA biplots from PCA performed on the Luminex® dataset coloured according to sample plate.	147
5.8	PCA biplots from PCA performed on the proteomic discovery dataset coloured according to clinical variables.	149
5.9	Logged abundance values of the 4 protein biomarkers identified in the discovery dataset coloured according to disease group.	150
5.10	Boxplot of the wDRS of the 4-protein signature and its ROC curve in the discovery cohort.	151
5.11	PCA biplots from PCA performed on the proteomic validation dataset coloured according to clinical variables.	152
5.12	Logged abundance values of the 4 protein biomarkers in the validation dataset coloured according to disease group.	153
5.13	Boxplot of the original wDRS, retrained wDRS and the simple DRS of the 4-protein signature and their ROC curves in the validation cohort.	154
6.1	PCA biplots from PCA performed on the RNA-Seq training dataset used in the multi-omics integration analysis coloured according to clinical variables.	163
6.2	Volcano plot showing log ₂ FC and BH-adjusted <i>p</i> -values from Differential Expression (DE) analysis comparing microbiologically confirmed TB vs. unlikely TB in the training set of the RNA-Seq data.	164
6.3	Logged expression values of the 6 candidate biomarker genes in the training set [Panel A] and test set [Panel B] of the RNA-Seq dataset coloured according to disease group. .	165
6.4	Boxplot of the wDRS of the 6-gene signature and its ROC curve in the training and test sets of the RNA-Seq dataset.	166

6.5	Principal Component Analysis (PCA) biplots from PCA performed on the training proteomic dataset used in the multi-omics integration analysis coloured according to clinical variables.	167
6.6	Logged expression values of the 5 candidate biomarkers in the training set [Panel A] and test set [Panel B] of the proteomic dataset coloured according to disease group. . . .	169
6.7	Boxplot of the wDRS of the 5-protein signature and its ROC curve in the training and test sets of the proteomic dataset.	169
6.8	Principal Component Analysis (PCA) biplots from PCA performed on the integrated RNA-Seq and proteomic training datasets coloured according to clinical variables. . . .	170
6.9	Most discriminant genes and proteins in the multi-omics training set.	171
6.10	Boxplot of the wDRS of the 5-feature signature and its ROC curve in the training and test sets of the multi-omics dataset.	172

LIST OF TABLES

1.1	Optimal TPPs for a biomarker-based diagnostic test for TB in adults and children. . . .	29
1.2	Current diagnostic tests for tuberculosis.	30
1.3	Gene expression biomarkers and/or biosignatures for child and adolescent TB published to date. Adapted from [84].	46
1.4	Protein biomarkers and/or biosignatures for child and adolescent TB published to date.	55
3.1	Clinical features of patients included in the discovery RNA-Seq, cross-platform validation RT-qPCR and independent validation RT-qPCR datasets of novel gene biomarkers for distinguishing between TB disease and unlikely TB in children.	94
3.2	$\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity of the 6 candidate gene biomarkers (individually and combined into a signature) identified in the discovery RNA-Seq dataset when contrasting microbiologically confirmed TB vs. unlikely TB. . .	101
3.3	Performance of the transcriptomic signature in the discovery RNA-Seq cohort broken down by clinical subgroups.	102
3.4	$\log_2FC$, BH-adjusted p -value, AUC, mean exon counts in the confirmed TB and unlikely TB groups, GC content and exon length of the 6 exons selected for each of the candidate biomarker genes identified in the discovery RNA-Seq dataset when contrasting microbiologically confirmed TB vs. unlikely TB.	103
3.5	$\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity of the 6 candidate biomarker genes measured using RT-qPCR in the assay optimisation phase.	103
3.6	$\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity of the 6 candidate biomarker genes (individually and combined into a signature) measured using RT-qPCR in the cross-platform validation phase.	107
3.7	$\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity of the 6 candidate biomarker genes (individually and combined into a signature) measured using RT-qPCR in the independent validation phase.	111
3.8	Performance of the transcriptomic signature in clinical subgroups.	112
4.1	Clinical features at enrollment of patients included in the longitudinal study to assess the changes of gene expression profile during treatment in children with TB disease or unlikely TB.	121
4.2	$\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity of the 6 candidate gene identified in Chapter 3 when contrasting microbiologically confirmed TB vs. unlikely TB before treatment, microbiologically confirmed TB vs. unlikely TB during treatment, and microbiologically confirmed TB during treatment vs. microbiologically confirmed TB before treatment.	130
5.1	Clinical features of patients included in the discovery and independent validation proteomic datasets for identifying novel paediatric TB protein biomarkers.	137

5.2	List of the 26 SDA proteins, sorted by descending $\log_2FC$, between microbiologically confirmed TB versus unlikely TB in the discovery SomaScan® dataset.	143
5.3	$\log_2FC$, BH-adjusted p -value and AUC of the 4 protein biomarkers (individually and combined into a signature) identified in the SomaScan® discovery experiment when measured by immunoassays in the same biological samples.	144
5.4	$\log_2FC$, BH-adjusted p -value and AUC of the 4 protein biomarkers (individually and combined into a signature) identified in the SomaScan® dataset when measured by immunoassays in an independent set of samples.	145
5.5	$\log_2FC$, BH-adjusted p -value and AUC of the 11 differentially abundant protein in the merged discovery dataset sorted by increasing BH-adjusted p -value.	150
5.6	$\log_2FC$, BH-adjusted p -value and AUC of the 4 candidate protein biomarkers (individually and combined into a signature) in the validation dataset.	153
5.7	Proteomic signature performance in clinical subgroups.	155
6.1	Clinical features of patients included in the transcriptomic and proteomic datasets, training set, and test set used to identify and test novel multi-omics biomarkers for distinguishing between TB disease and unlikely TB in children.	160
6.2	$\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity of the 6 candidate biomarkers in the training and test sets of the RNA-Seq dataset when contrasting microbiologically confirmed TB vs. unlikely TB.	165
6.3	$\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity of the 5 candidate biomarkers in the training and test sets of the proteomic dataset when contrasting microbiologically confirmed TB vs. unlikely TB.	168
6.4	Transcriptomic, proteomic and multi-omics signatures performance in the test set broken down by clinical subgroups.	173
B.1	SDE genes in the discovery RNA-Seq data when comparing microbiologically confirmed TB vs. unlikely TB.	197
C.1	SDE genes in the RNA-Seq data when comparing microbiologically confirmed TB vs. unlikely TB at baseline.	201
D.1	SDE genes in the RNA-Seq data when comparing microbiologically confirmed TB vs. unlikely TB at baseline.	205
E.1	SDE genes in the RNA-Seq data when comparing microbiologically confirmed TB after anti-TB treatment vs. microbiologically confirmed TB before anti-TB treatment.	206
F.1	SDE genes in the RNA-Seq data when comparing microbiologically confirmed TB vs. unlikely TB in the multi-omics cohort.	217
H.1	Copyright permissions information.	228

LIST OF ACRONYMS

Notation	Description
AECs	Airway Epithelial Cells
ART	AntiRetroviral Therapy
AUC	Area Under the ROC Curve
BAL	BronchoAlveolar Lavage
BCG	Bacillus Calmette–Guérin
BH	Benjamini-Hochberg
bp	base pairs
C_t	Cycle Threshold
cDNA	complementary DNA
CFP-10	Culture Filtrate Protein 10
CI	Confidence Interval
COVID-19	Coronavirus Disease 2019
CPM	Counts Per Million
CSV	Comma-Separated Values
DA	Differential Abundance
dcRT-MLPA	dual color Reverse Transcription Multiplex Ligation dependent Probe Amplification
DCs	Dendritic Cells
DE	Differential Expression
DIABLO	Data Integration Analysis for Biomarker discovery using Latent cOmponents
DNA	Deoxyribonucleic Acid
DRS	Disease Risk Score
DS	Drug-Susceptible
EB	empirical Bayes
ELISA	Enzyme-Linked Immunosorbent Assay
EPTB	Extra-Pulmonary Tuberculosis
ESAT-6	Early Secreted Antigenic Target 6
EU	European Union
FC	Fold Change
FN	False Negative

Notation	Description
FP	False Positive
FS	Forward Selection
FS-PLS	Forward Selection-Partial Least Squares
GC	Guanine-Cytosine
GLM	Generalised Linear Models
GO	Gene Ontology
HIV	Human Immunodeficiency Virus
HPC	High-Performance Computing
IFN- γ	Interferon Gamma
IGRA	Interferon-Gamma Release Assay
IQR	Interquartile range
KEGG	Kyoto Encyclopedia of Genes and Genomes
LAM	Lipoarabinomannan
LAMP	Loop-mediated isothermal Amplification
LASSO	Least Absolute Shrinkage and Selection Operator
LJ	Lowenstein Jensen
LMIC	Low- and Middle-Income Countries
LTBI	Latent Tuberculosis Infection
MDR	Multi-Drug Resistant
MGIT	Mycobacteria Growth Indicator Tube
MHC	Major Histocompatibility Complex
MIQE	Minimum Information for Publication of Quantitative Real-Time PCR Experiments
miRNA	MicroRNA
MLE	Maximum Likelihood Estimation
MRC	Medical Research Council
mRNA	messenger RNA
Mtb	Mycobacterium tuberculosis
MTBC	Mycobacterium tuberculosis complex
NAATs	Nucleic Acid Amplification Tests
ncRNA	noncoding RNA
NIH	National Institutes of Health
NK	Natural Killer

Notation	Description
NTM	Nontuberculous Mycobacteria
PBMC	Peripheral Blood Mononuclear Cells
PC	Principal Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PLS	Partial Least Squares
POCT	Point-of-care Test
PPD	Purified Protein Derivative
PTB	Pulmonary Tuberculosis
QC	Quality Control
REDCap	Research Electronic Data CAPture
RFU	Relative Fluorescence Unit
RIN	RNA Integrity Number
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
ROC	Receiver Operating Characteristic
rRNA	ribosomal RNA
RT-qPCR	Reverse Transcription - quantitative Polymerase Chain Reaction
RTI	Respiratory Tract Infection
SAM	Sequence Alignment Map
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SDA	Significantly Differentially Abundant
SDE	Significantly Differentially Expressed
SELDI	Surface-Enhanced Laser Desorption/Ionization
SELEX	Systematic Evolution of Ligands by EXponential enrichment
sGCCA	sparse Generalized Canonical Correlation Analysis
SOMAmer®	Slow Off-rate Modified Aptamer
STAR	Spliced Transcripts Alignment to a Reference
SVD	Single Value Decomposition
TB	Tuberculosis
TBM	Tuberculosis Meningitis

Notation	Description
TMM	Trimmed Mean of M-values
TN	True Negative
TNF- α	Tumor Necrosis Factor Alpha
TP	True Positive
TPPs	Target Product Profiles
tRNA	transfer RNA
TST	Tuberculin Skin Test
wDRS	weighted Disease Risk Score
WHO	World Health Organization

STATEMENT OF ORIGINALITY

I, Ortensia Vito, certify that this thesis titled “Discovery and validation of host transcriptomic and proteomic biomarkers for paediatric tuberculosis” is solely my own work and it has been done under the supervision of Dr Myrsini Kaforou and Prof Michael Levin. 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. All the outside sources have been appropriately referenced and collaborators cited in the text below.

All the datasets and clinical data were made available to me by my supervisors.

The recruitment of the patients and collection of biological samples were led by Prof Brian Eley (University of Cape Town, Cape Town, South Africa), Dr Suzanne Anderson (University College London, London, UK), Dr Andrew Brent (University of Oxford, Oxford, UK), Prof Beate Kampmann (London School of Hygiene and Tropical Medicine, London, UK) and Prof Heather Zar (University of Cape Town, Cape Town, South Africa).

Giselle d’Souza (Imperial College London, London, UK), Dr Samuel Nichols (Imperial College London, London, UK), Sara Hourmat (Imperial College London, London, UK), Dr Victoria Wright (Imperial College London, London, UK) and Dr Sandra Newton (Imperial College London, London, UK) performed the laboratory work for the RNA Sequencing (RNA-Seq).

Dr Luca Miglietta (Imperial College London, London, UK), Giselle d’Souza (Imperial College London, London, UK) and Dr Jesus Rodriguez Manzano (Imperial College London, London, UK) performed the laboratory work for the Reverse Transcription - quantitative Polymerase Chain Reaction (RT-qPCR) quantification of the gene targets.

Dr Dominic Habgood-Coote (Imperial College London, London, UK) carried out the pre-processing and Quality Control (QC) of the raw RNA-Seq data.

Dr Shea Hamilton (Imperial College London, London, UK), Dr Marçia Ivonne Peña Paz (Imperial College London, London, UK), Candice Snyders (Stellenbosch University, Cape Town, South Africa), Prof Gerhald Walzl (Stellenbosch University, Cape Town, South Africa) and Prof Novel N. Chegou (Stellenbosch University, Cape Town, South Africa) led and carried out the laboratory elements for the proteomic data generation.

Dr Clive Hoggart (Icahn School of Medicine, New York, USA) analysed the SELDI data and identified the best protein signature to distinguish TB from unlikely TB in children.

Prof Brian Eley (University of Cape Town, Cape Town, South Africa) wrote the case definition for Extra-Pulmonary Tuberculosis (EPTB) diagnosis described in Appendix A.

Forward Selection-Partial Least Squares (FS-PLS) was developed by Prof Lachlan Coin (Peter Doherty Institute, Melbourne, Australia).

Ortensia Vito

London, April 2023

COPYRIGHT STATEMENT

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

ACKNOWLEDGMENTS

Many are the people I would like to thank for making this thesis possible. First and foremost I would like to thank my main supervisor, Dr Myrsini Kaforou, for her continuous support, invaluable guidance and patience during my PhD. Myrsini, you have always believed in me more than I have believed in myself, you taught me so much and I owe you a very special thanks. You have been a friend to me, someone I could always count on and I feel very lucky to have worked with you. Many thanks to my other supervisor, Prof Michael Levin, for always being so passionate about research and for transferring this passion to me. Mike, I will always be grateful for your invaluable help and constant support.

I would like to thank the extraordinary people in the Section of Paediatric Infection Disease at Imperial College. I have been so lucky to work with you. You have made me feel welcome since the day I joined the Levin Research Group. A special thanks to Jane Webb, whose support and help are crucial for everyone in the Section. Jane, the Section would fall apart without you! Thanks to all the members of the Bioinformatics group for our weekly meetings, for your input and feedback, for the very stimulating discussions and for accelerating my learning curve in bioinformatics.

A big thank you goes to Prof Brian Eley, Prof Gerhald Walzl, Dr Suzanne Anderson, Dr Andrew Brent, Prof Beate Kampmann, Prof Heather Zar and everyone in the ILULU and TB-NIH consortia for allowing me to use your data. Thanks to Dr Myrsini Kaforou, Prof Michael Levin, Dr Shea Hamilton, Dr Heather Jackson and Dr Ivana Pennisi for reading sections of my thesis and for sharing with me your invaluable feedback.

I have been extremely lucky to have met fantastic people throughout my PhD, people whose friendship has been and still is so important to me. Thanks to Ivonne Peña Paz and Kay To for making my first years at Imperial full of laughs and fun. Thanks to Ivana Pennisi and Shea Hamilton for always listening and being there for me every time I needed a break or a laugh. You are among the funniest people I have ever met, and the best travel buddies ever. I cannot explain how much your friendship means to me. Thanks to Clare Andradi-Brown, Becca Womersley, Millie Gladstone, Stefan Ebmeier and Rachel Thompson. Thanks for our coffee breaks, our lunches and our chats.

I would like to thank my old friends, who I grew up with, who know me well and who have been there in every single phase of my life. Most importantly, I would like to thank my friends in London, who have seen the ups and downs of my PhD journey. Thanks for being understanding, for your patience and for always bringing a smile to my face when I needed it the most.

Above all, I would like to thank my wonderful family who has always encouraged me to explore the world and follow my dreams. To my parents and my sister, thank you for your unconditional love, for always supporting me in every way possible and for believing in me. You are the people I admire the most. This thesis is for you.

LIST OF PUBLICATIONS

Channon-Wells S.*, **Vito O.***, McArdle A.J.*, Seaby E.G., Patel H., Shah P., Pazukhina E., Wilson C., Broderick C., D'Souza G., Keren I., Nijman R., Tremoulet A.H., Munblit D., Ulloa-Gutierrez R., Carter M.J., Ramnarayan P., De T., Hoggart C., Whittaker E., Herberg J.A., Kaforou M., Cunningham A.J., Blyuss O., Levin M. on behalf of the BATS Consortium. Immunoglobulin, glucocorticoid, or combination therapy for multisystem inflammatory syndrome in children: a propensity-weighted cohort study. *The Lancet Rheumatology* 2023; 2665-9913

Jackson H., Rivero Calle I., Broderick C., Habgood-Coote D., d'Souza G., Nichols S., **Vito O.**, Gómez-Rial J., Rivero-Valesco C., Rodríguez-Núñez N., Barbeito-Castiñeiras G., Pérez-Freixo H., Acosta M. B-D., Cunningham A.J., Herberg J., Wright V.J., Gómez-Carballa A., Salas A., Levin M., Martinon-Torres F., Kaforou M. Characterisation of the blood RNA host response underpinning severity in COVID-19 patients. *Scientific Reports* 2022;12, 12216

Kaforou M., Broderick C., **Vito O.**, Levin M., Scriba T.J., Seddon J.A. Transcriptomics for child and adolescent tuberculosis. *Immunological Reviews* 2022;309(1):97-122

McArdle A.J.*, **Vito O.***, Patel H., Seaby E.G., Shah P., Wilson C., Broderick C., Nijman R., Tremoulet A.H., Munblit D., Ulloa-Gutierrez R., Carter M.J., De T., Hoggart C., Whittaker E., Herberg J.A., Kaforou M., Cunningham A.J., Levin M. for the BATS Consortium. Treatment of multisystem inflammatory syndrome in children. *the New England Journal of Medicine* 2021;385:11-22

Ouldali N.*, Son M.B.F.*, McArdle A.J.*, **Vito O.**, Vaugon E., Belot A. Leblanc C., Murray N.L., Patel M.M., Levin M., Randolph A.G., Angoulvant F for the BATS consortium, the Overcoming COVID-19 Investigators, and the French Covid-19 Pediatric Inflammation Consortium and Pandor study group. Immunomodulatory therapy for MIS-C: An Individual Patient Meta-Analysis. *Submitted to and accepted for publication by PEDIATRICS*

* Joint first author

Most of my publications are related to Coronavirus Disease 2019 (COVID-19) since I worked on addressing key questions about the diagnostics and management of COVID-19 and new syndromes associated with it during the pandemic.

CHAPTER 1

INTRODUCTION

1.1 Overview and chapter aims

Chapter 1 sets the background for this thesis, describing the context of my research and its relevance. It gives an overview of the global burden of Tuberculosis (TB) disease, with its main focus on paediatric TB, and of the World Health Organization (WHO) guidance to fight this infection. Sections 1.2, 1.3 and 1.4 outline respectively the biological aspects of TB, the host immune response to pathogens in general, and to the TB causative agent, *Mycobacterium tuberculosis* (*Mtb*). They are followed by Section 1.5, which outlines the factors that increase the risk of TB progression from an asymptomatic to a symptomatic state. Sections 1.6, 1.7, 1.8 and 1.9 describe the diagnostic tests currently used for TB, their advantages and limitations, and how host biomarkers could improve the diagnosis and reduce the global burden of this infectious disease, especially in children. They are followed by Sections 1.10 and 1.11, where I describe the various techniques used to detect and quantify the expression of genes and abundance of proteins respectively. Finally, Section 1.12 describes a novel way of combining transcriptomic and proteomic markers to achieve even better diagnostic performance with a multi-omics diagnostic signature.

1.2 Tuberculosis disease

Although TB is a largely curable disease, it is still one of the major causes of mortality and morbidity worldwide and the second leading cause of human death from a single infectious disease agent after Coronavirus Disease 2019 (COVID-19) [1]. In 2021, there were 10.6 million new TB cases worldwide and 1.6 million deaths, with the highest incidence in Low- and Middle-Income Countries (LMIC) [2] [Figure 1.1].

TB is an airborne infectious disease caused by the acid-fast bacillus *Mtb*, which is transmitted by inhalation of infectious particles released by coughing or sneezing from an infected and symptomatic person [3]. *Mtb* is primarily a pulmonary pathogen that, occasionally, can spread to other tissues and organs, such as lymph nodes, meninges and bones, causing Extra-Pulmonary Tuberculosis (EPTB). Individuals who inhale *Mtb* develop an immune response in the lungs that can either completely clear the bacteria or inhibit its growth. In the latter case, the bacteria become dormant and the person can remain asymptomatic with what is referred to as *Mtb* infection (sometimes termed Latent Tuberculosis Infection (LTBI)) [4], which affects one-fourth of the world population [5]. Out of all the people latently infected with TB, about 5%-10% will progress to the symptomatic TB disease years to decades after the primary infection [6].

This slow-growing and intracellular pathogenic bacteria was one of the earliest widespread human pathogens [7] and was first identified in 1882 by the German microbiologist Robert Koch [8] [Figure 1.2] following a TB epidemic that killed a quarter of the adult population worldwide. Driven by a search for

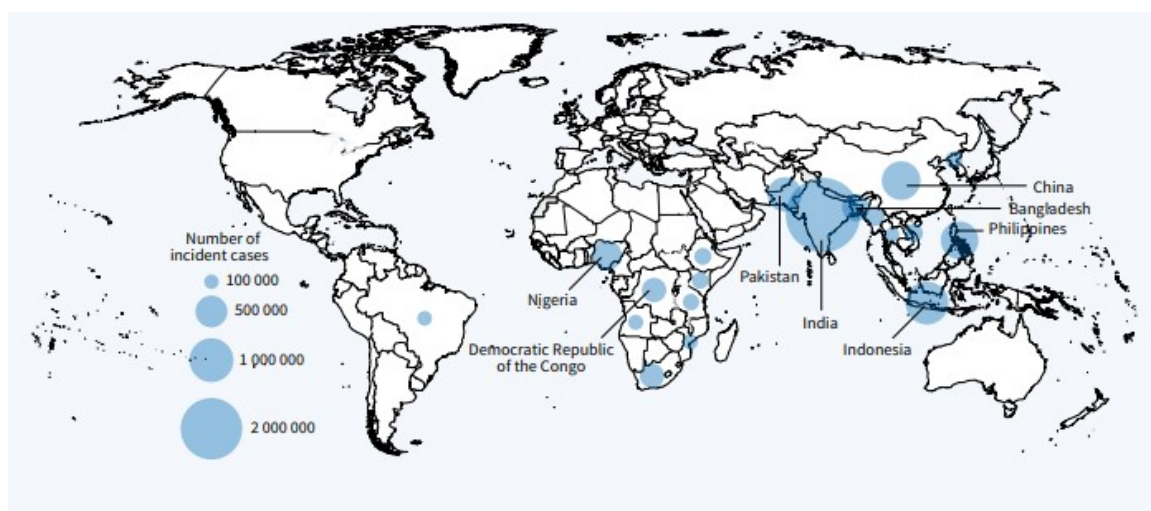


Figure 1.1: Estimated TB incidence in 2020 for countries with at least 100,000 incident cases [2]. The labelled countries have the highest incidence of TB cases worldwide and accounted for two-thirds of the global cases in 2021.

a cure, Koch extracted tuberculin, which contains several specific products from the tubercle bacillus, from cultures of *Mtb* [7]. Tuberculin did not cure TB but was the foundation of the first diagnostic test for this disease developed by Charles Mantoux in 1908. The first, and still widely used and only licensed [9] vaccine for TB, the Bacillus Calmette–Guérin (BCG) vaccine, was invented by and named after Albert Calmette and Camille Guérin in Lille (France). Through a process called passaging (repeated subculture), the *Mycobacterium bovis* bacterium mutated into a weakened strain (BCG), that was first administered to humans in 1921. Since its discovery, the BCG vaccine has been included in the childhood vaccination programmes of many countries, however, its efficacy in preventing TB disease varies depending on age, country and site of TB [10]. It was almost 40 years after the discovery of the BCG vaccine, in 1943, that the first effective antituberculosis agent, streptomycin, was isolated in the laboratory of Selman Waksman at Rutgers University (New Jersey, U.S.A.). In the following decades, new antituberculosis drugs (i.e., thiacetazone, para-aminosalicylic acid, isoniazid, pyrazinamide, cycloserine, ethionamide, rifampin, ethambutol etc.) came on the market [8]. The advent of every new drug led to the selection of mutations conferring resistance to them and developing Multi-Drug Resistant (MDR)-TB, which was and still is much harder to cure.

Drug-Susceptible (DS)-TB is, in fact, usually treated with a 2-month regimen of rifampicin, isoniazid, and pyrazinamide, followed by a 4-month regimen of rifampicin and isoniazid [11], whereas MDR-TB needs a longer regimen (up to 20 months) which can cause severe side effects. Recent trials have shown that specific subgroups of patients may benefit from a shorter regimen. In patients with DS-TB, less than 6 months of treatment might have a similar success rate [12, 13]. In May 2022, the WHO recommended that eligible people over 12 years old with pulmonary DS-TB can receive a 4-month regimen [14]. A shorter regimen was also enforced by the WHO [15] for paediatric TB after the SHINE trial [16]. The SHINE trial showed that 4 months was non-inferior to 6 months of treatment in children with non-severe and smear-negative DS-TB. MDR-TB requires a longer treatment, usually from 18 to 20 months, to be cured [11]. A few studies were successful in the challenge of assessing whether a shorter treatment could have a similar outcome than the 18-20 month treatment in patients with MDR-TB [17–19]. In 2022 the WHO recommended shorter regimens for subgroups of patients with MDR-TB [20]. The last 2 decades have seen major changes in the management of TB, especially in treatment

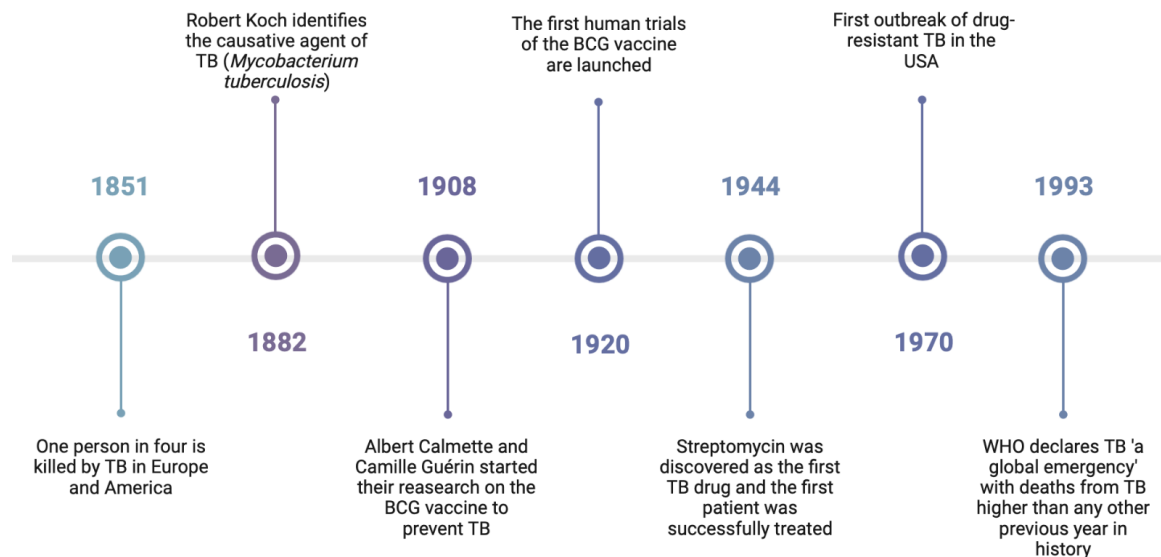


Figure 1.2: Timeline of TB: from *Mtb* discovery to a global emergency.

An outbreak of TB in 1851 killed one person in four in Europe and America. More than 30 years later, in 1882, the pathogen causing this disease was identified by Robert Koch. In the following years, a vaccine was developed and drugs to treat TB were discovered. Drug-resistant strains of TB led to the declaration of a global emergency by the WHO in 1993.

shortening with new drug combinations.

1.3 Host immune response to pathogens

The immune system is a complex network of biological processes that protects an organism from pathogens and diseases. It is composed of two main parts: the innate and adaptive immune systems. The innate immune system is the first to respond when a pathogen enters the organism's body, and it is made up of physical barriers, such as skin and epithelial layers, the eye's cornea and mucous membranes, along with many types of white blood cells, namely Natural Killer (NK) cells, granulocytes (i.e., neutrophils) and monocytes (i.e., Dendritic Cells (DCs) and macrophages). It is triggered when pathogens are identified by pattern recognition receptors [21]. However, it is non-specific, it does not confer long-lasting immunity against a pathogen and hence, it might be inadequate for the clearance of certain organisms under certain conditions. The adaptive immune system is triggered by the innate immune system with its response being slower (usually takes between 4-7 days), antigen-specific, and involves a type of white blood cells called lymphocytes (i.e., B cells and T cells) [22]. B cells are involved in the humoral immune response, or antibody-mediated immunity, whereas T cells are involved in cell-mediated immune response, where phagocytes, CD4⁺ T cells (helper T cells), CD8⁺ T cells (cytotoxic T cells), or various cytokines are activated in response to an antigen. When B cells and T cells are activated and begin to replicate, some of their offspring become long-lived memory cells. These memory cells remember each specific antigen encountered and can mount a strong response if the pathogen is detected again [23]. The immunological memory can either be short-term when immunity is not lifelong and the host might be re-infected by the same pathogen, or long-term memory when the adaptive immunity against that specific pathogen lasts a lifetime.

The host-pathogen interaction differs based on the type of foreign pathogen that crosses the host's physical barriers, in fact, a pathogen uses specific mechanisms to establish the infection in the host

and, in response, the host immune system uses specific mechanisms to clear that specific pathogen. For example, a virus, which can only live and propagate inside the host cells and hijack the host cell's biosynthetic machinery to enable the translation of viral proteins and the replication of the viral genome, can be difficult to detect by the host immune system. The Major Histocompatibility Complex (MHC) class I proteins play an important role in alerting the immune system about virally infected cells. MHC class I molecules are, in fact, expressed on the cell surface and present peptide fragments derived from intracellular proteins, including peptides derived from viral proteins when the cell is virally infected. CD8⁺ T cells monitor cell surface MHC class I molecules for peptides derived from viral proteins and eliminate infected cells. CD8⁺ T cells are therefore particularly important for intracellular pathogen recognition and elimination, whilst CD4⁺ T cells are involved in triggering the adaptive immune response and extracellular antigen recognition [24]. The host immune response is reinforced by the infected cells releasing interferons to limit the replication of viruses and alert proximal cells of the presence of the foreign pathogen.

If the immune system fails to eliminate a pathogen, this can multiply and survive within the host. Several factors can affect the host's immune response, including age, sex, infection history, genetics, and the pathogen's ability to survive within the host's body. Pathogens have evolved complex and efficient methods to mount an effective anti-immune response within the infected host and to overcome the innate and adaptive immune response causing diseases [25].

1.4 Host response to *Mycobacterium tuberculosis*

Once *Mtb* starts proliferating and propagating within the host, the infection develops into a disease with accompanying symptoms and clinical signs. The progression from TB infection to disease may be influenced by various factors, including the type of bacterial strain, the extent of exposure, co-infections, malnutrition, the host's immune response and treatment for autoimmune diseases [6]. The protective and pathogenic response to *Mtb*, which is a complex and multifaceted process, and the result of the interplay between bacterial virulence and host resistance is not yet fully understood. This lack of knowledge may be a significant challenge in developing an effective vaccine. Both innate and adaptive immune systems are involved in the fight against *Mtb*, with the outcomes of *Mtb* infection depending on the sum of the two immune system responses rather than on one particular cell type or function [26]. The adaptive immune response to *Mtb* is delayed compared to other infections [Figure 1.3]. This is likely due to the relatively small number of bacilli entering the host and their slow growth rate, which cause a limited production of antigens difficult to be detected by the host immune system in an early infection stage [6]. *Mtb* has also developed some evasion strategies to evade the host innate immune system. Manipulation of macrophage death pathways is one of the complex mechanisms used by *Mtb* to evade these host defences [27]. The bacilli inhibit apoptosis and induce necrosis of infected host immune cells, such as neutrophils and macrophages, promoting bacterial growth and migration [27].

The infection starts in the lung, where aerosol droplets infected with *Mtb* are deposited in the lower airways and the terminal alveolar tissue [Figure 1.4], after traversing the physical and anatomical barriers encountered. The Airway Epithelial Cells (AECs) are the first cells to come in contact with the bacteria. Even if their main role is to act as physical barriers, they might be considered “non-professional” immune cells [29]. The AECs, in fact, after perceiving the presence of *Mtb* through pattern recognition receptors, improve their antimicrobial capacity by secreting various microorganism-killing effectors

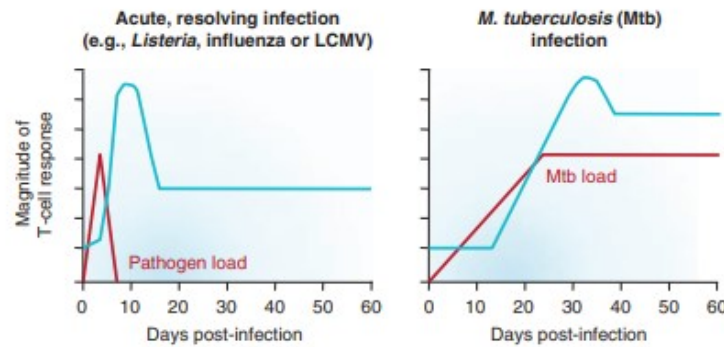


Figure 1.3: Time from infection to T cell response: a comparison between other pathogens and *Mtb* [28]. The host immune response to *Mtb* is delayed compared to other infections (i.e., listeria, influenza or lymphocytic choriomeningitis virus). T cell response starts usually 3-5 days after infection, with its peak around 7-8 days post-infection followed by established memory immune cell populations. With *Mtb* infection, T cell response initiates 8-10 days after infection, it peaks after several weeks and the bacteria is not usually cleared completely.

into the airway lumen to control the composition of the airway surface liquid [30]. The bacteria are then engulfed through phagocytosis by alveolar macrophages in the airways, where the host mounts an immune response. Neutrophils, macrophages and lung DCs are attracted to the site of infection probably by chemokine and cytokine signals from the infected cells [6], and surround the *Mtb*-infected immune cells starting the formation of a dynamic and spatially organised structure called a granuloma, the most characteristic lesion and hallmark structure of TB. Subsequently, T cells migrate to the site of infection and surround the granuloma, which is now formed by a necrotic centre surrounded by neutrophils, a macrophage layer and lymphocytes [6]. Each granuloma might be initiated by a single bacillus and reach 10^5 bacilli in its centre in 4-6 weeks [6]. At this point, the granuloma can inhibit the growth of the bacilli or the bacilli can escape and form new granulomas in the lung or other tissues, such as lymph nodes. The granuloma contains the bacteria preventing its spread within the host, but it has been shown that it can also provide a favourable environment and physical barrier that protect the pathogen from the immune system and might facilitate its survival into the host in a latent and asymptomatic form [31].

The adaptive immune response starts after 8-10 days when some of the *Mtb*-infected DCs migrate to the local lung-draining lymph nodes and activate the naïve T cell response. Naïve T cells are mature T cells that have not encountered their corresponding antigens yet. When they encounter the *Mtb*-infected DCs in the local lung-draining lymph nodes, their differentiation is driven toward a Th1 phenotype (i.e., $CD4^+$ T cells) [32]. $CD4^+$ T cells, which produce a proinflammatory response for killing intracellular pathogens, migrate back to the lung 14-17 days post-infection to enhance the activation of macrophages [33] and to contain the initial phase of bacterial multiplication by producing cytokines, such as Interferon Gamma ($IFN-\gamma$) and Tumor Necrosis Factor Alpha ($TNF-\alpha$), which have been shown to be protective against *Mtb* [32]. They might also be responsible for inducing $CD8^+$ T cell or B cell response, which might play a role in the prevention of TB and in producing migration signal for other cells [6]. $CD8^+$ T cells, which play an important role against intracellular pathogens, might recognise and destroy *Mtb*-infected cells [6]. At this point, the host can either kill the pathogen, or the bacteria can begin to multiply and grow, causing symptomatic TB disease. The infection can also become dormant and asymptomatic (*Mtb* infection) and eventually reactivate over a lifetime [34].

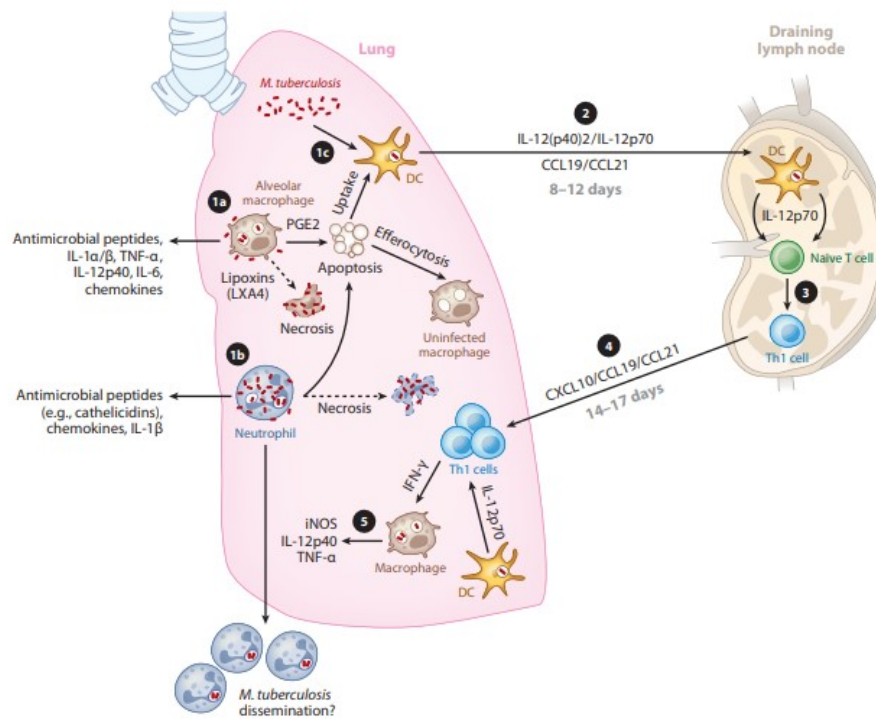


Figure 1.4: The host immune response to *Mtb* in the lungs [32].

The route of entry of *Mtb* into the body is via the respiratory tract. The pathogen gets into the lungs and is deposited in the lower airways and the alveolar tissue. It then reaches the local lung-draining lymph nodes and starts disseminating after 8-12 days from infection. This coincides with the activation of the naïve T cells, which differentiate toward a Th1 phenotype and migrate back to the lungs to activate an inflammatory response. The host immune response takes 18-20 days to reach an effective level and stop bacterial growth.

1.5 From latent infection to disease

One-third of the global population is latently infected with *Mtb* [3], and about 5%-10% progress from *Mtb* infection to TB disease in their life [35]. Changes in the balance between the host immune status and the bacterial pathogenicity can cause *Mtb* infection to become active at any time. This balance is influenced by several factors including the infectious dose, virulence, persistence of the pathogen, host health, co-morbidities, and the interplay between the innate and adaptive immune system [36]. Examples of risk factors are discussed below.

Co-infection with Human Immunodeficiency Virus (HIV) is one of the most well-known risk factors for progression from *Mtb* infection to TB disease [37] since it affects the host immune response to *Mtb*. Even if the activation of both CD4⁺ and CD8⁺ T cells has been seen in humans with TB disease, CD4⁺ T cells seem to be the most critical [38]. In response to the mycobacterial antigens, they produce IFN- γ which activates macrophages to kill the intracellular bacteria and control the infection [39, 40]. People living with HIV infection have a depletion of CD4⁺ T cells caused by the virus, and this leads to an increased risk of progression to TB disease. It has also been shown that HIV impairs TNF- α -mediated macrophage apoptotic response to *Mtb* and therefore facilitate bacterial survival [41].

Severe malnutrition is another risk factor for progressing from TB disease to *Mtb* infection because it causes a lack of nutrients, such as vitamins, minerals and trace elements, that boost the immune response against intracellular pathogens like *Mtb* [42].

Young age is an additional risk factor for developing TB disease. The global cases of paediatric TB were

aproximately 1.6 million in 2022, with over 200,000 deaths [2, 43]. However, it is very likely that the actual childhood TB cases and deaths are much higher than the estimated numbers since many cases go undiagnosed or unreported, especially in LMIC countries [1]. Diagnosis of TB disease is also more difficult in children compared to adults. Paediatric tuberculosis and its diagnosis will be discussed in more detail in Section 1.7.

1.6 Current diagnostic tests for tuberculosis

One of the major barriers to controlling and eliminating TB is the lack of accurate testing procedures globally. To improve the global TB testing, the WHO published the Target Product Profiles (TPPs) for an ideal Point-of-care Test (POCT) for TB in 2014 [44]. A POCT is a simple and rapid diagnostic test usually undertaken at the patient's bedside by non-laboratory trained personnel and with a very brief sample-to-result turnaround time. The WHO aim is to have accessible and affordable POCTs for peripheral centres of developing countries, where TB is endemic. Many of the current diagnostic tests are in fact restricted to high-infrastructure and central laboratories causing a delay in the diagnosis of TB, as the biological samples need to be transported before being processed and analysed. The TPPs delineated by the WHO concern four different tests needed for managing cases or suspected cases of TB: a sputum-based POCT, a non-sputum biomarker-based POCT, a triage test to screen patients with suspected TB and reduce the number that needs confirmatory tests, and a rapid drug-susceptibility test [44].

Based on the TPPs published by the WHO in 2014 [44], an ideal POCT for TB disease should simultaneously: have high sensitivity and high specificity; have treatment monitoring capability; detect *Mtb* in less than 20 minutes; test drug susceptibility in less than 30 minutes; and require no or minimal sample processing. Table 1.1 describes the optimal sensitivity and specificity for a biomarker-based diagnostic test for TB in adults and children.

Diagnostic measure	Clinical group	Optimal target
Adults		
Diagnostic sensitivity	PTB	Culture positive and smear positive ≥98%
		Culture positive and smear negative ≥68%
	PTB and HIV-infected	≥80%
	EPTB	≥80%
Diagnostic specificity	PTB or EPTB	≥98%
Children		
Diagnostic sensitivity	PTB	≥66%
Diagnostic specificity	PTB or EPTB	≥98%

Table 1.1: Optimal TPPs for a biomarker-based diagnostic test for TB in adults and children.

A triage test should have sensitivity >90% (optimal sensitivity >95%) and specificity of >70% (optimal specificity >80%). In this case, a lower specificity is acceptable. Sensitivity (i.e., not missing a case) takes higher priority over specificity (i.e., avoiding unnecessary treatments) because of the importance of initial evaluation of patients with possible TB disease, and the need for further testing to confirm the diagnosis.

Currently, several methods are used for TB diagnosis. These methods include medical imaging,

immune-based tests, which examine the host's immune response to the pathogen, and microbiological tests, which detect the pathogen (Table 1.2).

Test	Strengths (S) and Limitations (L)
Clinical imaging	
Chest radiography	S: Frontal and lateral views are very useful in PTB. High sensitivity. L: Images are not always clear. Low specificity.
Immune-based tests	
Tuberculin Skin Test (TST)	S: Very common and cheap reagent; Easy to use and to interpret the results. L: Cannot differentiate between TB disease and <i>Mtb</i> infection; Affected by BCG vaccination; Requires a second visit after 48-72 hours; Low specificity.
Interferon-Gamma Release Assay (IGRA)	S: Unaffected by BCG vaccination; Requires one single visit. L: Cannot differentiate between TB disease and <i>Mtb</i> infection.
Microbiological tests (detection of <i>Mycobacterium tuberculosis</i>)	
Smear microscopy	S: High specificity; Rapid (<1 hour) detection. L: Very low sensitivity; Highly operator-dependent; Labour intensive.
Grown on special media (TB culture test)	S: Diagnostic gold standard; High specificity; Species identification. L: 2-6 weeks required; Highly operator-dependent.
Rapid molecular tests	S: High specificity and sensitivity in diagnosis PTB; Detection of MDR-TB; Rapid (<2 hours) detection. L: High costs; Lower accuracy for EPTB diagnosis; Temperature-sensitive reagents; Regular maintenance and calibration needed.
Urine tests	S: Easy-to-use; Rapid (<30 minutes) detection; Biomarker-based (non-sputum); High specificity in HIV-infected patients; Diagnosis of PTB and EPTB. L: Lower sensitivity in HIV-uninfected patients.

Table 1.2: Current diagnostic tests for tuberculosis.

Chest radiography is used to identify abnormalities in the thoracic cavity, in particular in the lungs when PTB is present. This is a rapid imaging technique with a high sensitivity [45] in identifying any abnormalities that are consistent with TB disease. However, it has a low specificity since some of the lung abnormalities are also seen in other pulmonary pathologies and, therefore, they are not exclusively observed in patients with PTB. Moreover, it is limited by inter-reader variability and its very high equipment costs, by the technical expertise needed to perform and interpret radiographs, and shortages of material and supplies, which can hamper its use in peripheral settings, especially in LMIC [46]. Chest radiography is a useful diagnostic tool, but it is usually used to assist patient management in bacteriologically negative cases.

The Tuberculin Skin Test (TST) and Interferon-Gamma Release Assay (IGRA) tests are based on measuring the host's immune response to the pathogen. Their main shortcoming is that they cannot differentiate between TB disease and *Mtb* infection, hence the tests give a positive result if the bacteria live in a latent and asymptomatic state in the host.

The TST, which involves an intradermal injection of a small amount of a combination of proteins extracted from *Mtb*, called Purified Protein Derivative (PPD), into the patient's forearm, is very easy to perform, inexpensive and its results are very straightforward to interpret, but it lacks in specificity since the rate of false negatives is very high [47]. If the patient has been previously exposed to the bacteria, the skin is sensitive to tuberculin PPD and a small and hard red lump might appear at the site of injection within 48-74 hours. The size of the lump, measured in millimetres, indicates if *Mtb* lives in the host. If the patient has had the BCG vaccination, they might also have a mild skin reaction to the TST test.

The IGRA tests are whole-blood tests for TB based on the IFN- γ release after stimulation by some antigens, such as Early Secreted Antigenic Target 6 (ESAT-6) or Culture Filtrate Protein 10 (CFP-10)

which are more specific to *Mtb* than PPD used in TST [48], especially for BCG vaccinated people [49]. In fact, these antigens are encoded by genes located within the region of difference 1 (RD1) locus of the *Mtb* genome [50]. They are more specific than PPD for *Mtb* because they are not encoded in the genomes of any BCG vaccine strains [50]. The current IGRA tests measure either the amount of IFN- γ released by memory T lymphocytes (i.e., QuantiFERON-TB Gold test) or quantify individual memory T lymphocytes releasing IFN- γ in the blood (i.e., T-SPOT.TB test) after incubation with *Mtb*. Since these tests are *ex-vivo* assays, they can be repeated more than once with different aliquots of the same whole-blood sample, and therefore, they do not require the patient to have a follow-up visit to read the results. However, these tests are more expensive than TST and require a high blood volume, thus its use is limited in children. TST and IGRA tests are usually used as indicators of the presence of *Mtb* infection.

The first diagnostic tests for TB were based on the identification of the pathogen through staining of bacilli present in clinical specimens, primarily sputum, and making them visible with the use of light microscopy [51]. The sputum smear microscopy is still widely used since it is simple to perform, rapid and inexpensive, and it is the primary method to diagnose PTB in LMIC, where over 90% of the total TB cases and deaths worldwide occur [52]. This diagnostic test has a high specificity (>98%) but its sensitivity ranges from 20% to 80% [53, 54]. The low and variable sensitivity, which might be linked to the low amount of bacilli, the quality of sample collection or sample processing [55], is the main limitation of sputum smear microscopy. Moreover, it cannot differentiate between *Mtb* and other Nontuberculous Mycobacteria (NTM) [56].

A second method to identify the pathogen in a biological specimen is the TB culture test, which relies on the sample being sent to a laboratory, placed on special media to isolate and culture *Mtb*, and observing its growth. Culture tests performed on solid media, such as Lowenstein Jensen (LJ) medium, are time-consuming, take up to 8 weeks, and have a low sensitivity especially in the samples containing a small amount of bacilli [57]. In recent times, liquid culture media have been evaluated and introduced. The Mycobacteria Growth Indicator Tube (MGIT) system incubates and monitors 7 mL culture tubes. The tubes contain an oxygen quenching fluorescent compound embedded in silicon used as growth indicator. As the microorganisms grow in the media, oxygen gets exhausted allowing the fluorescence to be detected automatically over time [58]. Liquid media have been shown to be more accurate and to provide earlier results than solid medium [58], and hence they are currently widely used. Although the TB culture test is considered the current standard for *Mtb* detection in all specimen types for its high specificity and sensitivity, it can be costly, it requires laboratory facilities and specialised training for the operator, and it is prone to contamination. Additionally, the slow growth rate of the organism may result in a delay of up to 6 weeks in obtaining results.

More recently, the WHO have recommended the use of rapid molecular tests, such as Nucleic Acid Amplification Tests (NAATs) or Loop-mediated isothermal Amplification (LAMP), instead of the sputum smear microscopy.

NAATs are molecular tests to detect the Deoxyribonucleic Acid (DNA) of the Mycobacterium tuberculosis complex (MTBC) in the sputum specimen. They are accurate in detecting PTB, especially in HIV infected patients or in people with a paucibacillary disease (culture positive but smear negative) [59]. Xpert® MTB/RIF assay and the newer Xpert® MTB/RIF Ultra assay, both produced and commercialised by Cepheid, are currently widely used since they almost achieve all the requirements for an ideal POCT smear replacement test [60] and can detect MDR-TB. Xpert® MTB/RIF assay amplifies a *Mtb*-specific

region of the *rpoB* (RNA polymerase beta subunit) gene and uses mutation-specific probes to detect rifampicin resistance. The newer Xpert® MTB/RIF Ultra assay amplifies two additional genetic regions and differentiates silent mutations and mutations that cause rifampicin resistance. Its limit of detection is lower compared to the Xpert® MTB/RIF assay, making the new platform more sensitive and effective in detecting paucibacillary TB. Xpert® MTB/RIF and Xpert® MTB/RIF Ultra assays perform better in diagnosing PTB compared to the sputum smear microscopy [61] and they are performed in the automated GeneXpert System instrument. Despite their benefits in diagnosing TB, Xpert® MTB/RIF and Ultra have low sensitivity in diagnosing EPTB [62–65]. In addition, these tests cannot be used to monitor treatment response because they detect dead bacilli. The GeneXpert System is also expensive and requires constant electricity, calibration, regular maintenance, and storage of temperature-sensitive reagents [66]. These factors are particularly limiting in resource-limited settings, where inadequate sample processing and frequent power failures affect the performance of these tests [67–69].

WHO endorsed the use of LAMP for the diagnosis of PTB in adults with signs and symptoms of the disease [70]. LAMP is a technique where nucleic acid amplification occurs under isothermal conditions (i.e., at a constant temperature). To diagnose TB, LAMP targets and amplifies some regions of the *Mtb* complex genome. Due to its simplicity and low cost, LAMP might be a good option to be used as a screening or diagnostic test in LMIC. However, it has been reported that its sensitivity is lower than Xpert® MTB/RIF Ultra [71].

Urine tests for TB detect Lipoarabinomannan (LAM), a lipopolysaccharide of the mycobacterial cell wall released from *Mtb* and excreted in the urine. Abbott commercialised a rapid lateral flow POCT, the Alere Determine TB LAM (i.e., AlereLAM), able to screen for TB disease in HIV-infected patients. This test is easy to use and has a turnaround time of less than 30 minutes. Since it is not sputum-based, it is able to diagnose both PTB and EPTB with high specificity in HIV-infected patients. Since 2016, the WHO has endorsed this diagnostic test for people with symptoms of TB and living with HIV, even if its sensitivity is relatively low [72].

There is currently no diagnostic test that meets the WHO TPPs for the diagnosis of TB disease. Most of the diagnostic tests widely used today are optimised for respiratory samples and, therefore, have low sensitivity for EPTB, which comprises 15% of TB cases worldwide. As previously discussed, *Mtb* culture has high sensitivity and specificity with both pulmonary and extra-pulmonary samples, however, cultures are susceptible to contamination, they require quality assurance programmes, highly skilled staff, and biological level 3 facilities, which are only available in advanced central laboratories [73]. Moreover, results take up to 8 weeks and up to additional 6 weeks for drug-susceptibility testing [1]. The WHO *End TB Strategy* [74] seeks to reduce TB incidence by 90 %, TB mortality by 95% and costs for TB patients by 100% by 2035, compared with 2015 baseline levels. To achieve these ambitious targets new POCT for TB disease, new tests to identify which individuals will progress from infection to disease in the future, new vaccines and new drugs are required, especially in children where there are even more limitations than in adults.

1.7 Paediatric tuberculosis

The WHO estimates that 70 million children (<15 years of age) are currently infected with *Mtb*, 1.2 million progress to the disease each year and a quarter of them die [2]. The highest rate of childhood TB cases and deaths are in LMIC, where children are exposed to *Mtb* but not always evaluated for

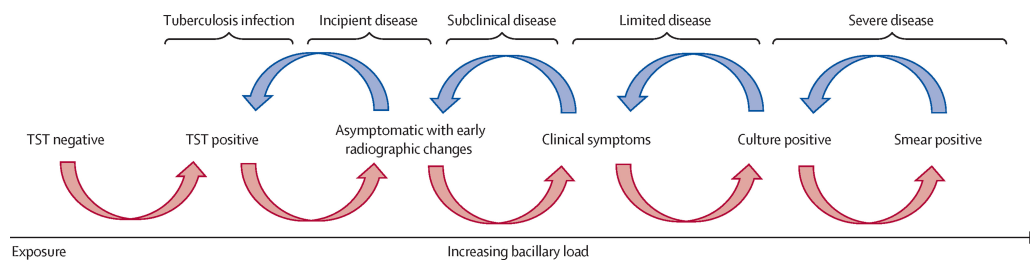


Figure 1.5: The dynamic spectrum from *Mtb* exposure to tuberculosis disease in children [83]. After exposure, children have an increase in bacterial loads over time, which can lead to a severe TB disease. The disease can then revert back to a latent and asymptomatic state (*Mtb* infection).

exposure or preventive therapy [75]. Although the BCG vaccine is included in childhood vaccination programmes in many countries, it is not effective enough to prevent this disease in young people. BCG vaccination, in fact, protects mostly young children from hematogenous TB including meningitis and miliary TB, but its efficacy in preventing PTB ranges up to 80% depending on sex, geographical location and pre-exposure to *Mtb* [76]. Typically, the risk of a child getting infected by *Mtb* depends on the probability, duration and proximity of exposure to an infected adult [77, 78]. Poverty, poor housing, urban environments and overcrowding are also associated with increased rates of transmission [79, 80].

There are many challenges in the diagnosis and treatment of TB that are unique to the paediatric population. Firstly, differences in the pathophysiology and clinical presentation of TB in children make the diagnosis more challenging than in adults [81] and definitions of latent infection and active disease are less clear cut [78]. Paediatric TB disease occurs more often in disseminated or extra-pulmonary forms compared to adults, and the initial presentation of the disease might have milder manifestation [82]. The TB pathogenesis in children has been always considered to have two dichotomous states: *Mtb* infection and TB disease. On the contrary to adults, after exposure to *Mtb*, children tend to develop TB disease rather than *Mtb* infection, due to an increase in bacterial loads over time. Childhood TB disease can then revert back to an asymptomatic and latent state. However, this dichotomous concept might be an extremely simplistic way to describe a dynamic spectrum of the disease [Figure 1.5]. Children can potentially move in both directions, from exposure (TST negative) to severe TB disease, and from TB disease to *Mtb* infection.

The risk of *Mtb* infection rises in a relatively linear way with age, reflecting cumulative exposure [84]. Young children have fewer antigen-presenting cells, with reduced functional responses, including phagocytosis and recruitment capacity, compared to older children and adults, and this can impair the T cell response to *Mtb* [36]. Recent studies have shown that non-conventional T cells, such as $\gamma\delta$ T cells, NK T cells, Th17 and regulatory T cells, are increased in young children with TB compared to healthy children [85] and might play a role in the host immune response to *Mtb*. On the contrary, the risk of progressing from *Mtb* infection to disease is higher in children younger than 5 years of age [86], it decreases in primary school age and increases again when children enter puberty [87]. Therefore, young children are at much higher risk and they progress more rapidly than adults from exposure to infection to clinically apparent disease [83, 88]. Children, because of the paucibacillary nature of childhood TB (i.e., low bacterial loads which cause negative smear results and positive *Mtb* culture), have usually less transmissible TB than adolescents and adults, but they are at greater risk of severe forms of this disease, particularly children younger than 5 years and without BCG vaccine [78, 89, 90].

The WHO has drawn a lot of attention to the TB-HIV co-infection and recommended that children with TB should be screened for HIV and conversely [36]. HIV increases the incidence of TB in children by a factor of around eight [91] and for young children living with HIV, TB is more aggressive and deadly [92]. Furthermore, whilst the WHO strongly recommend to BCG vaccinate HIV-uninfected neonates [93], they recommend delaying the vaccine in HIV-infected infants until they are clinically well and established on AntiRetroviral Therapy (ART) due to the risk of developing local or disseminated BCG disease after vaccination [76]. Immunocompromised children may be at risk of adverse effects from BCG vaccination, and in addition, the vaccine often elicits a weak immune response [92]. Early diagnosis and vaccination of infants living with TB-HIV co-infection can therefore be very challenging, especially in LMIC due to delayed testing or turnaround time of testing results, making even more difficult the access to effective TB treatment and preventive measures.

Since the beginning of the COVID-19 pandemic in early 2020, over 750 million confirmed cases of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) have been reported worldwide [<https://covid19.who.int>; accessed January 29th, 2023]. The COVID-19 pandemic has had a huge impact on TB care and prevention in children and adults, caused mainly by barriers faced in accessing healthcare facilities and services in high-burden countries [94]. In 2020, the TB notifications dropped by over 25% in children younger than 15 years old, with only 37% of children who developed TB being notified as diagnosed and/or treated [1]. TB notifications have improved in 2021, but they have not reached the pre-pandemic level achieved in 2019 [2]. Aside from the effects of the COVID-19 pandemic on the care and prevention of TB, it has been shown that co-infection of TB and COVID-19 in children carries a greater likelihood of severe illness and mortality compared to either disease alone [94–96].

The 80% of the worldwide paediatric TB deaths occur in children younger than 5 years old, with 96% of them occurring in children who remain untreated due to missed diagnoses [91]. The primary reasons for the high number of undiagnosed cases of paediatric TB are the limitations of the current diagnostic tests and the presence of symptoms and signs that resemble those of other prevalent paediatric conditions, which can complicate the clinical evaluation of the patient. Radiological investigations, for instance, yield non-specific observations with abnormalities that could be consistent with TB but also with other paediatric conditions [84]. The most widely used diagnostic tests for TB, which are based on culture or nucleic acid amplification to detect *Mtb* in respiratory or extra-pulmonary samples, require invasive procedures in children. Children are unable to spontaneously expectorate sputum, and induced sputum or gastric aspirate procedures are often needed for the collection of biological samples. Moreover, paucibacillary PTB is more difficult to diagnose by culture or smear microscopy, given the lower bacterial loads in the samples. *Mtb* culture detects only 30-62% of the childhood TB cases with results requiring up to 8 weeks [97]. Besides its limitations, *Mtb* culture is the gold standard for diagnosing paediatric TB. Tests for immunologic evidence of the presence of *Mtb* in the host (i.e., TST and IGRA tests) have limitations in both sensitivity and specificity for TB disease and *Mtb* infection. Additionally, accurately interpreting these tests in early infancy is further complicated by uncertainties regarding immune maturation or dysfunction in the presence of malnutrition [98], which has a higher prevalence in LMIC where TB is endemic. The nucleic acid amplification tests (Xpert® MTB/RIF and Ultra) can provide diagnostic results in a few hours, however, they detect only 65% of the *Mtb*-positive sputum and 73% of the *Mtb*-positive gastric samples detected by *Mtb* culture [99]. The NAATs may thus detect only 20-40% of all paediatric TB cases, and their sensitivity may be even lower in children living with HIV because of the higher rates of culture-negative TB cases [100].

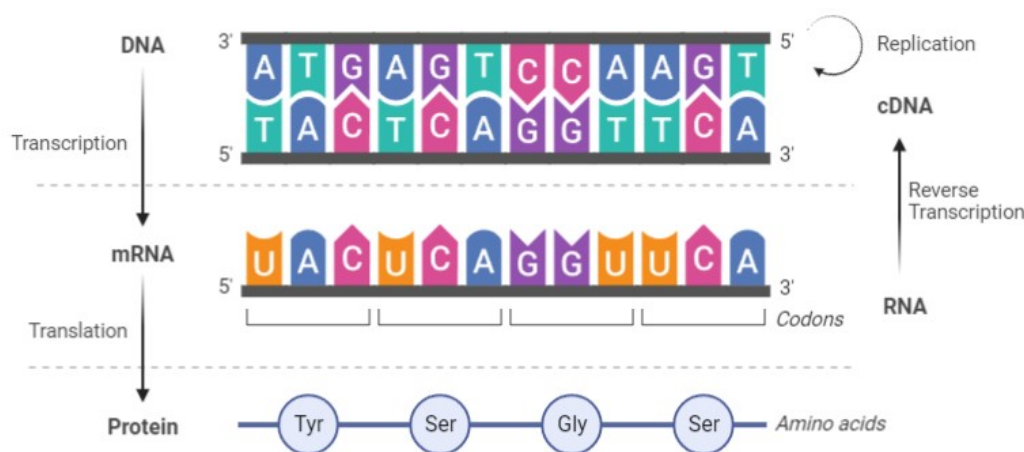


Figure 1.6: The central dogma of molecular biology.

The central dogma of molecular biology is a theory stating that genetic information flows only in one direction, from DNA to RNA, to a functional product (i.e., a protein). Only years after its publication, a reverse transcription step (from RNA to DNA) was described. DNA: Deoxyribonucleic Acid; RNA: Ribonucleic Acid; Tyr: Tyrosinase; Ser: Serine; Gly: Glycine. Adapted from “Central Dogma”, by BioRender.com (2022). Retrieved from <https://app.biorender.com/biorender-templates>.

The screening of paediatric TB is essential to control the disease spreading since a child with *Mtb* infection or TB disease is often an indicator of an infectious adult case in the household. Prevention and treatment of paediatric TB will reduce the future global disease burden, and new and more accurate tools will likely be key steps in the global TB elimination and in reaching the targets of the WHO *End TB Strategy* [74] by 2035. To lessen the global burden of TB, the WHO has identified the need for new tools to identify children with TB disease, with *Mtb* infection who are at risk of progressing to TB disease, vaccines that provide long-lasting protection, and new drugs. However, developing these tools can be challenging and time-consuming due to the difficulties in detecting *Mtb* in biological samples. As an alternative approach, the specific host immune response to *Mtb* can be used to distinguish between different clinical groups. In the past few decades, growing attention has been drawn to transcriptomic and proteomic host biomarkers specific to TB, which can be used to develop new non-sputum diagnostic tests for TB in both adults and children. In order to illustrate why genes and proteins can serve as effective biomarkers for diagnosing TB, as well as other diseases, it is essential to define the processes of Deoxyribonucleic Acid (DNA) replication, transcription, and translation.

1.8 The central dogma of molecular biology

The central dogma of molecular biology defines the processes of DNA replication, transcription, and translation [101] and provides an explanation of how genetic information flows within a biological system [Figure 1.6]. The central dogma states that genetic information flows only in one direction, from DNA to Ribonucleic Acid (RNA), to proteins [102]. Almost 20 years later, two scientific teams, one led by Howard Temin at the University of Wisconsin and the other led by David Baltimore at the Massachusetts Institute of Technology, independently identified new enzymes able to transcribe an RNA templates into complementary DNA (cDNA) molecules [103]. This process is called reverse transcriptase and is explained in more detail in Section 1.10.1 of this chapter.

DNA is a long polymer that contains the hereditary materials in humans and almost all other organisms. It is composed of two polynucleotide chains that coil around each other to form a double helix [104].

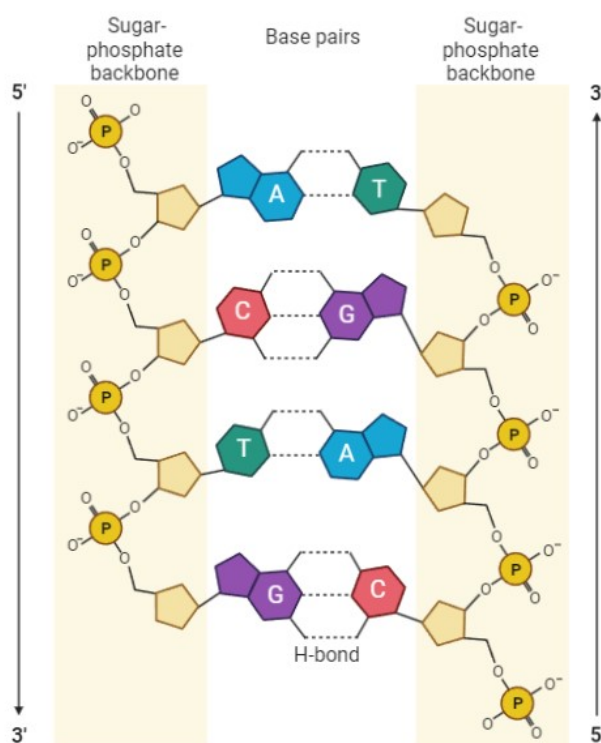


Figure 1.7: Molecular structure of the DNA.

The DNA is a polymer with a double helical structure made up of two strands linked together by hydrogen bonds. Each strand is composed of a sequence of 4 different nucleotides. A: Adenine; G: Guanine; C: Cytosine; T: Thymine; P: Phosphate group.. Reprinted from “Molecular Structure of the DNA”, by BioRender.com (2022). Retrieved from <https://app.biorender.com/biorender-templates>.

The two DNA strands are known as polynucleotides because they are made up of single units called nucleotides. Each nucleotide is composed of a deoxyribose sugar molecule to which is attached a phosphate group and one of the four nitrogenous bases: two purines (adenine [A] and guanine [G]) and two pyrimidines (cytosine [C] and thymine [T]) [104]. The nucleotides are joined together by covalent bonds between the phosphate group and the sugar of two consecutive units, forming a phosphate-sugar backbone from which the nitrogenous bases protrude. The two strands are linked by hydrogen bonds between the bases. The sequencing of this bonding is specific with adenine bonding only with thymine, and cytosine only with guanine [Figure 1.7]. DNA and RNA strands are directional, with the two ends named 5' (usually pronounced “five-prime end”) and 3' (usually pronounced “three-prime end”). The “five-prime end” and “three-prime end” indicate the carbon number in the nucleotide pentose-sugar-ring. Directionality is an end-to-end chemical orientation, meaning that the nucleic acids are synthesised and read in the 5'–to–3' direction because the polymerases involved in the replication of the nucleic acids use the energy produced by breaking nucleoside triphosphate bonds to add nucleotides in the new strand [105]. The order of the nucleotides is also important since it determines the unique genetic code of each living organism and contains the instructions for producing functional products, usually proteins.

In eukaryotes, DNA is organized into chromosomes [105]. The configuration of the DNA molecule is highly stable, allowing it to act as a template for the replication of new DNA molecules, as well as for the production (transcription) of the related RNA molecule. Genes are segments of DNA that contain the set of instructions to produce proteins. The human genome contains over 20,000 protein-coding genes, but a large part of DNA is non-coding, hence it does not serve as patterns for protein production.

An important property of the DNA is that it can replicate and make copies of itself. The two strands are separated during the DNA replication and each strand in the double helix can serve as a pattern for duplicating the sequence of bases since they are complementary and store the same biological information. DNA replication occurs in three phases: the separation of the double helix and the breaking of the hydrogen bonds by, respectively, an initiator protein and the helicase; the synthesis of a primer to start the replication of the DNA; and the addition of complementary nucleotides by the DNA polymerase.

To transform the genetic information stored in the DNA into functional products, transcription and translation are needed. DNA is firstly converted into RNA via transcription, which is then converted into proteins via translation. Although the full set of DNA is always present in all cells of an organism, not all functional products are always needed at every given moment. Therefore genes are expressed and proteins are produced only when the cells need the functional products.

RNA is essential in various biological roles such as coding, decoding, regulation and expression of genes. It is a nucleic acid that has structural similarities to DNA, but unlike it, RNA is usually single-stranded. Moreover, RNA has the sugar ribose and the uracil [U] base, instead of the deoxyribose and the thymine [T] base found in DNA. The presence of the ribose sugar makes the RNA prone to hydrolysis and less stable than DNA. Out of the many types of RNA, the ones involved in protein synthesis are the messenger RNA (mRNA), the transfer RNA (tRNA) and the ribosomal RNA (rRNA). RNA strands are created using DNA strands as a template during transcription, which is carried out by the RNA polymerase and a number of accessory proteins, the transcription factors. The transcription factors can bind to specific DNA sequences to recruit RNA polymerase molecules to an appropriate transcription site [106]. Together, the transcription factors and RNA polymerase form the transcription initiation complex, which enables the RNA polymerase to start the RNA synthesis by matching complementary bases to the original DNA strand. The RNA molecule is elongated and, once the strand is completely synthesized, transcription is terminated [Figure 1.8]. In eukaryotes, once the RNA product has been synthesised, the non-protein-coding sections (i.e., introns) are removed, leaving just the protein-coding regions (i.e., exons), forming in the mRNA molecules [105]. Exons from the same genes can be joined in different combinations leading to different mRNAs during the process of alternative splicing.

Transcription factors have an important role in the synthesis of RNA since they can either promote the transcription by recruiting RNA polymerase or suppress it by preventing the RNA polymerase from binding to the promoter region. Therefore, transcription factors ensure that genes are transcribed only if their functional products are required. mRNA carries the same information as DNA but is not used for long-term storage and can freely exit the cell nucleus.

The newly formed mRNA molecule is a complementary copy of the DNA template, it contains the same biological information and serves as a blueprint for protein synthesis during translation. Proteins are chains of amino acids linked by peptide bonds. In an mRNA molecule, the instructions for building polypeptides are RNA nucleotides read in codons (i.e., groups of three nucleotides). tRNA and ribosomes have key roles in the translation process [106]. tRNA are molecules that connect mRNA codons to the amino acids they encode. One end of each tRNA has a sequence of three nucleotides (i.e. anticodon), which can bind to specific mRNA codons. The other end of the tRNA carries the amino acid specified by the codon. Proteins are produced in ribosomes, which are made up of ribosomal proteins and rRNA. The ribosome provides a set of slots where tRNA molecules can find their matching codons on the mRNA template and deliver their amino acids [106]. They also act as an enzyme, catalyzing the

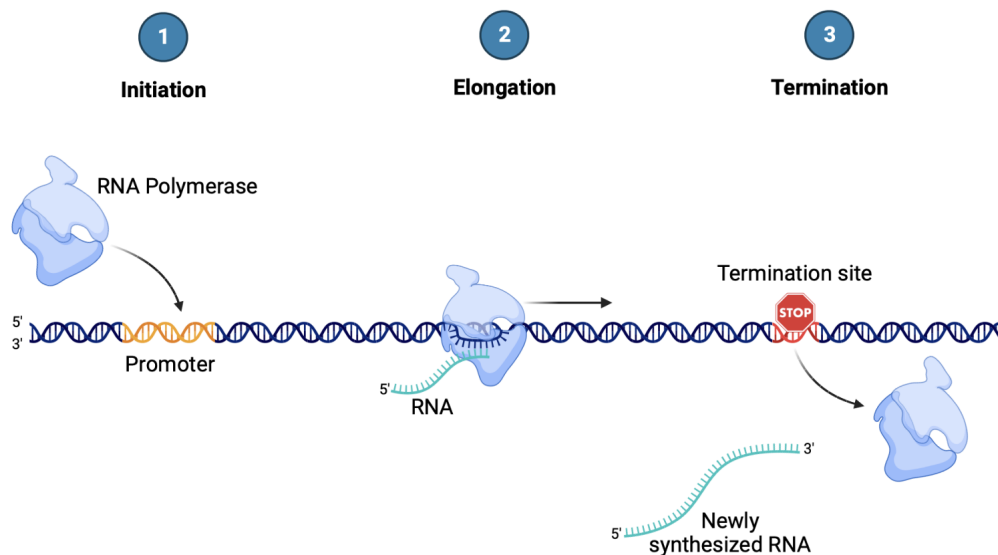


Figure 1.8: Transcription from DNA to RNA.

The DNA is transformed into mRNA via transcription, which is carried out by RNA polymerase and a number of accessory proteins, the transcription factors. Transcription is made up of three steps: initiation, elongation and termination. Reprinted from “Eukaryotic Transcription (Landscape)”, by BioRender.com (2022). Retrieved from <https://app.biorender.com/biorender-templates>.

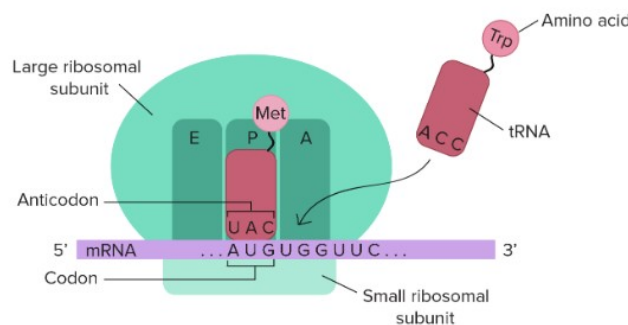


Figure 1.9: Translation from mRNA to proteins.

The mRNA is transformed into proteins via translation, which is carried out by tRNA polymerase into subunits called ribosomes. *tRNA: transfer RNA*. Created with BioRender.com.

chemical reaction that links amino acids together to make a chain [Figure 1.9].

Finally, the completed polypeptide chain is released from the ribosome and folded into its final protein state [106]. Proteins are required for the structure, function, and regulation of the body's tissues and organs. Human cells make nearly 100,000 different types of proteins, each with its own unique mRNA sequence.

The presence or severity of diseases alters gene expression and protein abundance. The activation of the immune system caused by a specific pathogen activates antibodies and other proteins, such as chemokines and cytokines, changing the abundance of the mRNA precursor molecules as well. The immune response, and therefore the expression of some genes and abundance of some proteins, elicited in the host differs based on the specific pathogen. If the host immune response to a specific pathogen is well-characterised, the changes in biomarkers, such as genes or proteins, after infection can be used to assess the presence in the host of that specific pathogen.

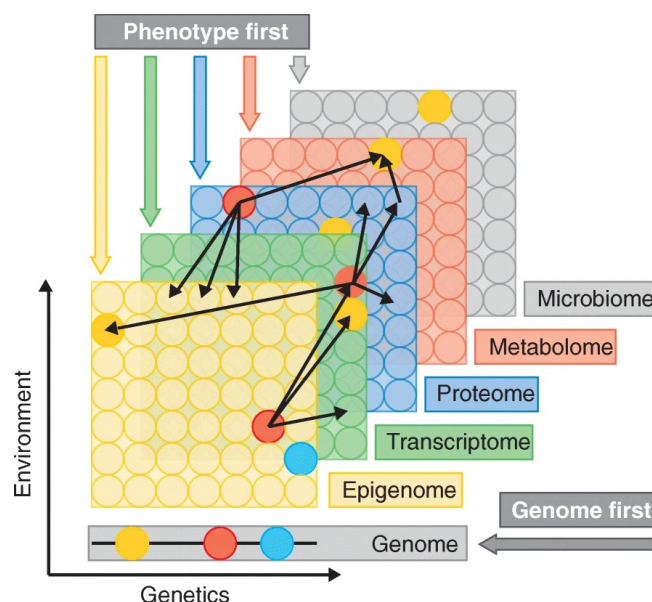


Figure 1.10: Multiple omics data types and approaches to disease research [109].

All data layers are not single and isolated entities, they interact with each other. The arrows represent potential interactions or correlations detected between molecules in different layers.

1.9 Biomarkers as diagnostic test

Since it is difficult to diagnose TB by detecting the pathogen, the focus has been moved to the detection of the host immune response elicited by the pathogen. As outlined previously in this chapter [Section 1.3], the host immune response changes based on the type of pathogen that is causing the infection, and this leads to specific changes in levels of molecules involved in the host immune system to challenge the infectious agent. Variations in gene expression and protein abundance are, therefore, a host's defensive reaction to the presence of a pathogen to ensure that the necessary proteins or metabolites are available for the immune system to neutralise the external agent. A gene or a protein can be, thus, defined as a biomarker, that is to say, a biological marker (substance, structure or process) that can predict the incidence, severity or outcome of a disease [107]. The changes in the level of biomarkers for a specific disease are usually captured by untargeted host molecular profiling, or -omics, in tissues from the site of disease [84]. The -omics notion refers to high-throughput biochemical assays that measure comprehensively and simultaneously molecules of the same type from a biological sample, providing a complete view of the molecules' profile in the biological system [108]. The analysis of -omics data is essential to better understand biological processes [Figure 1.10] such as the dynamics of infection and inflammation.

Genomics, epigenomics, transcriptomics, proteomics, metabolomics and microbiome fall all under the umbrella of -omics. They have been extensively studied as a single molecular layer (i.e., single-omics), but recently the concept of integrating single-omics datasets has become more popular. Multi-omics integration can, in fact, explore the interactions between different molecular levels and lead to a comprehensive representation of the biological system. This thesis will focus on the study of gene expression (i.e., transcriptome) and protein abundance (i.e., proteome) as single-omics, along with their integration to identify new multi-omics biomarkers for paediatric TB. Bulk transcriptomics is the study of the complete set of RNA transcripts that are produced by the genome in a specific tissue and under specific circumstances (i.e., transcriptome), and it quantifies the gene expression. Proteomics is

the study of the complete set of proteins produced by an organism (i.e., proteome), and it quantifies the abundance of the proteins. Studying the host transcriptome or proteome of the tissue from the site of the disease can elucidate the local pathogenic characteristics of the disease and the host immune response at a tissue or cellular level. For instance, type I/II interferon-mediated cytokine responses and T-cell activation have been revealed to be strongly activated in adults with TB by the analysis of host transcriptomics of BronchoAlveolar Lavage (BAL) or sputum samples [110] and might be therefore used as biomarkers. Due to difficulties in collecting samples from the site of infection, especially in children, the use of biomarkers from peripheral blood, since it is the main route for transportation of immune cells and it is easier to sample, can give a more complete picture of the pathological and immunological changes in the whole biological organism [111]. Presence of disease [112–115], prognostic information [116, 117] and effect on therapeutic response [118] have been found to be reflected in the gene expression pattern of blood cells.

Various techniques have been developed for characterising and quantifying RNA transcripts. These methods can be broadly categorised into those characterising a small number of genes (low-throughput), or a large number of genes (high-throughput). High-throughput methods are usually used as “hypothesis-free” approaches, where no hypothesis is known but all data are acquired and analysed to generate a hypothesis or characterise the RNA profile in an exhaustive manner [119]. Contrastingly, low-throughput methods are usually “hypothesis-driven”, thus they are used to validate or disprove a given hypothesis. High-throughput methods, such as microarray and RNA Sequencing (RNA-Seq), are usually used in biomarker discovery studies, whilst low-throughput approaches are used in biomarker validation studies.

With the advances in both high-throughput methods and computational technologies, it has become easier to assess which molecules change their levels after an infection caused by a specific pathogen. By knowing the molecules and the biological pathways they map to and comparing them between different disease groups, it can be inferred which biological processes are disrupted by the pathogen, unravelling the host response to the infection. The disease manifestation and how the host’s immune system responds to the same pathogen can vary significantly among individuals. This variability may be attributed to factors such as the host’s geographic location, race, ethnicity, gender, and immune system, which can be influenced by age, coexisting medical conditions, or concurrent infections [56]. Analysis of single-omics datasets is a way to take into account this heterogeneity and overcome it. Starting from a cohort of patients with known diagnoses, single-omics technologies are used to find biomarkers able to distinguish between disease groups. Once the biomarkers have been identified and validated on independent cohorts, they can subsequently be used on a POCT. The transcriptomic and proteomic technologies used to generate the data analysed for this thesis are described respectively in Sections 1.10 and 1.11.

1.10 Measuring gene expression

The ribose sugar, which has a highly reactive hydroxyl group (-OH), makes the RNA less chemically stable and more prone to heat degradation than DNA. RNA stabilising systems are therefore needed for sample collection and storage in order to preserve the biological sample. The reagents used to stabilise the RNA preserve the gene expression profiles that reflect the blood’s state at the moment of sampling. The two main commercial systems for stabilising and storing RNA from blood are the *PAXgene Blood*

RNA System (PreAnalytiX QIAGEN/BD, Hombrechtikon, Switzerland) and the *Tempus Blood RNA System* (Applied Biosystems, Foster City, CA, USA). These two systems have the same purpose of preserving the RNA in the biological samples, but they have different protocols to isolate the RNA and this, along with the different reagents used, results in biases in RNA quantification results [120, 121]. Good practice would be to avoid the mixed use of different RNA stabilising systems in the same experiment. In case this is not possible or very difficult to achieve, the differences between the different RNA stabilising systems need to be taken into consideration in study designs and meta-analyses [122]. Protocols for purification and isolation of RNA are equally important to ensure good RNA quality, integrity and yield, and to minimise potential DNA contamination [123]. RNA quality is usually measured by the RNA Integrity Number (RIN), a number between 1 and 10 measured using gel electrophoresis and calculated from the analysis of the ratios of 28S to 18S ribosomal bands, which are two subunits of rRNA. The RIN needs to be above 7-8 to ensure a successful experiment [124].

Transcriptomic profiling was a huge step forward for biomarker discovery, especially in children, since it can be achieved with a very small amount of blood. More recently, some studies have shown that in vitro whole blood stimulation with *Mtb* antigen peptides can enhance the diagnostic potential of single gene markers or unveil some transcriptomic signals that cannot be detected in unstimulated samples [125].

Sections 1.10.1, 1.10.2 and 1.10.3 describe respectively the Reverse Transcription - quantitative Polymerase Chain Reaction (RT-qPCR), microarray and RNA-Seq assay, which are currently the most used transcriptomic profiling techniques for transcriptomic biomarker discovery and validation.

1.10.1 Reverse Transcription - quantitative Polymerase Chain Reaction (RT-qPCR)

RT-qPCR is a targeted technique for the detection and quantification of RNA. It is the combination of three steps:

1. The reverse transcriptase (RT)-dependent conversion of total RNA or mRNA into cDNA
2. The amplification of cDNA using Polymerase Chain Reaction (PCR)
3. The detection and quantification of amplification products in real-time

Reverse transcription is a process based on reverse transcriptase enzymes to generate cDNA using a strand of RNA as a template. As explained in Section 1.8, the original central dogma of molecular biology states that the genetic information progresses just in one direction: DNA is transcribed into RNA, which is then translated into protein. This concept was revolutionised when reverse transcriptase enzymes, which convert RNA templates into cDNA, were identified in the 1970s.

Once cDNA is transcribed by reverse transcriptases, the core part of the RT-qPCR technique takes place. PCR, which Kary Mullis received the Nobel Prize in Chemistry in 1993 for, is highly sensitive and widely used to detect and amplify DNA and RNA sequences in just a few hours, compared to other methodologies previously used to clone and amplify nucleic acids that could take up to days. From a very small amount of biological sample, the DNA sequences are exponentially amplified in a series of cycles of temperature changes (i.e., thermal cycling) to allow temperature-dependent chemical reactions to happen. In the first step of PCR, the high temperature breaks the hydrogen bonds between the two DNA strands and separates them in the denaturation phase, which happens at 95°C. The temperature then goes down to 60°C, and the primers, short DNA molecules that are complementary to

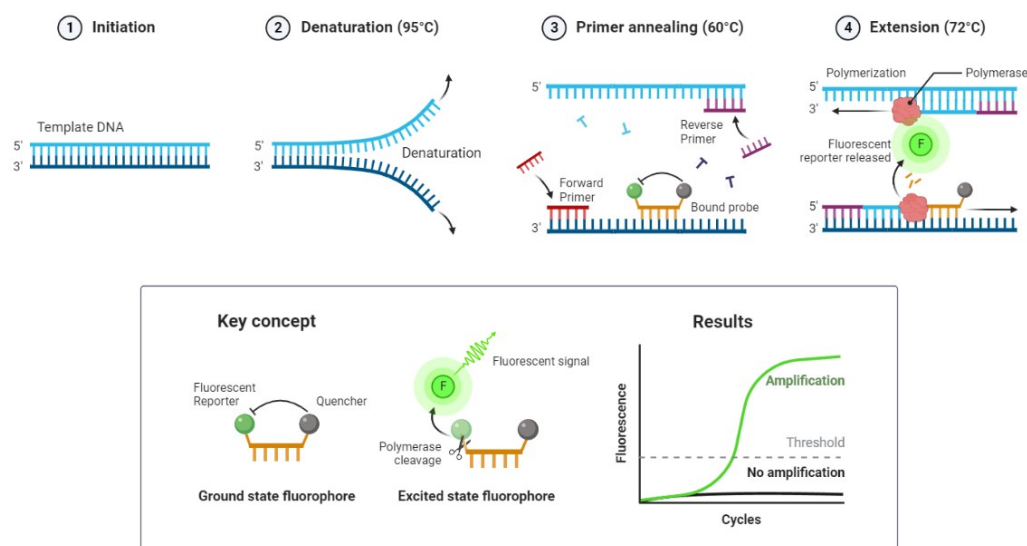


Figure 1.11: Polymerase Chain Reaction (PCR).

The PCR steps include denaturation, annealing and extension in order to amplify the DNA sequences of interest from a very small amount of nucleic acids. Adapted from “Fluorescent Probe-Based Real-Time PCR”, by BioRender.com (2022). Retrieved from <https://app.biorender.com/biorender-templates>.

the DNA sequences of interest, align to the 3' end of the single DNA strands during the annealing phase. This is followed by extension, where the DNA polymerase binds the primer and adds progressively complementary nucleotides that are free in the solution. In this step, the two DNA strands are used as templates to generate new copies of the original DNA sequences. Since most DNA polymerases that are used for PCR work best at 68 - 72°C, the extension temperature should be within this range. Each cycle of amplification, which involves denaturation, annealing, and extension, is repeated 20-40 times and amplifies both the original and newly produced DNA sequences, achieving an exponential amplification of the target sequences.

The real-time quantification of the amplified DNA is possible by using fluorescent labelling and by measuring the fluorescence at each cycle [Figure 1.11]. The widely used output value of an RT-qPCR run is the Cycle Threshold (C_t), defined as the number of cycles required for the fluorescent signal to cross the user-definable threshold or to exceed the background level. Whereas the fluorescence is proportional to the amount of amplified DNA, C_t values are inversely proportional to the amount of target nucleic acid since the more target nucleic acid is present in the sample and the fewer cycles are needed for the fluorescence to cross the threshold. The maximum C_t value in a real-time assay is the number of amplification cycles the assay undergoes, usually 40.

RT-qPCR is highly sensitive, reproducible, accurate and takes only a few hours to be performed. However, variability of RNA templates, assay designs and protocols, as well as inappropriate data normalisation and inconsistent data analysis might be huge limitations for this technique. Moreover, it is a targeted technique, meaning that just a few mRNA sequences can be measured in one single experiment. This has been overcome by the new high-throughput RNA quantification methods (i.e., microarray or RNA-Seq), which quantify many more RNA transcripts. High-throughput techniques are usually preferred for the discovery of novel candidate genes and pathways, whereas RT-qPCR is still widely used for validation studies.

1.10.2 Microarray

Once the human genome sequence was completed in 2001 [126, 127] and the DNA sequences that encode genes were published, many studies and experiments started focusing on identifying the regions of DNA that control normal or disease states. One of the first and most common tools used for this aim was the DNA microarray. Microarray is a multiplex lab-on-chip, a device that integrates several laboratory functions in a single and small integrated circuit to achieve automatic and high-throughput screening. The concept of microarray was first published in 1983 [128] when Tse Wen Chang described the antibody microarrays. There are several types of microarray, with the two-channel DNA microarray being the most used to detect the expression of thousands of genes simultaneously in two different disease groups. A microarray assay is a collection of specific nucleic acid sequences (i.e., probes) attached to a solid surface and used to detect the set of mRNA transcripts (i.e., targets). mRNA transcripts are typically collected from both experimental and reference samples (usually healthy patients) in order to compare the two groups. Since mRNA degrades easily, it is necessary to be converted into a more stable cDNA form synthesised by the single-stranded RNA template. Each cDNA sequence is then labelled with different fluorescent dyes for the different disease groups in order to discriminate the over-expressed and/or under-expressed genes in the group of interest.

The core principle of microarray is the hybridisation step between two DNA strands, the fluorescently labelled targets (cDNA) and the probes, through hydrogen bonds. A washing step is needed to wash off the non-specific bindings and ensure that just the strongly paired strands remain hybridised. The microarray is then scanned to measure the expression of each gene printed on the slide by reading the fluorescence of the labelled targets. The measurements are finally transformed into numerical values proportional to the abundance of the mRNA transcripts in the sample. A general DNA microarray experiment is summarised in Figure 1.12.

The microarray assay gives an accurate picture of gene expression in the biological sample at the time of collection quantifying the expression of tens of thousands of genes simultaneously and in a high-throughput manner. As a result, the protocols required for processing, normalising and analysing microarray data are complex and computationally very demanding. Whilst microarray revolutionised transcriptomic profiling, this technology relies on probes from known gene targets, meaning that new transcripts cannot be detected. Newer high-throughput platforms, which are not hybridization-based approaches, have overcome this limitation not relying anymore on already known nucleic sequences. An example of these new technologies is the RNA-Seq assay, which is described in the next section [Section 1.10.3].

1.10.3 RNA Sequencing

RNA-Seq, in contrast to microarray, is a sequence-based approach that directly determines the cDNA sequence, and therefore it does not need probes. The RNA-Seq technique, since it is not a hybridization-based approach, measures the level of known protein-coding mRNA, along with other noncoding RNA (ncRNA) molecules which have important functional roles and regulate gene expression at a post-transcriptional level (i.e., small ncRNA and long ncRNA) [130]. RNA-Seq gives a more detailed and specific view of the gene expression and measures the exon/intron boundaries, alternative splicing and allele-specific expression. All the steps involved in an RNA-Seq experiments are summarised in Figure 1.13.

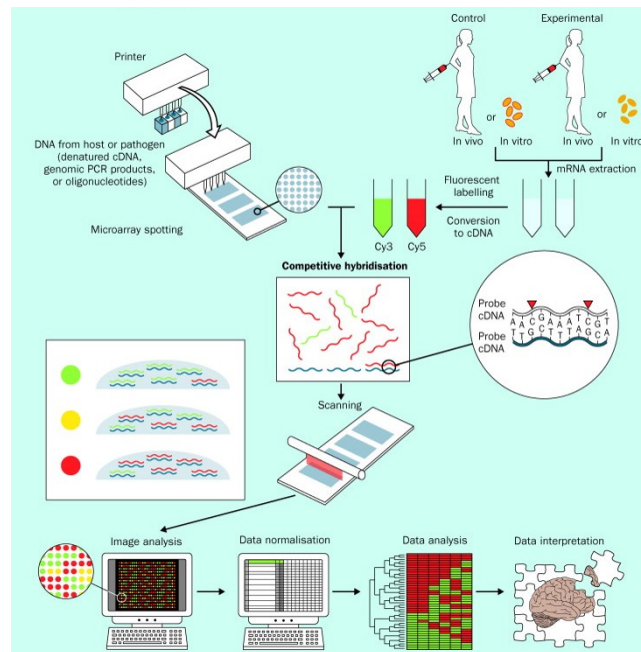


Figure 1.12: Overview of the steps involved in DNA microarray experiments [129].

DNA probes are printed on the microarray surface and bind to the cDNA sequences complementary to the RNA transcripts extracted from the biological samples. Once the hybridisation and washing-off steps are performed, the microarray plate is scanned and fluorescence detected. The intensities of the fluorescent dyes at each spot are proportional to the amount of fluorescent-labelled cDNA that has hybridised to the DNA probe at that specific spot.

A typical RNA-Seq experiment starts with isolating the RNA from the biological sample and assessing whether the extracted RNA has sufficient quality and is not degraded. The RNA-Seq preparation library is the step following RNA isolation. The sequencing library involves isolating the RNA species and molecules of interest, reverse-transcribing from RNA to cDNA, amplifying primer cDNA molecules, and ligating sequencing adaptors to them. It is very important to remove rRNA before library construction since rRNA accounts for over 95% of the total RNA in a cell, and it might consume the bulk of sequencing reads and limit the detection of other less-abundant RNA molecules. Along with rRNA depletion, many protocols focus on enriching for mRNA molecules before library preparation. There are several choices in library design protocols and the final protocol should be carefully chosen based on the desired output data, in fact, the library designs differ in their capacity to enhance the detection of certain transcripts while limiting the ability to detect other transcripts. The final and common step of the different sequencing library preparations is the conversion from RNA to cDNA fragments, which are immobilised by adaptors, short oligonucleotides that bind to one end or both ends of cDNA fragments. The library is then amplified by PCR and sequenced.

The sequencing step changes based on the sequencing platform chosen. The first sequencing method was invented in 1977 by and named after Fred Sanger, who converted the sequencing problem into a problem of measuring mass. This method was based on a sequencing-by-synthesis approach [132]. This technique was low-throughput since the mass measuring process was time-consuming. Moreover, it could only be used for sequences below 700 base pairs (bp) in length due to the difficulties in detecting differences in mass between long sequences when a single nucleotide is added.

The second-generation sequencing, pioneered by Illumina, modified the sequencing-by-synthesis principle and transformed it into a high-throughput assay. This sequencing approach runs several

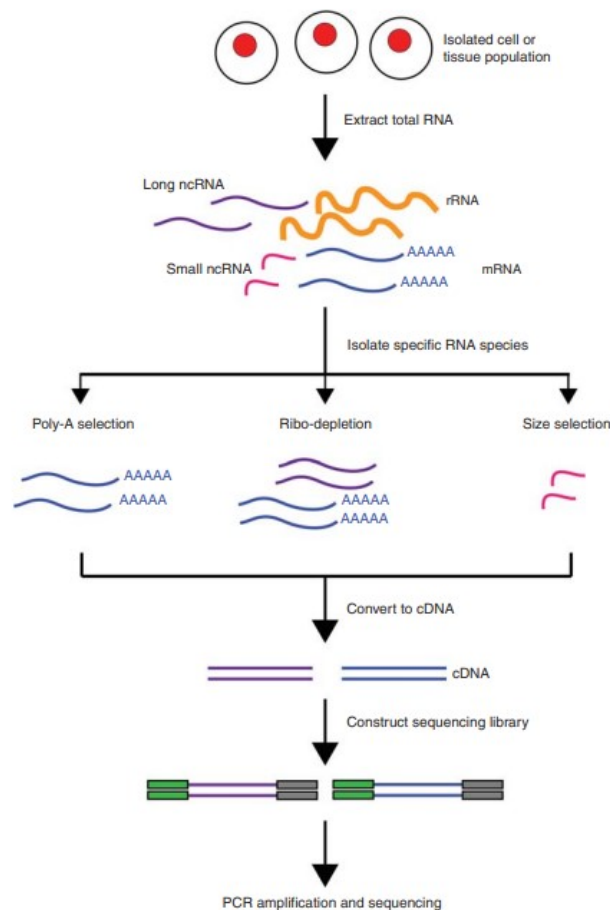


Figure 1.13: Overview of the steps involved in a RNA-Seq experiment [131].

RNA is extracted from the biological sample, isolated using a specific protocol, and converted to cDNA by reverse transcription. Sequencing adaptors are then ligated to the ends of the cDNA fragments. Following amplification by PCR, the RNA-Seq library can be sequenced. *rRNA*: ribosomal RNA; *mRNA*: messenger RNA; *ncRNA*: noncoding RNA; *cDNA*: complementary DNA; *PCR*: Polymerase Chain Reaction.

syntheses in parallel and uses only fluorescently labelled reversible terminators, which are nucleotides that end the synthesis, but once cleaved, they can bind to new bases [133]. The need to measure masses was therefore transformed into an easier and quicker process of detecting light by the use of reversible termination. Once the templates have been sequenced base-by-base, the reads are either mapped to a reference genome or transcriptome, or assembled *de novo* if the sequence doesn't match with any known regions of the genome.

The choice of the biological sample is one of the most important initial considerations when planning an RNA-Seq experiment. Each biological sample or tissue has hundreds of cell types, and a bulk RNA-Seq experiment will look at the whole set of cell types present at the time of sampling. Moreover, the heterogeneity and purity of the tissue need to be taken into consideration since they might impact the gene expression quantification. Recent advances in RNA-Seq include single-cell sequencing, which captures the transcriptomic composition of individual cell populations and, therefore, captures the gene expression at the single-cell level.

The RNA-Seq method has revolutionised transcriptomics by measuring the quantity of RNA molecules through the sequencing of cDNA and allowing the entire transcriptome to be surveyed in a very high-throughput and quantitative manner. However, like other high-throughput sequencing technologies, the data analysis workflow is complex and computationally expensive since it consists of the processing

of large amounts of data and many pre-processing steps to transform the single sequences into gene counts. The ability to detect novel regions of the genome, to provide information on exon usage and sequence variations (i.e., single nucleotide polymorphisms), to reveal the precise location of transcription boundaries to a single-base resolution [134], the wide dynamic range and capability to detect very low and high abundant transcripts, made RNA-Seq become the most widely used technique for transcriptomic studies and replace microarray for many applications. Moreover, RNA-Seq has a resolution of a single-base, low background noise, since DNA sequences can be unambiguously mapped to unique regions of the genome, and requires a low amount of RNA [134].

The transcriptomic techniques described in Sections 1.10.1, 1.10.2 and 1.10.3, have been used in the past few decades to discover and validate new biomarkers and biosignatures for childhood TB. An overview is given in the next section [Section 1.10.4].

1.10.4 Transcriptomic signatures for paediatric tuberculosis

Transcriptomic biomarkers and parsimonious signatures might be a powerful tool for TB diagnosis, prognosis and treatment outcome, especially in low-resource clinical settings where the disease is endemic, and in children or adolescents where the clinical challenges are several. Most of the research studies that have been published on novel diagnostic, prognostic and treatment response biomarkers focused on adult TB. The few studies that focused on the disease in children are listed in Table 1.3. The diagnostic performance of biomarkers is usually measured in terms of Area Under the ROC Curve (AUC), sensitivity and specificity, which are described in detail in Chapter 2. Since Table 1.3 describes the performance of biomarkers and signatures in terms of AUC, sensitivity and specificity, it is necessary to give a very brief description of them in this section. AUC, sensitivity and specificity are percentages calculated based on the true positive and the false positive rates. The higher they are, the better the performance is. The Confidence Interval (CI) reflects the variation of the estimated measure.

Table 1.3: Gene expression biomarkers and/or biosignatures for child and adolescent TB published to date. Adapted from [84].

Publication	Country	Description of the study	Clinical cohort	Main findings
Verhagen (2013) [135]	Venezuela	Discovery (microarray) and validation (RT-qPCR) of gene expression signatures to distinguish children with TB from children with <i>Mtb</i> infection, healthy and with non-TB pneumonia (validation only)	Discovery cohort: 9 TB patients, 9 with <i>Mtb</i> infection, 9 healthy controls; Validation cohort: 20 with <i>Mtb</i> infection, 16 healthy controls and 18 with non-TB pneumonia	A 116-gene signature for TB vs. <i>Mtb</i> infection and healthy controls, and a 9-gene signature for TB vs. <i>Mtb</i> infection showed an average prediction error of 11%. An optimised set of 5 genes had 4% false positive rates for healthy controls and 11% for non-TB pneumonia cases
Dhanasekaran (2013) [136]	India	Whole-blood mRNA from 210 children was examined by dcRT-MLPA for the expression of 45 genes	13 children with TB disease, 90 with <i>Mtb</i> infection, 107 uninfected controls	A single gene discriminated between TB and <i>Mtb</i> infection (AUC of 78%), a 5-gene signature discriminated TB disease from controls (AUC of 92%). An 11-gene signature distinguished <i>Mtb</i> infection from controls (AUC of 72%)

Continues on next page

Publication	Country	Description of the study	Clinical cohort	Main findings
Anderson (2014) [137]	South Africa, Malawi, Kenya	Discovery and validation of gene expression biomarkers from microarray data to distinguish TB from <i>Mtb</i> infection and TB from other diseases	149 children with microbiologically confirmed TB, 44 with unconfirmed TB, 71 with <i>Mtb</i> infection and 139 with other diseases	A 51-transcript signature for TB vs. other diseases (sensitivity of 82.9% and specificity of 83.6% in the independent validation cohort) and a 42-transcript signature for TB vs. <i>Mtb</i> infection (sensitivity of 94.3% and specificity of 100.0% in the independent validation cohort)
Li (2015) [138]	China	Quantification by RT-qPCR of 7 genes in PBMC after ESAT-6 stimulation in children with PTB, EPTB and healthy controls	39 children with TB (3 smear culture positive and 26 culture negative) and 25 healthy controls	The expression of <i>IL-9</i> separated children with TB vs. healthy controls with an AUC of 92% after ESAT-6 stimulation
Wang (2015) [139]	China	RT-qPCR was used to quantify <i>miR-31</i> expression in PBMCs from children with TB and healthy controls	65 children with TB and 60 healthy controls	The expression of <i>miRNA-31</i> distinguished children with TB from healthy controls with sensitivity of 98.5% and a specificity of 86.7%
Zak (2016) [140]	South Africa, The Gambia	Discovery using blood RNA-Seq, and validation (RNA-Seq and RT-qPCR) of a signature to predict progression of <i>Mtb</i> infection to disease in adolescents (12-18 years)	46 progressors and 107 matched controls	A 16-gene signature for TB progression which had a sensitivity of 53.7% and specificity of 82.8% in the 12 months preceding TB in independent South African and Gambian cohorts
Fletcher (2016) [141]	South Africa	Infants were vaccinated with BCG at birth and followed for 2 years. Blood was collected at 10 weeks. Host responses from the 10-week samples were compared between those who developed TB disease within 2 years and controls who remained healthy	5,726 infants were recruited. 29 cases of microbiologically confirmed TB were compared to 110 controls (55 household controls and 55 community controls)	Gene expression analysis did not show a difference between cases and controls
Jenum (2016) [142]	India	Targeted analysis of transcriptional immune biomarkers in <i>Mtb</i> -antigen stimulated whole blood using dcRT-MLPA assay	88 children with intra-thoracic TB (6 months - 15 years); 40 microbiologically confirmed, 48 unconfirmed and 39 asymptomatic	An 8-gene biomarker signature separated children with TB from asymptomatic siblings (AUC 88%) in stimulated blood. 12 genes were found associated with clinical groups towards culture-positive TB or towards a decreased likelihood of TB disease on the TB disease spectrum
Zhou (2016) [143]	China	Identification of circulating miRNAs that can differentiate between TB and healthy controls	14 culture-positive TB cases, 14 culture-negative TB cases and 25 children with TB and 21 healthy controls for validation	An 8-miRNA signature provided 95.8% sensitivity and 100% specificity for the discrimination of children with TB vs. uninfected healthy controls

Continues on next page

Publication	Country	Description of the study	Clinical cohort	Main findings
Gjoen (2017) [144]	India	Selection and optimisation of 2 signatures for TB vs. asymptomatic household controls, and other symptomatic non-TB cases, from a set of 198 genes using dcRT-MLPA assay	71 TB cases (36 definite/35 probable) and 36 asymptomatic household controls, and 26 symptomatic non-TB cases	A 7- and a 10-transcript signature with AUC of 94% in separating TB cases from symptomatic non-TB cases regardless of culture status, and 100% sensitivity for definite TB
Hemingway (2017) [145]	South Africa	Longitudinal microarray blood gene expression analysis in children with TBM and comparison with children with PTB	9 children with TBM, (4 timepoints) and 9 healthy controls; 13 children with TBM and 28 with PTB	Reduced abundance of 68% of SDE genes in TBM vs. healthy controls. The difference in abundance was less in PTB than in TBM
Rohlwink (2019) [146]	South Africa	RNA-Seq on whole blood as well as on ventricular and lumbar cerebrospinal fluid of paediatric patients treated for TBM	20 TBM cases, 7 non TBM infection controls, and 24 healthy controls	2230 genes were SDE in TBM cases vs. healthy controls in blood, and 312 genes were SDE in ventricular cerebrospinal fluid in TBM vs. infection controls. TB disease processes differ between the periphery and the central nervous system, and within brain compartments
Penn-Nicholson (2020) [147]	South Africa, The Gambia, Ethiopia, Peru, Brazil	Identification of a parsimonious signature from the RNA-Seq [140] data using RT-qPCR data, and subsequent validation as a signature for diagnosis, progression and treatment response	46 progressors and 107 matched controls (Adolescent cohort study)	A 6-gene transcriptomic signature of TB disease risk, diagnosis and treatment response
Tornheim (2020) [148]	India	Longitudinal RNA-Seq from whole blood in cases during treatment and controls for the identification of SDE genes	16 TB cases and 32 TB-exposed controls	A 71-gene signature for TB diagnosis and a 25 gene signature for treatment response
Johnson (2021) [149]	India	Performance of TB gene signatures in malnourished individuals (including children) with TB and <i>Mtb</i> infection	23 severely malnourished individuals with TB and 15 severely malnourished TST positive household contacts	4,913 SDE protein-coding genes in TB vs. <i>Mtb</i> infection in malnourished individuals; 56.9% of the genes overlap with the 45 TB gene signatures included in the paper

Transcriptomic signatures for diagnostic use

Transcriptomics has been widely used to answer several research questions and identify key molecules in biological processes or diseases. Transcriptomic biomarkers can provide an understanding of the dynamics of infection or inflammation, monitor the response to treatments, give insights into the prognosis, or diagnose diseases.

The first transcriptomic paediatric signature for diagnosing TB in children was published in 2013 [135]. In this study a 116-gene signature was identified on a training set of 27 HIV-uninfected Warao Amerindian children (9 TB disease cases, 9 *Mtb* infection cases and 9 healthy controls) and validated in publicly available adult datasets [150, 151]. The 116-gene signature was then reduced to 10 genes by running random forest bootstrapping and was validated by RT-qPCR in the discovery cohort comparing

the 9 TB cases to an independent cohort of 54 children (20 *Mtb* infection cases, 16 healthy controls and 18 non-TB pneumonia cases). Decision tree analysis indicated that 5 genes (S100 Calcium Binding Protein P [*S100P*], Hemoglobin Subunit Delta [*HBD*], Phosphatidylinositol Glycan Anchor Biosynthesis Class C [*PIGC*], Cholinergic Receptor Muscarinic 2 [*CHRM2*], and Acyl-CoA Thioesterase 7 [*ACOT7*]) were contributing the most to the signature since they were able to correctly classify 78% of the TB cases with 11% of false positive rate for the non-TB pneumonia cases, 4% for the healthy controls and 0 for the *Mtb* infection group.

The second study searching for a paediatric transcriptomic TB signature was published in 2014 [137]. In this study, the authors identified a transcriptomic signature to differentiate microbiologically confirmed TB disease from *Mtb* infection, and a second signature to distinguish microbiologically confirmed TB from other respiratory infections with similar clinical manifestations. In this work, microarrays were used on a cohort of HIV-infected and HIV-uninfected children (<15 years of age) from 3 African countries (South Africa, Malawi and Kenya). The elastic net algorithm was used on the discovery cohort (South Africa and Malawi) to select the combination of genes with the highest diagnostic accuracy, and the findings were then validated in the independent Kenyan cohort. The 51-transcript signature selected to discriminate TB disease from other respiratory infections had a sensitivity of 82.9% (CI_{95%} 68.6%-94.3%) and specificity of 83.6% (CI_{95%} 74.6%-92.7%) in the validation cohort. Following a similar pipeline, a 42-transcript signature discriminated TB disease from *Mtb* infection with a sensitivity of 94.3% (CI_{95%} 85.7%-100.0%) and a specificity of 100% in the Kenyan cohort.

Most of the studies on diagnostic TB gene signatures in children published after 2014 were based on targeted techniques and included just a limited set of genes or transcripts. In 2017, a dual color Reverse Transcription Multiplex Ligation dependent Probe Amplification (dcRT-MLPA) assay was used to assess the performance of 198 genes (including 4 reference genes used to normalise the data) on a training cohort of 19 children with microbiologically confirmed TB, 28 children with unconfirmed TB disease, and 36 asymptomatic household controls [144]. dcRT-MLPA is a high-throughput technique that can rapidly profile mRNA expression of multiple host genes with a dynamic range and sensitivity comparable to RT-qPCR and RNA-Seq [152]. Two different feature selection algorithms were used to avoid overfitting. The first statistical approach used Least Absolute Shrinkage and Selection Operator (LASSO) [153] on the 198 genes and on the training set selecting a 10-transcript signature. The second statistical approach consisted of a combination of logistic regression, used as a filter to retain just the biomarkers with a cross-validated prediction ≥ 0.7 , and LASSO, used as a feature selection algorithm. The two identified signatures were then evaluated in the test set (17 children with microbiologically confirmed TB, 30 children with unconfirmed TB disease, and 26 symptomatic non-TB cases), without any retrospective optimisation. The 10-transcript signature achieved an AUC of 94% (CI_{95%} 88%-100%) in the test set, with a sensitivity of 91.7% (CI_{95%} 71.5%-98.5%) and a specificity of 88.5% (CI_{95%} 68.7%-96.9%). The 7-transcript signature achieved an identical performance - AUC of 94% (CI_{95%} 88%-100%) in the test set, with a sensitivity of 91.7% (CI_{95%} 71.5%-98.5%) and a specificity of 80.8% (CI_{95%} 60.0%-92.7%).

In 2016, a multicohort meta-analysis of 14 datasets identified and validated a 3-gene signature (guanylate binding protein 5 [*GBP5*], dual specificity phosphatase 3 [*DUSP3*], and Krüppel-like factor 2 [*KLF2*]) [154] on 2572 samples from 10 countries. The training cohort included only HIV-uninfected adults, with paediatric and HIV-infected patients only included in the validation cohort. The 3-gene signature had an overall AUC of 84.0% (CI_{95%} 80.0%-95.0%) in discriminating between TB disease

versus other diseases, and an AUC of 88.0% (CI_{95%} 84.0%-92.0%) in discriminating between TB disease versus *Mtb* infection. The 3-gene signature achieved a sensitivity of 68.6% and a specificity of 74.0% in the paediatric South African and Malawian cohort [137] when contrasting TB vs. other diseases, and respectively of 77.1% and 87.5% in the Kenyan cohort [137]. The lower performance of the 3-gene signature in the paediatric cohort compared to its performance in the adult cohorts highlights the need for new studies and new paediatric-specific TB signatures that account for the differences in the nature of the disease and host immune response. This 3-gene signature is the only one that have been translated into a POCT prototype so far. Cepheid (the Xpert MTB Host Response [MTB-HR] prototype) developed a fingerstick blood test that generates a score based on the mRNA of the 3 genes [155]. The Cepheid GeneXpert MTB-HR prototype was tested on 195 adults with symptoms compatible with TB recruited in South Africa, The Gambia, Uganda, and Vietnam. The prototype discriminated between TB disease and unlikely TB with an AUC of 94% (CI_{95%} 91%-97%), sensitivity of 87% (CI_{95%} 77%-93%) and specificity of 94% (CI_{95%} 88%-97%). When sensitivity was set at 90%, specificity was 86% (CI_{95%} 75%-97%), meeting the WHO TPPs for a TB triage test.

Many more studies focused on discriminating between paediatric TB disease and *Mtb* infection or healthy uninfected control in unstimulated or antigen-specific stimulated samples (Table 1.3). These two comparisons are clinically less relevant, but important as a proof-of-concept. Since this is a comparison between disease status and health status, the host response in the disease group is more perturbed compared to the healthy controls and fewer genes are needed to achieve optimal performance.

To better understand how important a new diagnostic tool for TB in children is, and also to put these novel transcriptomic signatures into context, an overview of the performances of the current diagnostic tests for paediatric TB is given below.

- The **chest X-ray** can provide a sensitivity of 71.4% (CI_{95%} 58.8%-84.1%) and specificity of 50.0% (CI_{95%} 41.9%-58.1%) when comparing microbiologically confirmed TB versus unlikely TB in children living with HIV [156]
- **IGRA assays** achieve a sensitivity of 75.0% (CI_{95%} 63.0%-84.0%) in diagnosing microbiologically confirmed TB in children, and a sensitivity of 66.0% (CI_{95%} 53.0%-68.0%) when the microbiologically confirmed and unconfirmed TB cases are pooled together [157]
- **Xpert MTB/RIF** achieves a sensitivity of 64.4% (CI_{95%} 55.3%-72.9%) and specificity of 99% (CI_{95%} 98.1%-99.5%) in diagnosing microbiologically confirmed PTB cases in sputum samples, but sensitivity and specificity change, respectively to 19.7% (CI_{95%} 12.1%-30.4%) and 100.0% (CI_{95%} 99.8%-100.0%), when both microbiologically confirmed and unconfirmed TB children are included [158]
- **Xpert MTB/RIF Ultra** achieves a sensitivity of 72.8% (CI_{95%} 64.7%-79.6%) and specificity of 97.5% (CI_{95%} 95.8%-98.5%) in diagnosing microbiologically confirmed PTB cases in sputum samples, but sensitivity and specificity change, respectively to 23.5% (CI_{95%} 20.0%-27.4%) and 99.2% (CI_{95%} 96.9%-99.8%), when both microbiologically confirmed and unconfirmed TB children are included [158]

In 2022, the WHO suggested a clinical decision-making algorithm to diagnose PTB in children younger than 10 years of age with a sensitivity of 85.0% and specificity of 37.0% [15].

Transcriptomic signatures as prognostic tool

A prognostic tool to identify children who are at higher risk to progress from *Mtb* infection to TB disease would help reduce the TB burden and meet the WHO *End TB Strategy* [74] goals. In 2016, 153 South African adolescents were recruited into a study and whole blood samples were collected across different time points prior to TB diagnosis [140]. The biological samples underwent RNA-Seq and 16 gene biomarkers were identified and combined into a signature with the aim of discriminating between progressors and non-progressors. The performance of the signature was inversely proportional to the time to diagnosis - the signature was performing better when the samples were collected closer to the disease diagnosis. The signature achieved an AUC of 79.0% (CI_{95%} 76.0%-82.0%) at 6 months prior to diagnosis. The 16-gene signature was then cross-platform validated by RT-qPCR in the discovery cohort, and validated by the same targeted technique in other African adult cohorts (from South Africa, Ethiopia and The Gambia).

In 2018, the 16-gene signature [140] was reduced to a 6-gene signature (RISK6) by selecting the smallest set of transcripts with the best prognostic performance based on RT-qPCR data [147]. The signature achieved an AUC of 87.6% (CI_{95%} 82.8%-92.4%) in the samples of the discovery cohort collected within 12 months from the disease diagnosis. The AUC of the signature dropped to 74.0% (CI_{95%} 66.0%-82.0%) in samples collected more the 12 months before the disease diagnosis. The findings were finally validated in an adult multi-country cohort [159], and a similar performance was described.

These two studies describe how progressors, especially closer to the disease diagnosis, have a host response and gene profile similar to TB cases, and furthermore, show that the closer the disease diagnosis is, the more similar to TB disease the host response is.

Currently, there are no reliable tools to assess the risk of a child progressing from *Mtb* infection to TB disease. Children younger than 5 years of age with a positive IGRA or TST are at higher risk to progress to TB disease compared to children that test negative, but only the 20% of the IGRA/TST positive will progress to TB disease [90]. Therefore, the current prognostic tests used have a high sensitivity but very low specificity [84].

Transcriptomic signatures to assess the tuberculosis treatment response

To improve disease outcomes and identify treatment failure, better biomarkers are needed, since the currently available tests have very low accuracy in both adults and children [160]. The conversion of sputum smear or culture from positive to negative after 2 months of treatment is the most reliable test currently used in adults. Although their sensitivity and specificity are limited (24%-40% and 85% respectively) [160], especially in children. The growing interest in host-based TB diagnostic has led to the discovery of genes to monitor TB treatment response. A 5-gene signature (SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily D, Member 3 [*SMARCD3*], Uncoupling Protein 2 [*UCP2*], MAP7 Domain Containing 3 [*MAP7D3*], STT3 Oligosaccharyltransferase Complex Catalytic Subunit A [*STT3A*], and *RP11-295G20.2*) was identified to complement GeneXpert in monitoring TB treatment and predicting failure in adults [161]. A 6-gene signature (Guanylate Binding Protein 2 [*GBP2*], Fc Gamma Receptor 1b [*FCGR1B*], Serpin Family G Member 1 [*SERPING1*], Tubulin Gamma Complex Component 6 [*TUBGCP6*], TRNA Methyltransferase 2 Homolog A [*TRMT2A*], and Short Chain Dehydrogenase/Reductase Family 39U Member 1 [*SDR39U1*]) was also evaluated

to track the treatment response in adults [147, 159]. However, studies in paediatric cohorts are very limited.

A recent longitudinal study published in 2020 [148], identified a list of 25 Significantly Differentially Expressed (SDE) genes for treatment response in 16 Indian children with TB disease and 32 Indian TB exposed children. All the 25 SDE genes were overexpressed at baseline. The list was then evaluated in three adult cohorts and achieved AUCs of respectively 90.0% (CI_{95%} 66.0%-82.0%) [161], 72.0% (CI_{95%} 85.0%-94.0%) [162] and 50.0% (CI_{95%} 36.0%-63.0%) [163].

1.11 Measuring protein abundance

In the past few decades, proteomics has gained increasing importance in monitoring disease-related changes in the protein composition of biological fluids such as blood plasma, serum, bronchial lavage fluid or urine. Proteomics typically gives a better understanding of an organism. In fact, the level of transcription of a gene gives only a rough estimate of its level of expression into a protein. An mRNA produced in abundance may be degraded rapidly or translated inefficiently, resulting in a small amount of protein. However, proteomic techniques have a limited dynamic range compared to some transcriptomic techniques. RT-qPCR, microarrays and RNA-Seq, for instance, have an amplification step where the cDNA is amplified and can therefore be measure more easily.

Human plasma and serum are the most frequently used biological samples since they are more stable and show less variability according to the time of collection and individual habits compared to other biological fluids. Serum and plasma represent a rich source for protein biomarker identification [164]. Plasma, which represents about 55% of the total blood volume, is the liquid component of the blood in which blood cells are suspended. An anticoagulant is usually added to the samples immediately after collection to avoid clotting, followed by gentle centrifugation to remove the blood cells [165, 166]. Serum collection, on the other hand, is done in the absence of an anticoagulant and involves clot activator and centrifugation. The centrifugation step is needed to remove the blood cells and the fibrin clots. Plasma and serum, therefore, have different proteomic profiles since plasma contains fibrinogen protein and other clotting factors that are removed from serum. The choice to collect plasma or serum usually depends on the purpose of the study. In some cases, serum can be more useful because of the absence of an anticoagulant, in other cases, plasma collection might be preferred for its higher stability or for the study of the clotting cascade [167].

Blood constituents, primarily proteins, reflect diverse physiological and pathological states, therefore they are of great clinical significance. Upon infection by *Mtb*, specific proteins are released by both the bacilli and the host-infected tissues. Most of these released proteins are measurable in plasma and serum, however, the enormous complexity and dynamic range of plasma and serum proteins make the detection of potential low-abundant biomarkers difficult due to the presence of very high-abundant proteins [164]. Moreover, some serum and plasma proteins are highly variable in concentration between individuals and might obscure disease-specific biomarkers. To control this variability, large and heterogeneous cohorts of samples need to be analysed and the cellular source of the proteins of interest needs to be taken into account (i.e., secreted proteins should be targeted).

The recent developments in proteomics, especially in quantitative proteomics, allow the identification of novel biomarkers with high diagnostic accuracy. Techniques for detecting and quantifying proteins present in a biological sample can be classified, likewise transcriptomic techniques, in untargeted or

targeted, in high- and low-throughput methods. The most widely used proteomics techniques for biomarker discovery and validation are described in detail in Sections 1.11.1, 1.11.2 and 1.11.3.

1.11.1 The Enzyme-Linked Immunosorbent Assay (ELISA)

The most used tool, and the gold standard of clinical laboratories for the detection of single analytes, is the Enzyme-Linked Immunosorbent Assay (ELISA) [168]. This immunological assay is commonly used to detect and quantify soluble substances such as antibodies, antigens, hormones, peptides, proteins and glycoproteins in biological samples. In ELISA assays, the antigen is immobilised on a solid surface, making it easy to wash away non-bound material. Although many variants of ELISA assays have been developed (i.e., direct, indirect, competitive and sandwich), the sandwich ELISA is the most widely used for its high specificity, sensitivity and flexibility. A sandwich ELISA measures an antigen between two layers of antibodies (capture and detection antibody). The detection antibody is linked to a reporter enzyme, of which activity is finally measured and it is proportional to the abundance of the protein of interest (signal measurement).

ELISA assays are generally carried out in 96-well or 384-well plates, allowing multiple samples to be measured in a single experiment, and they do not need sample pretreatment [169]. Each ELISA quantifies a specific protein, and kits for a variety of proteins are widely available. The high specificity comes from the highly specific antibody-antigen interaction.

The main limitation of this assay is its characteristic to be singleplex, therefore multiple ELISA immunoassays need to be performed to measure multiple proteins. This becomes tedious and impractical owing to substantial demands on sample volume requirement (which can be as high as 50 μL of sample per analyte), turnaround time and cost.

1.11.2 The Luminex® assay

A Luminex® assay is an antibody-based, targeted technique used for detecting up to 100 protein analytes in biological fluids. This technology is a bead-based immunoassay and is summarised in Figure 1.14. It uses colour-coded beads internally dyed with different proportions of red and infrared fluorophores that correspond to a distinct spectral signature or bead region. The sample is incubated with a mixture of magnetic beads, pre-coated with analyte-specific capture antibodies that bind to the analytes of interest. Detection antibodies are then added to the mixture in order to bind to the specific bead-protein complexes forming an antibody-antigen sandwich. The magnetic beads are read on a dual-laser flow-based detection instrument, able to sort the beads by their spectral properties and fluorescence. One laser determines the analytes that have been detected and a second laser determines the magnitude of the fluorescence, which is proportional to protein abundance.

The output of the assay is the fluorescence emitted by the detection antibodies bound to the analytes of interest. Through a 5-parameter logistic regression, the concentrations of the target analytes are calculated from the fluorescence detected and a standard curve (concentration versus fluorescence) is generated for each protein of interest. The standard curve also sets the dynamic range and limits of detection of each protein in the experiment.

Luminex® assays are widely used mainly in validation experiments, where a few target proteins need to be quantified. They require a small sample volume since multiple analytes can be measured in a single well with less than 25 μL , minimise experimental variability with multiple data points derived

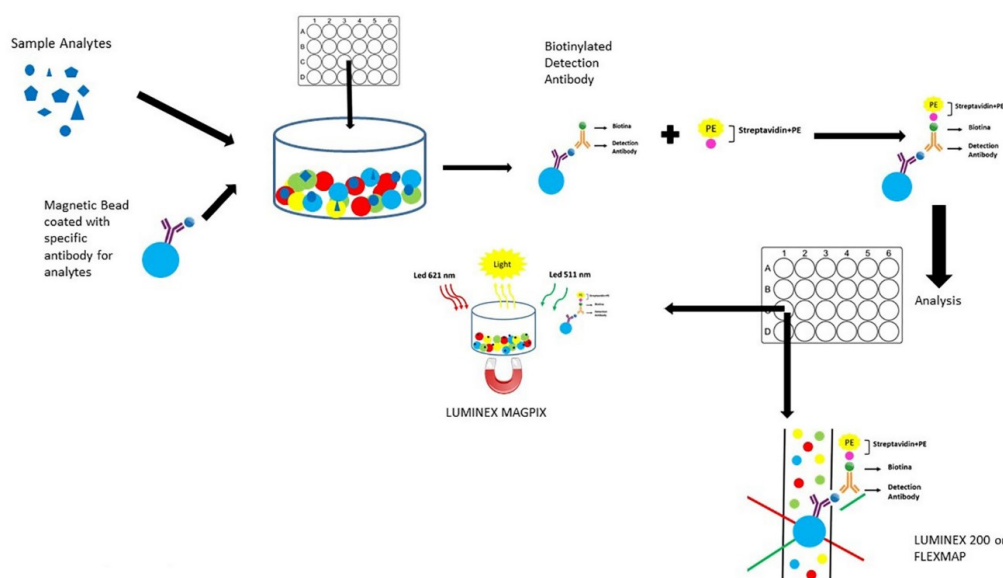


Figure 1.14: Overview of the Luminex® assay [170].

In step 1, the sample is incubated with a mixture of magnetic beads, pre-coated with analyte-specific capture antibodies that bind to the analytes of interest. In step 2, the biotinylated detection antibodies are added to the mixture in order to bind to the specific bead-protein complexes forming an antibody-antigen sandwich. Phycoerythrin (PE)-conjugated streptavidin is then added, and it binds to the biotinylated detection antibodies. In step 3, the magnetic beads are read on a dual-laser flow-based detection instrument, able to sort them by their spectral properties and fluorescence and detect the protein abundance.

from a single experiment, reducing time and cost compared to singleplex ELISA, and custom plates can be designed.

1.11.3 The SomaScan® assay

The SomaScan® assay is a targeted, multiplexed proteomic tool produced by SomaLogic. It currently detects more than 7,000 protein analytes in a small quantity of biological fluid such as serum, plasma or cerebrospinal fluid. The data analysed in this thesis were generated by the SomaScan® 1.3K multiplexed assay, which contained 1,301 human proteins and is the assay described below. This technology relies upon protein-capture Slow Off-rate Modified Aptamer (SOMAmer®) molecules [171]. Aptamers are artificial and short single-stranded DNA or RNA that, like antibodies, bind to their targets with high affinity and specificity. Aptamers are more stable than antibodies, more temperature resistant (stable at room temperature) and have a longer shelf life. Moreover, they are produced through an inexpensive, simple and short process, and unlike antibodies, aptamers do not need an immune response for their production [172]. Since aptamers are chemically synthesised, there is less batch-to-batch variation across experiments. Chemical modifications at any desired position in the nucleotide chain can also be introduced to optimise their target recognition and stability. Although antibodies can be chemically modified, site-specific modifications are extremely difficult [173, 174]. Most aptamers originate from Systematic Evolution of Ligands by EXponential enrichment (SELEX), which are a family of *in-vitro* experiments that select the best aptamers from a starting DNA library [171].

The SomaScan® assay is highly sensitive and highly multiplexed given the dual nature of nucleic acid ligands: they spontaneously fold into precise three-dimensional structures very specific to their targets through shape complementarity, and they are also capable of highly specific hybridisation

to complementary sequences arrayed on surfaces. The assay is run at three different dilutions (i.e., 0.005%, 1% and 40%) to measure high-, medium- and low-abundant proteins.

The SomaScan® assay is high-throughput, specific, sensitive and with a large dynamic range. It currently measures more than 7,000 proteins from a 55 μ L sample and, since it is not based on variable polyclonal antibodies, the results are reproducible and with a low coefficient of variation between runs. Its main limitations are the high cost, the sequences of aptamers that are not publicly available, and the limited flexibility since the assay cannot be run locally but samples need to be shipped to SomaLogic.

1.11.4 Proteomic signatures for paediatric tuberculosis

Proteomics has gained increasing interest in the past few decades because proteins are more complex, dynamic, and reflect biological functions more closely than genes. However, protein biomarker research requires the analysis of complex biological fluids (i.e., serum and plasma) with some biomarkers of interest often hard to detect due to their very low concentration compared to thousands of other proteins present in the biological sample. Adult pulmonary TB studies have shown the potential of blood-based host protein biomarkers in the diagnosis of TB disease. 74 host biomarkers were evaluated in 55 South African adults with or without HIV co-infection [175]. A signature of 8 biomarkers (i.e., NCAM, CRP, SAP, IP-10, ferritin, TPA, I-309, and MIG) showed a diagnostic AUC $\geq 80\%$ in discriminating TB disease from other respiratory disease. A seven-marker signature of C reactive protein, transthyretin, IFN- γ , complement factor H, apolipoprotein-A1, inducible protein 10 and serum amyloid A was identified and tested in a cohort of 716 adults recruited in 5 African countries [176]. The signature had a sensitivity of 93.8% and a specificity of 73.3% in the test set regardless of HIV status. Host-derived biomarkers, previously identified in QuantiFERON supernatants, were also investigated in a large pan-African cohort [177]. A signature comprising unstimulated IFN- γ , MIP-1 β , TGF- α and antigen-specific levels of TGF- α and VEGF had an AUC of 81%, sensitivity of 64.2% and specificity of 82.7% in the test set respectively.

Studies in children are either lacking, do not compare TB cases to relevant control groups, or have a very small sample size. A list of published protein biomarkers or signatures for paediatric TB is reported in Table 1.4.

Table 1.4: Protein biomarkers and/or biosignatures for child and adolescent TB published to date.

Publication	Country	Description of the study	Clinical cohort	Main findings
Alsleben (2012) [178]	Germany	<i>Mtb</i> antigen-stimulated IP-10 response in children with TB disease and <i>Mtb</i> infection was compared to the response from children with Nontuberculous Mycobacteria (NTM) lymphadenopathy and Respiratory Tract Infection (RTI) by ELISA. Agreement between IP-10 and QuantiFERON-TB Gold In-Tube test results was assessed, along with IP-10 association with age	11 children with TB disease, 14 with <i>Mtb</i> infection, 8 with NTM lymphadenopathy, 15 with RTI	Antigen-stimulated IP-10 and IFN- γ plasma concentrations were highly correlated and significantly higher in the TB disease and <i>Mtb</i> infection groups compared to the NTM and RTI groups. No significant difference was observed between children with TB disease and <i>Mtb</i> infection in the IP-10 or IFN- γ plasma concentrations. Compared to IFN- γ , the IP-10 levels were higher and IP-10 was released independently of age

Continues on next page

Publication	Country	Description of the study	Clinical cohort	Main findings
Chegou (2013) [179]	South Africa	12 host markers previously identified as potential TB diagnostic markers were evaluated in 76 children (29% HIV-infected) to assess their diagnostic power	19 children with microbiologically confirmed TB disease, 57 TB-exposed children (26 QuantiFERON-TB In-Tube positive and 31 <i>Mtb</i> uninfected)	IFN- α 2, IL-1Ra, sCD40L and VEGF were identified as the best biomarkers to diagnose TB disease. IL-1Ra, IP-10 and VEGF were identified as the best biomarkers to discriminate between TB and <i>Mtb</i> infection in QuantiFERON-TB In-Tube positive children. When markers were used in combinations of four, 84% of all children were accurately classified into their respective groups (TB disease or no TB), after leave-one-out cross validation
Pavan Kumar (2013) [180]	India	Circulating tissue inflammation, oxidative stress, innate immune activation, fibrosis markers and cytokine response were examined by immunoassays to assess their differences between children with PTB or EPTB and healthy controls	36 children with TB disease (14 with PTB and 22 with EPTB), 19 healthy controls	Levels of MMP-1, TIMP-3, CRP, HO-1, IL-21 were significantly higher in children with PTB and EPTB compared to healthy controls. MMP-7, MMP-8, TIMP-1, alpha-2m, haptoglobin, IL-23, and TGF- β were significantly higher in children with PTB compared to healthy controls. MMP/TIMP, acute-phase proteins, and HO-1 levels in PTB children were not different from their levels in EPTB children, with the exception of MMP-7. LPS and LBP were significantly higher in children with PTB compared to children with EPTB
Latorre (2014) [181]	Spain	IP-10 was evaluated as a biomarker for TB disease and <i>Mtb</i> infection, and compared to QuantiFERON-TB Gold In-Tube (QFN-G-IT) and T-SPOT.TB test results	12 children with TB disease, 81 healthy but with TB-contact and 137 healthy controls	IP-10 assay had comparable performance to QFN-G-IT and T-SPOT.TB tests for diagnosis of <i>Mtb</i> infection. When QFN-G-IT and IP-10 assays were combined, the diagnostic power increased compared to the independent assays. After antigen-specific stimulation, the levels of IP-10 were statistically higher in the TB disease group compared to the other two groups

Continues on next page

Publication	Country	Description of the study	Clinical cohort	Main findings
Tebruegge (2015) [182]	Australia	Levels of 18 proteins were measured by Luminex® after three different antigen-specific stimulations to identify <i>Mtb</i> -specific cytokine biomarkers able to diagnose TB disease and distinguish it from <i>Mtb</i> infection	6 children with TB disease, 16 with <i>Mtb</i> infection, 75 healthy controls and 28 with a TST induration >10 mm and a negative QFT-GIT assay result	IFN- γ , IP-10, TNF- α , IL-1 α , IL-2, IL-13 and MIP-1 β were statistically different between the four groups irrespective of the antigenic stimulant, with the highest concentration in the TB disease group. TNF- α , IL-1ra and IL-10 responses were the biomarkers with the greatest ability to discriminate between <i>Mtb</i> infection and TB disease
Sudbury (2019) [183]	Peru	Plasma levels of IFN- γ , IL-10, IL-13, IL-1ra, IL-2, TNF- α , IP-10 and MIP-1 β were measured by Luminex® to assess their performance as candidate biomarkers for TB in children	203 children (<15 years of age) were included in the study: 47 sick controls, 98 healthy controls, 5 unconfirmed TB and 53 <i>Mtb</i> infection. All the children were HIV-uninfected	IL-2, IL-13 and IP-10 emerged as the best signature to discriminate TB, either unconfirmed TB or <i>Mtb</i> infection, from the controls with a sensitivity of 84.2% (CI _{95%} : 72.1%-92.5%) and specificity of 82.4% (CI _{95%} : 74.9%-88.4%). IL-1ra and TNF- α levels were statistically different between unconfirmed TB and <i>Mtb</i> infection with an AUC respectively of 91.92% and 70.38%
Manyelo (2019) [184]	South Africa	69 host biomarkers were evaluated in serum samples by immunoassay (either ELISA or Luminex®) to find the best signature to distinguish Tuberculosis Meningitis (TBM) from other types of meningitis	23 children with TBM and 24 with other types of meningitis	A 3-protein signature (Complement factor D/adipsin, Ab 42 and IL-10) distinguished TBM from non-TBM samples with an AUC of 84%
Albuquerque (2019) [185]	India	Levels of cytokines, acute-phase proteins and tissue remodelling proteins were examined by Luminex® or ELISA to identify biomarkers that could accurately distinguish different PTB from EPTB and TB from healthy controls	14 children with PTB, 22 with EPTB and 19 healthy controls. All the children were HIV-uninfected	4 biomarkers (MMP-1, HO-1, TIMP-3 and IL-21) were significantly different between TB disease and healthy controls. The combination of 3 biomarkers (MMP-7, LBP and CRP) could distinguish PTB from EPTB and healthy controls with an accuracy of 87.5%

Continues on next page

Publication	Country	Description of the study	Clinical cohort	Main findings
Togun (2020) [186]	The Gambia	431 children were included in the study. 27 cytokines were run on Luminex® to find the best biomarkers to distinguish TB (microbiologically confirmed or unconfirmed) from other diseases with similar symptomatology	44 children with microbiologically confirmed TB, 60 with unconfirmed TB, 327 with other diseases with symptoms suggestive of TB. All children were HIV-uninfected	A 3-protein signature (IL-1ra, IL-7 and IP-10) was derived by using logistic regression analysis combined with LASSO penalty on the training set to compare microbiologically confirmed TB to the other disease group. The signature achieved an AUC of 74% (CI _{95%} : 67%-81%) with sensitivity of 72.2% (CI _{95%} : 60.4%-82.1%) and specificity of 75% (CI _{95%} : 64.9%-83.4%) in the test set
Druszczynska (2021) [187]	Poland	Levels of 53 Th17-related cytokines and inflammatory mediators in sera were analysed by immunoassay in 216 BCG-vaccinated children to identify biosignatures for TB and <i>Mtb</i> infection	15 children with TB, 50 with <i>Mtb</i> infection and 151 healthy controls. All children were HIV-uninfected and BCG vaccinated	The most discriminating cytokine between the 3 disease groups was MMP-2, with AUCs of 84.8%, 72.1% and 70.1% respectively for healthy controls versus TB, healthy controls versus <i>Mtb</i> infection and <i>Mtb</i> infection versus TB. Levels of BAFF/TNFSF13B, OPN and LIGHT/TNFSF14 were also significantly different between the three disease groups and had a good discrimination power
Pavan Kumar (2021) [188]	India	Plasma levels of 5 CC (CCL1, CCL2, CCL3, CCL4 and CCL11) and 5 CXC (CXCL1 CXCL2, CXCL9, CXCL10 and CXCL11) chemokines were measured in a paediatric cohort to assess their ability in diagnosing TB	44 children with microbiologically confirmed TB, 47 with unconfirmed TB and 76 with unlikely TB. All children were HIV-uninfected	CCL1, CCL3, CXCL1, CXCL2 and CXCL10 were significantly higher in the TB group (both microbiologically confirmed and unconfirmed TB) in comparison to unlikely TB children. CCL1, CXCL1 and CXCL10 showed sensitivity and specificity >80%, with CCL1 being the single marker with the highest diagnostic ability. The combination of the three markers showed optimal diagnostic performance. CCL1, CXCL1, CXCL2 and CXCL10 exhibited a significant reduction in abundance following anti-TB treatment

The first proteomic studies on TB biomarkers were mainly based on targeted proteomics to assess the discriminatory ability of single inflammatory markers between TB disease and other disease groups, such as *Mtb* infection or other respiratory infections. In a paediatric German cohort, Alsleben and colleagues [178] showed that IP-10 and IFN- γ levels were not significantly different between children with TB disease and *Mtb* infection, but their levels were higher in those two disease groups compared to children with Respiratory Tract Infection (RTI) or Nontuberculous Mycobacteria (NTM)

lymphadenopathy. The ability of IP-10 in distinguishing TB disease from *Mtb* infection in children was also assessed by Latorre and colleagues [181] in a Spanish paediatric cohort. While the study setting and cohort were different, the two research groups found very similar results. Chegou and colleagues [179] evaluated 12 previously identified proteomic biomarkers for TB disease in a paediatric South-African cohort. 29% of the children in the cohort were HIV-infected. A combination of 4 proteins showed good diagnostic power with 84% of the children correctly classified in their respective diagnostic groups. In an Indian paediatric study [180], the authors used immunoassays to assess which ones of a pool of markers were statistically different between children with PTB, children with EPTB and healthy controls. They showed that some markers were able to differentiate between PTB and EPTB, and some others between TB disease, regardless of the disease site, and healthy controls. Tebruegge and colleagues [182] assessed the diagnostic ability of 18 proteins in distinguishing TB disease and *Mtb* infection in children after antigen-specific stimulation of the samples. They found that 3 proteins (TNF- α , IL-1ra and IL-10) were best at discriminating between TB disease and *Mtb* infection, all of them with higher abundance in the TB group.

More recently, researchers have been focusing on proteomic signatures rather than single biomarkers, since single biomarkers do not reach the TPPs for a test for paediatric TB. However, there are very few published studies on proteomic signatures for paediatric TB disease [183–188] and none of the published diagnostic signatures meets the criteria for sensitivity and specificity published by the WHO.

1.12 Multi-omics data integration

1.12.1 Overview

New high-throughput technologies generate a series of -omics data, such as genomics, transcriptomics, proteomics and metabolomics, and provide opportunities to study different molecular levels in biological systems. Each layer of -omics provides functional and biological insights at a single molecular level. The biological molecules interact with each other forming complex networks, which can provide a comprehensive representation of the biological system. Therefore, there is a need to analyse more than one layer of -omics, either across experiments on the same samples (vertical integration), or across studies on the same variables (horizontal integration). In the past few years, several studies have shown that combining -omics datasets helps to better understand the biology of the systems. For instance, in a transcriptomics microarray study focusing on atherosclerotic lesions, it has been shown that the Renin-Angiotensin system needs to be studied on local tissue physiology to understand how it impacts the global organisation of the system in the human body as a whole and how it plays an important role in several diseases [189]. In the same study, the results have then been microbiologically confirmed and enriched with more elaborate information by analysing the molecules on protein and metabolic levels, giving more insights into the Renin-Angiotensin system. Another study showed how the integration of microarray data (mRNA and miRNA), proteomics and metabolomics from a small cohort improves the accuracy of molecular classification of chronic obstructive pulmonary disease compared to the single-omics analysis [190]. The mean accuracy of the group prediction increased linearly from 0.28 with one -omics dataset to 0.9 with the integration of all the -omics datasets. To analyse such a big variety of datasets, machine learning techniques have been successfully used to learn processing rules, identify patterns and make decisions. Machine learning, with network theory and statistical modelling, is capable of detecting hidden patterns in the molecular network. However,

multi-omics data integration is highly challenging due to its high computational cost, technical limits, the complexity of biological systems and the heterogeneity of several analytical platforms, as each -omics dataset captures a separate aspect of biology.

1.12.2 Multi-omics data

Multi-omics refers to the integrative study of the following single omic data:

- **Genomics** is the study of the genome of an organism. Its main aim is to determine the whole sequence of DNA and study its structure, function and evolution. Genome is the complete set of DNA (genes, noncoding DNA, mitochondrial DNA and chloroplast DNA) and it holds all genetic information of an organism [191]
- **Epigenome** consists of all the chemical compounds that have been added to the genome as a way to regulate gene expression [192]. The epigenetic changes are modifications that do not change the DNA sequence but affect gene activity. Whereas the individual genome remains static, the epigenome can be altered by environmental conditions [193]
- **Transcriptome** is the full range of transcripts in a cell or collection of cells [134]. The term usually refers to all the types of RNA, but it might sometimes refer just to mRNA. The transcriptome might be considered the expression of the genome since it captures the expressed genes in a given condition. It is a dynamic system responsive to external conditions. The main aim of transcriptomics is to identify genes that show a significant difference between different conditions
- **Proteome** is the set of proteins in an organism. It provides insights into protein abundance, post-translational modifications and protein-protein interactions, which carry biological information that is not accessible by genomics or transcriptomics [194]. The proteome also actively changes in response to external or biological conditions
- **Metabolome** is the complete range of all the low molecular weight metabolites produced by cells during metabolism. Properties and concentration of enzymes regulate the level of metabolites, and therefore they provide a direct functional image of cellular activity and physiological status [195]

The integration of all the different omics layers might explain the molecular dynamics underlying the pathophysiology of human diseases and lead to novel strategies for early detection or treatment of them [Figure 1.15].

1.12.3 Networks in molecular biology

Multi-omics interactions might be described and represented by multilayer dynamic networks. Networks are widely used in biology as an intuitive representation of relationships between different molecules of a biological system, where nodes are particles such as genes, RNAs and proteins, and edges are links between molecular pairs [197]. Multilayer networks are connected by both within-layer links and between-layer links [198], they have been introduced in molecular biology to understand the mechanisms from different molecular levels and integrate them in a single network [199]. They also provide a flexible and easy tool where each -omics can be described by a layer of the network and interactions between different -omics can be captured by the between-layer links [Figure 1.16]. As biological systems are dynamic entities responsive to different internal or external stimuli or changes

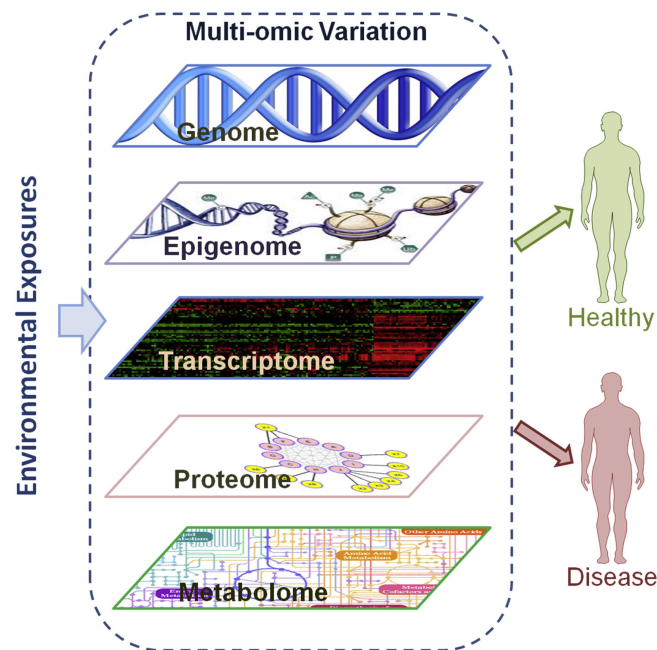


Figure 1.15: Conceptual model of multi-omics and human disease [196].

Overview of omics technologies and their ability to differentiate between a healthy individual and a patient with a specific disease.

over time, a dynamic multilayer network, which describes the state of the network at every time point, should be used to describe them.

1.12.4 Multi-omics data integration tools

The interest in multi-omics approaches led to the development of several tools, methods and platforms to analyse, visualise and interpret multi-omics data. The objectives of integrative analysis and biological questions can be summarised into the following [201]:

- Study of molecular interactions between different omics data types to discover new mechanisms about a biological process or a disease
- Clustering of samples in order to improve patient stratification and disease subtyping
- Prediction of an outcome based on predictive models to classify and stratify patients based on their -omics data

Most of the published algorithms focus on the first two objectives and on merging two types of -omics [202] with DNA sequencing data, gene expression and protein abundance being the most common. An extensive list of available multi-omics tools is represented in Figure 1.17, where they are coloured according to their aim (biomarker prediction/identification, disease subtyping or disease insights) and grouped by their approach (similarity, correlation, network, Bayesian, fusion and multivariate). Few tools like iClusterPlus and MixOmics [203] have more than one objective, and others such as SNF [<http://compbio.cs.toronto.edu/SNF/SNF/Software.html>; accessed February 10th, 2023] and PARADIGM [<http://paradigm.five3genomics.com/>; accessed February 10th, 2023] use a combination of the approach categories.

In -omics data integration, the presence of missing values is a common situation and difficult to handle because most statistical methods cannot be directly applied to incomplete datasets. Some integration

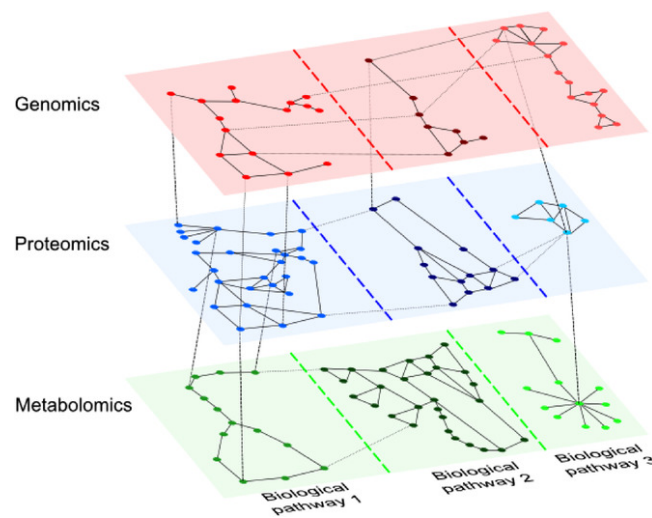


Figure 1.16: Example of a multilayer network used in the context of biology [200].

The different layers are the different types of data (i.e., genomic, proteomic or metabolomic), and the interactions between them and the biological pathways are the links between the different omics.

tools handle missing data using imputation methods, others require them to be removed or handled in the pre-processing steps. In both situations, the presence of missing values is often something to be addressed. Figure 1.17 organises the different tools into six groups:

- Similarity-based approaches label the samples of a test set based on their similarity to a training set. They are a generalisation of the minimal distance (MD) methods, so the similarity-based approaches do not require to access all the features of the samples, provided that the similarity function is well-defined [205]
- Correlation-based approaches aim to find the association between the features by using correlation methods
- Network-based approaches take into account the relationships, previously known or predicted, between biological variables and use graph measurements (i.e., degree and connectivity) and graph algorithms (i.e., sub-networks identification) to identify biological information
- Bayesian approaches predict the probability distribution from an inferred assumption about the initial probability distribution, by using statistical models and Bayes' rule
- Multivariate methods explain and clarify relationships among multiple variables simultaneously. The general purpose is to learn more about the relationship between several explanatory variables and one response variable
- Fusion-based approaches work on merging two or more raw datasets in order to have better performances

MixOmics [203], which can be used for biomarker prediction and disease subtyping, is becoming one of the most popular tools. It integrates multi-omics datasets to cluster samples by using dimension reduction and regression techniques, correlation or supervised analysis, furthermore, it uses several supervised multivariate algorithms to select the best biomarkers [203]. This integration tool is largely used for its range of multivariate methods and graphical outputs.

For the aim of this thesis, the MixOmics R package was used for identifying diagnostic multi-omics

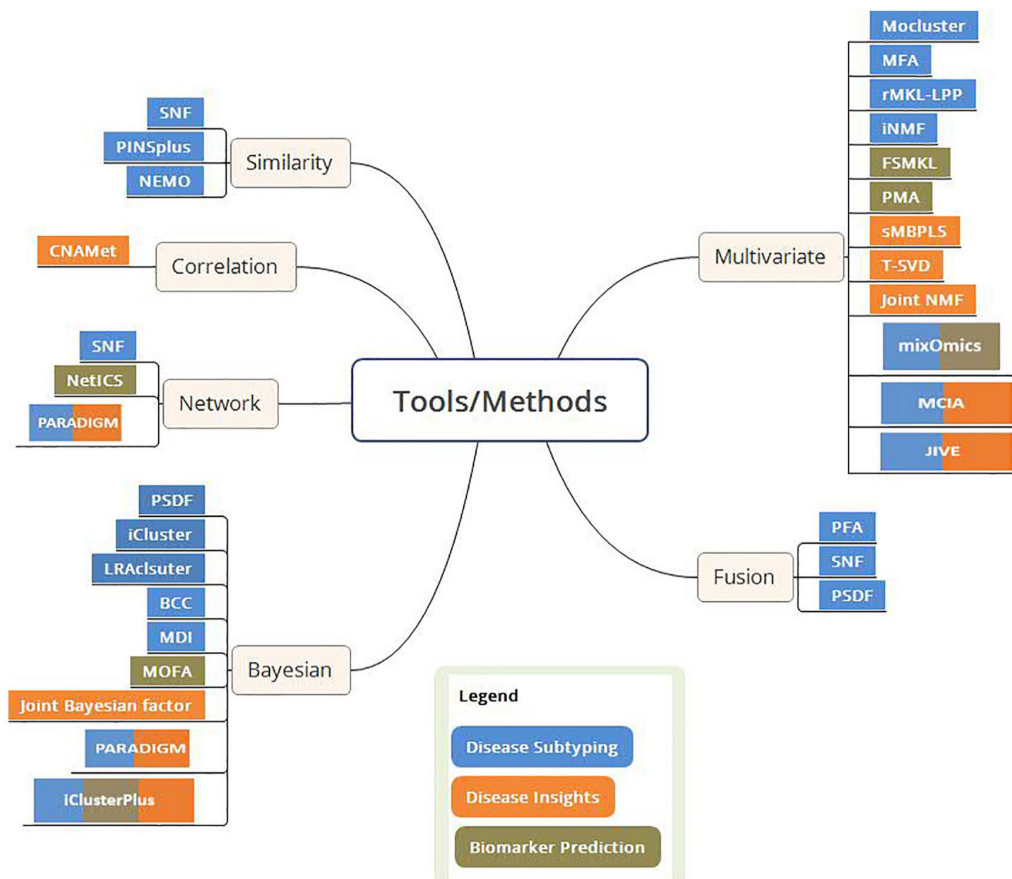


Figure 1.17: Overview of multi-omics data integration tools [204].

The tools/methods are grouped based on their approach and are coloured according to their applications.

biomarkers for paediatric TB. The data analysis pipeline is described in more detail in Section 2.8.

1.13 Summary and thesis contribution

Diagnosing TB, especially in children, is a major healthcare challenge worldwide. The gold standard diagnostic test remains the *Mtb* culture, which involves growing the bacteria on different substances and assessing whether they are present in the biological sample. The main advantages of *Mtb* culture are the high specificity (i.e., the correct diagnosis of patients not having TB) and the identification of the bacterium species. However, the culture test has a lower sensitivity, which causes true TB cases left undiagnosed, it is highly operator dependent, needs expensive laboratory facilities and reagents to be used, trained personnel and it requires up to 6 weeks for results due to the slow growth rate of the bacteria. Its sensitivity is even lower in children because of their low bacteria loads, making the microbiological diagnosis of this disease even harder. Diagnosis on clinical grounds (i.e., unconfirmed TB) in children is also challenging mainly due to a milder presentation of the disease compared to adults [82] and to symptoms similar to other paediatric diseases. Moreover, the collection of biological samples from children might be very invasive since they are unable to spontaneously expectorate and procedures such as induced sputum or gastric aspiration are often required [88].

Since the detection of the pathogen in biological samples is challenging, the focus has been moved to the detection of changes in the host immune system during infection or disease. The host in fact mounts an immune response following infection, changing the expression of specific genes or production of

specific proteins, which is unique to a specific foreign pathogen. Therefore, unique patterns of genes or proteins can be used as diagnostic biomarkers to develop whole-blood based POCT.

Host omic profiles have been shown to aid in the diagnosis of paediatric TB [135, 137, 140, 145, 146, 148, 149, 183–188]. Many of the published diagnostic TB signatures, however, consist of several genes or proteins, making the development of a POCT technically difficult and more expensive. A parsimonious diagnostic TB signature with high sensitivity and specificity is needed to develop a non-sputum-based, rapid and cheap diagnostic or triage test for LMIC where TB is still endemic.

In Chapter 3, I present the discovery and subsequent validation of a new 6-gene signature for diagnosing paediatric TB. The 6 candidate biomarkers were identified in an RNA-Seq dataset, cross-platform validated by RT-qPCR and finally validated in independent cohorts by RT-qPCR. This is the first time a gene signature has been discovered in a large multi-country paediatric RNA-Seq dataset and validated in independent cohorts by RT-qPCR. The 6 genes that I have identified can accurately differentiate between microbiologically confirmed TB and unlikely TB in children and could be developed into a rapid POCT. In Chapter 3, I have also described the results from a pathway enrichment analysis carried out on the Significantly Differentially Expressed (SDE) genes between the two disease groups of interest. These biological insights could be used by other researchers to guide future studies.

In Chapter 4, I describe how the gene expression profile of children with microbiologically confirmed TB responds to anti-TB treatment. Other than as potential diagnostic biomarkers, genes have also been used to explore the response of adults to anti-TB treatment. It has been shown that marker genes generally normalise on completion of anti-TB treatment in adults [206]. Such changes have not previously been documented in children on anti-TB treatment. Chapter 4 fills this gap and, for the first time, assesses how the performance of the gene signature described and identified in Chapter 3 changes after 3 months of treatment in children with TB disease and unlikely TB.

Chapter 5 describes the discovery and subsequent validation of a new 4-protein signature for paediatric TB. The 4-protein signature has a good diagnostic performance in distinguishing between microbiologically confirmed TB vs. unlikely TB and could potentially be developed into a rapid POCT. To my knowledge, there are no similar studies with comparable results, especially in paediatric cohorts.

In chapter 6, I describe a novel and still much-unexplored field of multi-omics integration for biomarker discovery. A 5-feature signature, composed of 2 genes and 3 proteins, has an optimal performance in diagnosing TB disease in children and outperforms the single-omics transcriptomic and proteomic signatures.

In this thesis, I have provided considerable contributions to the ongoing effort to improve the diagnosis of TB in children. I have identified candidate biomarkers for childhood TB, firstly analysing the transcriptomic and proteomic datasets as a single layer of -omics, and then exploring a completely new path by integrating the two single-omics datasets and performing a multi-omics analysis to identify multi-omics biomarkers.

CHAPTER 2

MATERIALS AND METHODS

2.1 Overview and chapter aims

Pre-processing and normalisation of biological data are key components in the analysis workflow of raw data to ensure accurate interpretation of the results. An -omics dataset typically contains more than one sample and more than one variable (i.e., gene or protein). Normalisation is needed for the elimination of noise and batch effect prior to statistical analysis.

This chapter starts with the description of the online platform created to merge, store and electronically access the clinical metadata used in my thesis in Section 2.2, followed by Sections 2.3 and 2.4 which describe the case definitions and the clinical features of the samples used to generate the biological data respectively.

Sections 2.5, 2.6 and 2.7 outline the computational and statistical methodologies for transcriptomic and proteomic data analysis used in this thesis, pre-processing of raw data, Quality Control (QC) and normalisation, as well as analytical approaches. Methods used for multi-omics data integration are described in Section 2.8.

Five types of -omics datasets were analysed as part of this thesis and described in Chapter 2: Reverse Transcription - quantitative Polymerase Chain Reaction (RT-qPCR), RNA Sequencing (RNA-Seq), Enzyme-Linked Immunosorbent Assay (ELISA), Luminex® and SomaScan® assays.

2.2 Clinical REDCap database

Research Electronic Data CAPture (REDCap) is a web-based application developed for collecting, storing and disseminating data for clinical research in online databases. It was created in 2004 at Vanderbilt University (Nashville, Tennessee, USA) with the aim of making a user-friendly interface available to researchers with no background knowledge or technical experience in data management [207]. REDCap has been designed to provide a secure environment for the storage of highly sensitive information, it is in fact compliant with many data access regulations worldwide, and hence highly secure. Being a metadata-driven application, REDCap can be seen as a server containing individual online databases, referred to as projects, which share the same design and rendering engine, with the metadata being the only difference.

The key features of REDCap [207] make each individual project user-friendly, highly secure and accessible simultaneously by many institutions and departments worldwide. Researchers' freedom and autonomy in creating new REDCap projects with no or limited support by people with experience in data management were probably the main reasons for its success worldwide.

Some limitations should however be recognised. The research data are stored in a flat-table structure, which makes it inefficient from a storage standpoint compared to a relational database. A relational

database, which is the most widely used database structure, is based on a relational model where several tables have pre-defined relationships to avoid information redundancy, which is instead very often present in REDCap projects. Moreover, the REDCap metadata creation process, which is researcher-led, does not encourage standards-based data identifiers and, hence, does not follow the traditional database naming convention.

Besides its limitations, REDCap was chosen as the web tool to store the clinical metadata needed to analyse the data described in this thesis for its user-friendly and intuitive interface. REDCap, in fact, merges the presentation layer (i.e., front-end) and the data access layer (i.e., back-end) in a single platform. Back-end and front-end are usually different but integrated parts of an online application. The core of the front-end is to ensure that users with no data management experience can easily interact with the data. The back-end handles the actual data flow from the data source to the front-end layer, with data QC and transformation to have them in a format easily readable by the front-end application. Whilst generally these two layers are kept separated and are managed by two different software or tools, REDCap is able to merge back-end and front-end because of its flat-table structure.

Clinical data for 4,958 biological samples were uploaded and stored in a REDCap project. The samples were classified into 4 disease groups following the case definitions described in Section 2.3. Some of their clinical features are described in Section 2.4.

2.3 Case definitions

After clinical evaluation, children presenting to healthcare facilities for suspected TB were assigned to groups according to whether the diagnosis was microbiologically confirmed TB, unconfirmed TB, *Mtb* infection or diseases other than TB, which are referred to as unlikely TB in this thesis.

The children were classified based on the phenotyping algorithm described by Graham et al (2015) [208], which included only the definition of PTB. It was subsequently expanded also to include EPTB diagnosis [Appendix A]. Figure 2.1 describes the case definitions used to classify the biological samples described in this thesis.

A child, to be classified as microbiologically confirmed TB, needed to have a bacteriological confirmation, either by culture or NAATs from at least 1 specimen obtained from the site of TB (i.e., respiratory specimen in case of PTB or body fluid or tissue specimen from the EPTB site). A child was diagnosed with unconfirmed TB when the bacteriological confirmation was not obtained and at least 2 of the following were met:

- Symptoms/signs suggestive of TB
- Close TB exposure or immunological evidence of *Mtb* infection
- Positive response to TB treatment
- Chest radiography consistent with PTB OR other evidence/criteria at the site of EPTB described in Appendix A

When the bacteriological confirmation was not obtained and the criteria for “unconfirmed TB” were not met, a symptomatic child was classified as unlikely TB. A healthy child with no clinical features of TB disease and a positive IGRA or TST result was classified as having *Mtb* infection.

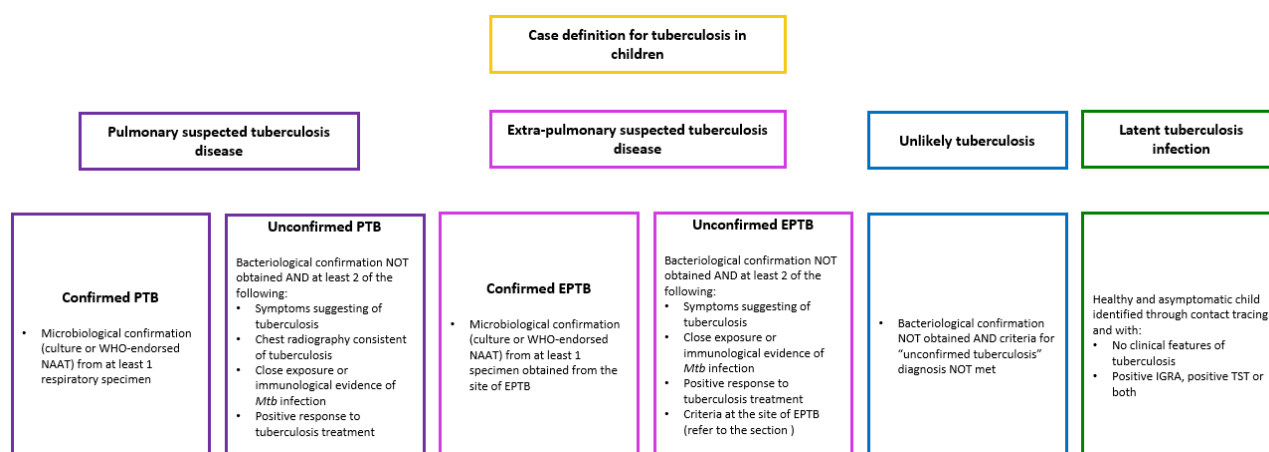


Figure 2.1: Diagnostic algorithm for case definitions.

Bacteriological confirmation from at least one specimen from the site of TB was needed to classify a patient into the microbiologically confirmed TB group. If the bacteriological confirmation was not obtained but the patient had other evidence of TB disease, the child was diagnosed with unconfirmed TB. When the criteria classification for microbiologically confirmed or unconfirmed TB were not met, but the child was symptomatic, then the classification was unlikely TB. A healthy child with positive IGRA or TST was classified as having *Mtb* infection. *Mtb* = *Mycobacterium tuberculosis*; IGRA: Interferon-Gamma Release Assay; TST: Tuberculin Skin Test.

2.4 Description of the clinical cohorts

The paediatric samples used to generate the data analysed and presented in this thesis were collected, after written informed consent was obtained from parents or guardians of the children, in four Sub-Saharan countries (South Africa, Malawi, Kenya and The Gambia) as shown in Figure 2.2.

Whole blood was collected in either PAXgene Blood RNA Tubes (PreAnalytiX) or serum-separating tubes at recruitment. PAXgene tubes were frozen within 6 hours after collection, stored at -80°C and RNA was extracted using PAXgene Blood RNA Kits. Serum was processed and stored at -80°C .

South African and Malawian children were recruited from February 2008 to January 2011 into the European Union (EU)-funded ILULU study, which aimed to identify transcriptomic and proteomic signatures to differentiate TB disease from unlikely TB in both children (<15 years old) and adults. The study was approved by the research ethics committees of the University of Cape Town, South Africa; the University of Malawi, College of Medicine; the Liverpool School of Tropical Medicine; and Imperial College London. Patients with suspected TB or close TB contact were referred to the study team for investigation by the clinician. Recruited patients were classified into 4 diagnostic groups as described in Section 2.3.

The recruitment of the Kenyan cohort was part of the TB KIDS study funded by the Wellcome Trust, which aims were to investigate the performance of clinical and laboratory tools for childhood TB diagnosis in Kenya. The study was approved by the research ethics committee of the Kenya Medical Research Institute. Whole blood samples were collected at enrollment from August 2009 to December 2011. Children (<15 years old) with suspected TB disease or who had been in close contact with a smear-positive PTB case were referred for outpatient clinical investigation, followed by further investigations and case management as described in Section 2.3. A more detailed description of the three cohorts has been published in the Anderson et al paper [137] in 2014.

The Gambian cohort was recruited from February 2012 to June 2017, and described in previously

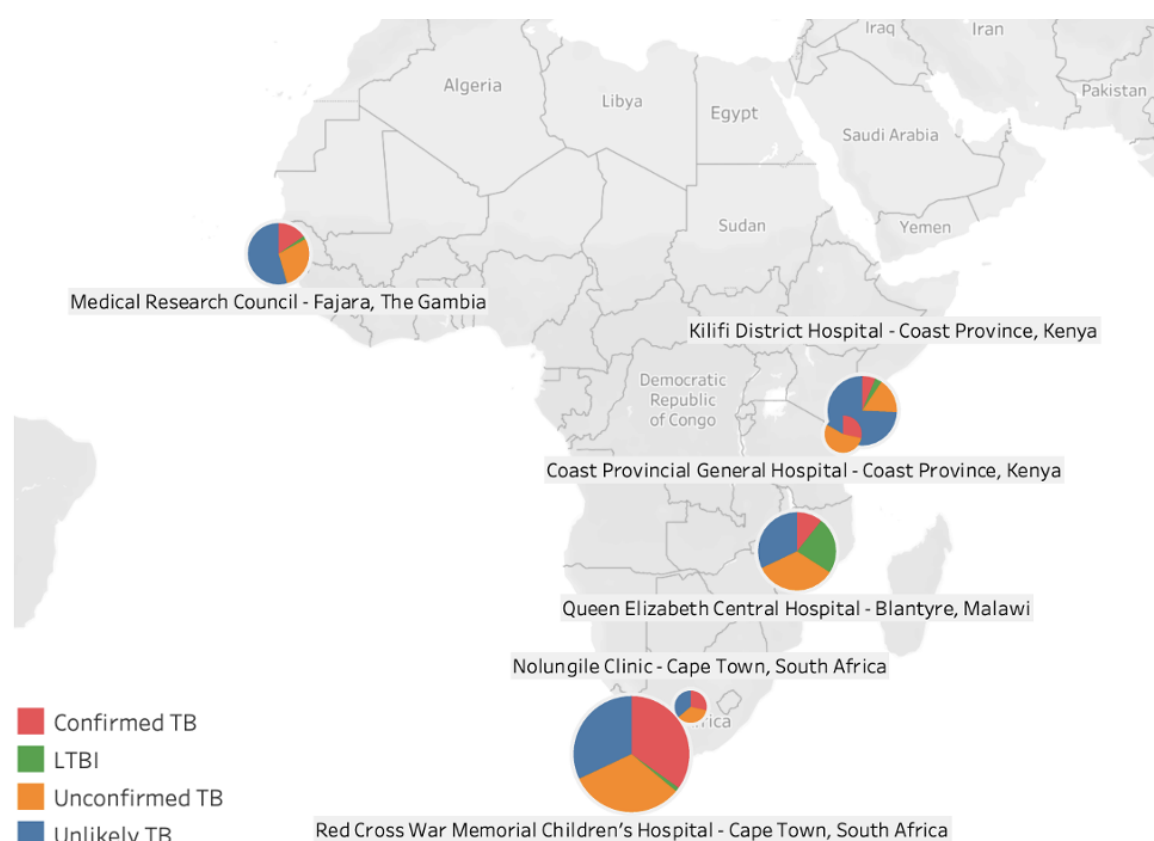


Figure 2.2: Maps of the study population.

Each pie chart represents a recruitment centre. The size of the pie chart is proportional to the number of children recruited in that specific recruitment centre. Each colour in the pie chart (i.e., slice) represents a disease group. *LTBI*: Latent Tuberculosis Infection. Created with Tableau Desktop 2022.3.

published works [186, 209]. Briefly, this cohort was recruited by The Childhood TB group of the Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine. Ethical approval was obtained from the Gambia Government/Medical Research Council (MRC) joint ethics committee. The aim of the study was to identify a diagnostic signature to distinguish TB disease from unlikely TB in children. The recruited children (<15 years old) were HIV-uninfected, symptomatic with suspected PTB disease and had documented household exposure to an adult with smear-positive TB. Further evaluations, as described in Section 2.3, were carried out at a dedicated childhood TB clinic to classify children into the 4 diagnostic groups.

A NIH-funded study set up at Red Cross War Memorial Children's Hospital in Cape Town (South Africa) recruited a cohort of children (<15 years old) with suspected TB disease from February 2009 to March 2017. The Research Ethics Committee of the University of Cape Town approved the study. The recruited children were followed up and samples (RNA, nasopharyngeal swabs, stool and urine) were taken at enrollment, 3 and 6 months to assess the response to TB therapy or recovery in the absence of TB therapy.

Clinically, the four African countries present some differences. At the time of recruitment, South Africa, Malawi and Kenya had a very high paediatric HIV epidemic [91], and in the last two malnutrition was endemic [Figure 2.3]. As described in Section 1.7, malnutrition causes immune dysfunction and may be associated with worse outcomes for a child with TB disease [210], hence it is an important clinical variable to take into consideration in TB studies. Malnutrition, which is usually used as a synonym for

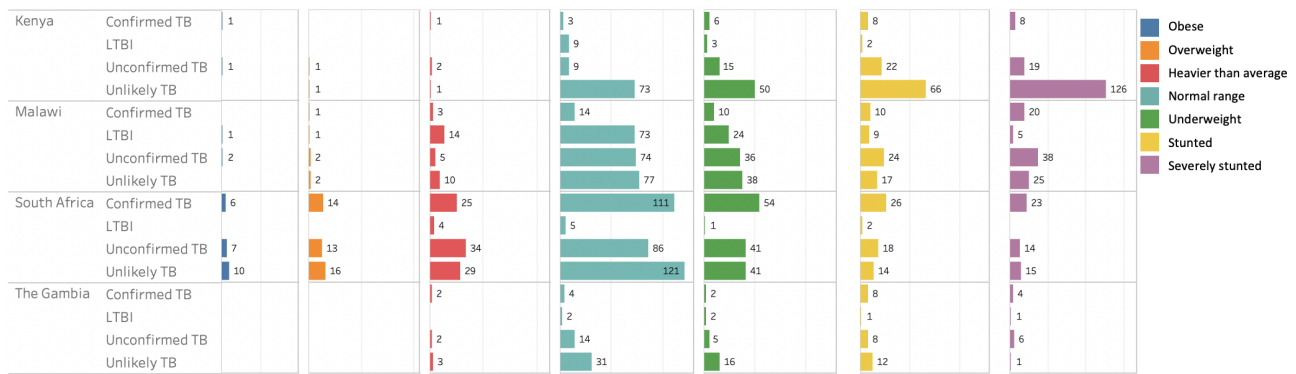


Figure 2.3: Z-score of the children in the study population broken down by country.

From left to right, the horizontal bars represent the number of children in the groups derived from the z-scores. The bars are broken down by country and disease group. *Created with Tableau Desktop 2022.3.*

under-nutrition or deficiency in essential nutrients [211], was categorised based on z-scores calculated as weight-for-length/height. For this purpose, the WHO published some charts [212] for calculating the z-score based not just on the weight and length/height of the child but also on sex and age. In our cohorts, z-scores grouped the children into seven groups (i.e., obese, overweight, heavier than average, normal range, underweight, stunted and severely stunted). The number of malnourished children, as previously mentioned, resulted to be higher in Kenya and Malawi compared to South Africa and The Gambia.

HIV co-infection is another risk factor for worse outcomes and mortality. The Gambia has a very low HIV incidence compared to the other countries in the study and, while the children recruited in South Africa, Malawi and Kenya included HIV-infected and HIV-uninfected patients, all the samples collected in the Gambian cohort were only HIV-uninfected [Figure 2.4]. The highest percentage of HIV-infected children was in the Malawian population. Around the time of recruitment, there was a high HIV incidence in this country with 11% of the population living with HIV [213]. Due to the increased availability of antiretrovirals, the number of HIV cases in Malawi, and more generally in LMIC, is currently decreasing and the survival rate of people living with HIV is increasing [214]. However, as highlighted by the WHO, it is critical to include both HIV-infected and -uninfected children in biomarker discovery studies to identify highly accurate biomarkers in the two clinical subgroups since HIV infection increases the incident of TB in children with a more aggressive and deadly disease [92].

The different cohorts included patients with microbiologically confirmed TB (both PTB and EPTB), unconfirmed TB (both PTB and EPTB), *Mtb* infection and unlikely TB. For the aim of this thesis, the most relevant disease groups were microbiologically confirmed TB and unlikely TB, which were used to identify the best biomarkers to diagnose TB disease in children. Children with unconfirmed TB, in fact, fulfil clinical and radiological criteria for TB disease but without microbiological confirmation. This group is most likely a mixture of microbiologically confirmed TB and unlikely TB. Hence it was excluded from biomarker identification in this thesis. Follow-up visits to assess whether an unconfirmed TB case had microbiological confirmation of the disease were not carried out, leaving the diagnosis of the children in the unconfirmed TB group uncertain.

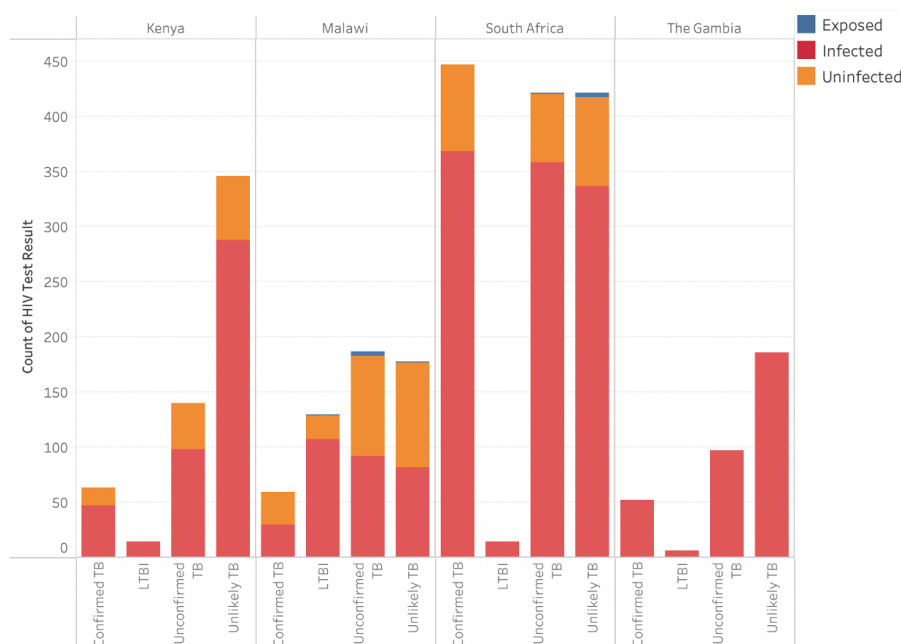


Figure 2.4: HIV status (exposed, infected or uninfected) in the study population broken down by country and disease group.

From left to right, the vertical bars represent the number of children in each disease group and country. The single bars are coloured according to the HIV status. *Created with Tableau Desktop 2022.3.*

2.5 Data analysis of biological data

As each -omics dataset was generated differently and from different types of patient samples, a tailored-made data analysis pipeline was used to ensure the robustness of the results. Whilst some steps in the various pipelines are indeed unique to specific -omics datasets, some others are in common. All the data analysis pipelines described in this chapter have as input the raw -omics dataset and as output a list of statistically significant features (i.e., genes or proteins) from which a feature selection algorithm identifies the most promising biomarkers.

A general data analysis pipeline for the identification of biomarkers is described in Figure 2.5 and it is made up of:

1. Pre-processing - The raw data are transformed into an analysis-ready format along with QC steps to ensure the quality of the data
2. Normalisation - Biases or errors, systematic or otherwise, are corrected to make samples more comparable
3. Statistically different features identification - Statistical models are used to compare samples and identify the features that are statistically different between the groups of interest
4. Feature selection - The most discriminant features between the groups of interest are selected and combined into a predictive model

The first key step for the analysis of -omics datasets is to transform the raw data into a matrix format. The complexity of this step varies between -omics approaches, with some platforms returning output data in an analysis-ready format, and other platforms requiring computationally demanding steps and specialist knowledge. In general, targeted -omics techniques, such as RT-qPCR, SomaScan®, Luminex®

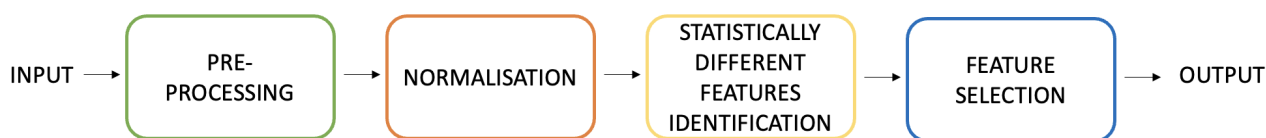


Figure 2.5: Data analysis pipeline for biological data.

Raw biological data, regardless of the -omics technique used to generate them, need to go through some common steps in the data analysis pipeline. Whilst the processing from raw data to an analysis-ready format and the normalisation are usually unique for each -omics technique, some steps of the pre-processing, the identification of statistically significant features and the feature selection steps are usually performed in a similar manner.

and ELISA, require more straightforward steps to process the raw data since the target features do not need to be identified. Contrastingly, untargeted techniques, such as RNA-Seq, require extra steps in the pipeline to first identify the quantified features.

Out of the five types of -omics datasets presented in this thesis, RNA-Seq [Section 1.10.3] is the only one not depending on known target molecules. It quantifies all RNAs in the biological samples and needs extra steps to identify the target molecules and have the data in an analysis-ready format. Somascan® [Section 1.11.3] is a targeted technique based on aptamers which bind the molecules of interest. For this technology, image processing software transforms the images received from the scanners into raw data files by numerically encoding the measured intensity for each aptamer. RT-qPCR [Section 1.10.1], ELISA [Section 1.11.1] and Luminex® [Section 1.11.2] are targeted techniques which measure a specific set of targets. Whilst RT-qPCR quantifies the levels of genes of interest, ELISA and Luminex® measure the abundance of proteins of interest. They all measure the fluorescence of the targets, which is proportional to the expression of the genes or the abundance of the proteins of interest. Whilst commercial software tools can be used for downstream analyses of the raw data, open-source software packages for the R Statistical Computing Language [215] were used to read and analyse the files or data returned from the -omics technologies in this thesis.

2.5.1 RT-qPCR data generation and pre-processing

RT-qPCR is a highly sensitive targeted transcriptomic technique aimed at measuring the expression of sparse sets of genes. It generally includes technical replicates and control samples.

Technical replicates help protect the data - if one amplification fails, other wells may generate a good signal. The Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines suggest carrying out an RT-qPCR experiment in duplicates or triplicates [216]. The MIQE guidelines are the minimum information required for the publication of RT-qPCR experiments with the aim of ensuring the integrity of the scientific literature, promoting consistency between laboratories and increasing experimental transparency. In the RT-qPCR experiments performed to generate the data presented in this thesis, triplicates were preferred over duplicates since, if one of the three wells failed, the other two could be kept in the downstream analysis. In the case of duplicate samples, if one of the two wells fails, both of them should be excluded since there is uncertainty on which of the two wells has generated the correct signal. The output signal of technical replicates was averaged.

RT-qPCR controls were used to validate the results and to ensure that the reactions occurred correctly. Negative amplification controls do not have any DNA templates and therefore no amplification is expected. The negative controls were used to ensure that none of the reagents had been contaminated.

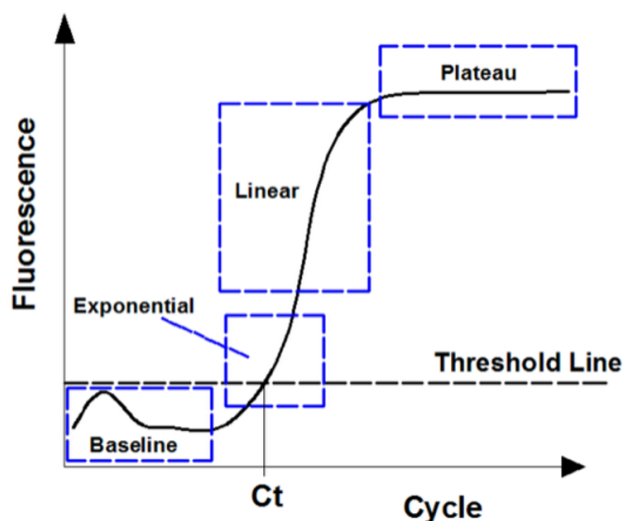


Figure 2.6: Theoretical plot of a PCR output curve [218].

A general PCR output curve plots the PCR cycle number against the fluorescence emitted. For a positive reaction, the amplification curve can be divided into four phases: baseline, exponential, linear and plateau phases. The C_t value is the number of cycles needed for the fluorescence to reach the user-selected threshold. PCR: Polymerase Chain Reaction; C_t : Cycle Threshold.

On the other hand, positive controls contain the DNA template and therefore show a positive amplification reaction. Adding positive controls in the reaction is important to determine the sensitivity of the experiment. Internal controls were also used in RT-qPCR experiments for the correct interpretation of the results. According to the MIQE guidelines [216], these controls are referred to as reference genes. Reference genes are cellular maintenance genes which regulate basic cellular functions [217] and have stable expression across the different groups of interest. These genes are used to normalise the raw data. The MIQE guidelines [216] recommend a minimum of three reference genes for normalisation.

C_t values are the numerical outputs of an RT-qPCR experiment. PCR can be divided into four phases: baseline, exponential, linear and plateau phases [Figure 2.6]. In the baseline phase, the reaction has not taken place yet, therefore the fluorescence remains at the background level. In the exponential phase, the signal increases exponentially since the reagents are still available and not limited. During the linear phase, the amplification continues but at a slower pace due to the PCR reagents becoming limited until they become depleted and the fluorescence signal remains unchanged in the plateau phase. The C_t value is the number of cycles needed for the fluorescence to reach the user-selected threshold or the background noise, and it was the value exported and subsequently analysed for each gene and sample.

2.5.2 RNA-Sequencing data generation and pre-processing

High-throughput or next-generation sequencing technology has revolutionised genomic research, overcoming targeted techniques such as microarray and RT-qPCR. As described in Section 1.10.3, the current RNA-Seq technologies sequence the whole transcriptome rapidly, at a low error rate and affordable price. RNA-Seq is becoming more and more popular for its ability to sequence transcripts in an untargeted manner, in contrast to microarray and RT-qPCR that need probes to bind the transcripts of interest.

Processing and analysis of next-generation sequencing data are more demanding in terms of computation

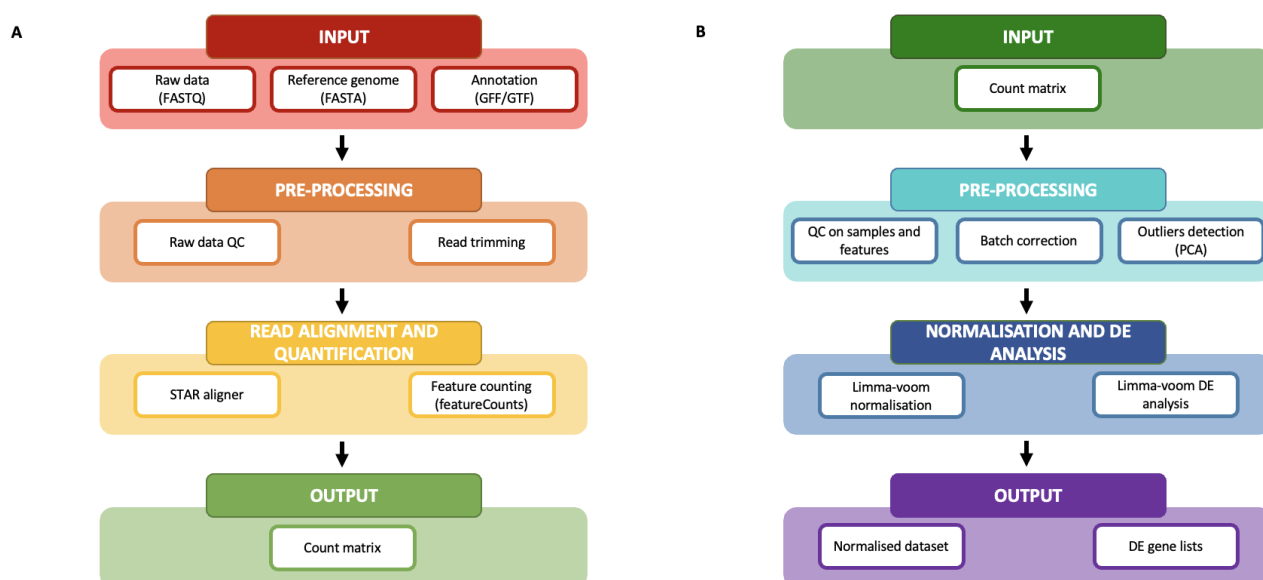


Figure 2.7: RNA-Seq data analysis pipeline.

Panel A describes the steps of the pre-processing, QC and alignment performed on a HPC resources. This transforms raw data into a count matrix, which is easier to work with. The output count matrix is then usually pre-processed, normalised and analysed locally (Panel B). *RNA-Seq* = *RNA Sequencing*; *QC* = *Quality Control*; *HPC* = *High-Performance Computing*; *PCA* = *Principal Component Analysis*; *DE* = *Differential Expression*

in comparison to previous quantitation techniques. Next-generation sequencing data are generally processed through multiple steps and on High-Performance Computing (HPC) resources due to their computational cost. Furthermore, they require strong programming knowledge. One of the main advantages of RNA-Seq is its ability to assemble *de novo* transcripts. In fact, it quantifies all the RNA transcripts present in the biological sample at the time of sampling not depending on predefined target probes. However, the pre-processing of the data is less straightforward since the target features need to be first identified. RNA-Seq raw data, which are the output of the sequencing machine, are sequencing reads stored in FASTQ files. Reads, which are sequences of base pairs corresponding to DNA fragments, are then aligned to a reference genome, usually using available aligners such as STAR [219], HISAT2 [220] or Subread [221], and counted into annotated genes. These steps are needed to transform the raw data (i.e., sequences of nucleotides) into analysis-ready format data (i.e., the count matrix). The count matrix is easier to work with since the nucleotide sequences are transformed into gene counts and moreover, the computational cost is reduced from several terabytes needed for storing FASTQ files to some megabytes for storing a count matrix. Figure 2.7 describes the data analysis pipeline used to pre-process, QC and normalise the RNA-Seq data, of which results are presented in this thesis.

Pre-processing and QC of RNA-Seq raw data Each step of the pre-processing and QC of raw RNA-Seq data requires a different tool and, hence, computational experience. With the huge success of sequencing techniques, several tools to pre-process and analyse sequencing data have been developed and made available. Below I describe the ones used to process the RNA-Seq data in this thesis.

The raw reads were stored in compressed FASTQ files. FASTQ format is a text-based format used for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Quality scores are numerical values assigned to every base by the sequencer and describe the probability of sequencing errors. Due to the technical limitations of each sequencing platform, errors

such as incorrect nucleotides being called can be introduced. The higher the quality score, the lower the probability of a sequencing error. Sequence quality control is, therefore, an essential first step in the analysis pipeline. FastQC [222], which is one of the several tools available to QC RNA-Seq raw data, was used in this thesis. FastQC generated a report with graphs and plots for each sample in the experiment. Since many samples underwent sequencing in the same experiment, the output of FastQC was collated by MultiQC [223], a tool to summarise QC information for multiple samples into a single report, as well as provide tab-delimited files containing all the FastQC statistics.

The second step of the pre-processing pipeline was the adaptor and quality trimmings, along with the removal of very short reads. This step, which is needed to clean up the data and reduce unwanted noise, consisted of three sub-steps:

1. Trimming of the sequencing adaptors - Adaptors, which are short and chemically synthesised DNA sequences used for the library preparation [Section 1.10.3], usually get sequenced and 90% to 20% of their reads get removed before saving the FASTQ files. The adaptor trimming was needed to remove the remaining adaptor reads
2. Trimming of the low-quality bases at the ends - The quality of the sequenced bases tends to drop at the ends of the read, hence the probability of a sequencing error and the base being sequenced incorrectly is higher. Bases at the ends with low-quality scores needed to be removed
3. Discarding any sequences shorter than 20 base pairs (bp) - Short sequences can align to multiple and incorrect locations on the reference genome introducing noise, and therefore needed to be removed

FastQC and MultiQC were used again to assess whether the quality of the data had improved after the adaptor and quality trimmings.

Spliced Transcripts Alignment to a Reference (STAR) [219] was subsequently used to align the cleaned-up data to the reference genome (version 89 Ensembl GCh38 human genome). Along with the pre-processed FASTQ files and a FASTA file with the reference genome, a GFF/GTF file with the annotations was needed to align the data. Whilst the reference genome is a description of the whole genome sequence, the annotation file describes where the genetic elements (i.e., introns, exons, transcripts, genes) are located in the genome. In contrast to some other tools, STAR was designed to align the non-contiguous RNA-Seq reads directly to the reference genome in order to mimic the process of splicing together non-contiguous exons to create mature transcripts. Along with annotated and novel genes, exons and splice junctions, STAR is capable of discovering more complex RNA sequence arrangements, such as chimeric and circular RNA [224]. The algorithm is based on a two-step process:

1. Seed searching - STAR starts with searching for the longest sequence that exactly matches one or more locations on the reference genome. This will be called *seed1*. This process is repeated several times labelling *seed2*, *seed3* etc the subsequential longest sequences matching the reference genome. Sequential searching of only the unmapped portions of reads is the core step that makes STAR faster than other aligners based on algorithms that often search for the entire read sequence before splitting reads and performing iterative rounds of mapping. If STAR does not find an exact matching sequence for each part of the read due to mismatches or indels, the previous *seed* will be extended
2. Clustering, stitching, and scoring - The seeds are then stuck together to create a complete read.

This is done by clustering the seeds together based on proximity to a set of not multi-mapping seeds (anchor seeds), and based on scores and penalties for mismatches, insertions, deletions and junction gaps. The combination of seeds with the highest score will be the best and final alignment of the read [219]

Once the alignment was complete, the final results were saved in Sequence Alignment Map (SAM) or BAM (compressed binary version of a SAM files) files. When 70% - 90% of the reads match somewhere on the genome, the quality of the data can be defined as satisfactory. Along with other advantages already mentioned in this section, speed and accuracy were among the main reasons why STAR was used to align the raw RNA-Seq data for this thesis.

featureCounts [225] was then used to count the number of aligned reads associated with each gene in each sample. The output of pre-processing and QC of the raw RNA-Seq data was a $p \times n$ count matrix, with p genes and n samples.

2.5.3 ELISA data generation and pre-processing

ELISA is a single-plex antibody-based proteomic platform which measures one analyte of interest in biological samples [Section 1.11.1]. The ELISA reader quantifies the fluorescence emitted by the reporter enzyme, which is used to calculate the concentration values and standard curves.

The generation of the ELISA standard curves is very similar to the process used in the Luminex® assay. It is therefore described in the following paragraph.

2.5.4 Luminex® data generation and pre-processing

Luminex® is a multi-plex antibody-based proteomic platform which can measure several proteins of interest in the same sample plate [Section 1.11.2]. Even if, theoretically, the maximum number of cytokines detected in a single Luminex® plate can be up to 100 [226], most of the Luminex® kits detect up to 40 cytokines because of limitations due to interference issues between the target analytes, as well as their different dynamic ranges. The number of analytes detected is often extended by using multiple assay types (i.e., increasing the number of analytes measured for each sample) and different panels against the same samples (i.e., increasing the sample size for each analyte).

Luminex® assays are targeted antibody-based techniques where a reader quantifies the fluorescence emitted by the capture labelled antibodies, and the software calculates the standard curves and concentration values.

Biological samples, standards, blanks and controls are usually assayed in a 96-well plate format, with 8 rows and 12 columns. Blanks are negative controls with no analytes and are used to establish the background signal/readout. The fluorescence of blank controls is subtracted from the fluorescence of the other samples before fitting the standard curve to control for the background noise. Standards, which are serial dilutions of a reference sample with a known concentration of each analyte, are used to plot a standard curve [Figure 2.8]. For each analyte, the standards with the lowest and highest concentration set the dynamic range, the extreme points of the standard curve and the lowest and highest limits of detection. If the fluorescence intensity for a specific analyte in a specific sample is too low (i.e., the reader is unable to detect the signal) or too intense (i.e., the reader gets saturated), the reader fails to quantify the fluorescence intensity and this point falls outside the dynamic range on the

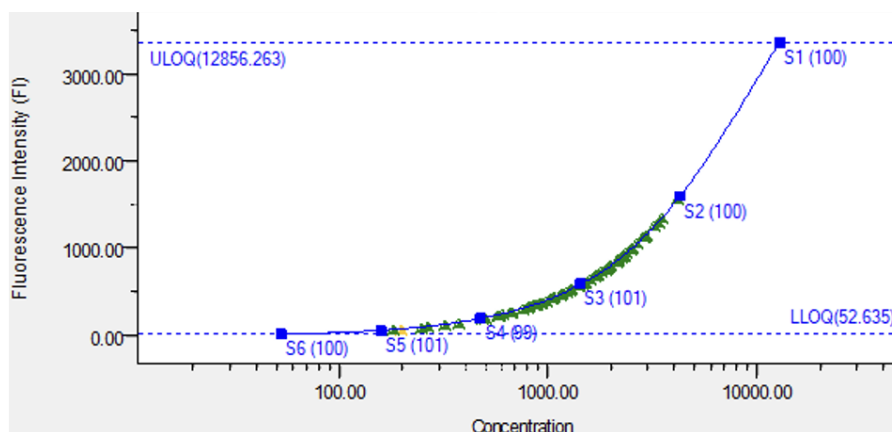


Figure 2.8: Theoretical plot of a Luminex® output curve.

The standard curve represents the fluorescent intensity on a linear scale plotted against the concentration on a logarithmic scale for each analyte. The lowest and highest concentrations of cytokine standards are the extreme points of the standard curve and define the dynamic range of the analyte (i.e., lowest and highest limit of detection). The output signal detected by a Luminex® reader is the Fluorescence Intensity (FI) over all beads containing the corresponding analyte that is captured by the reader. The fluorescence, which is proportional to the abundance of the target protein, is then transformed into a concentration through a sigmoid 4- or 5-parameter logistic regression model.

standard curve.

The Luminex® output data were exported as “.CSV” files for each plate. Each file had the values of fluorescence intensity and concentration for each analyte in each sample, along with the standard curves for each analyte in each plate and other values needed to assess the quality and reliability of the data.

2.5.5 SomaScan® data generation and pre-processing

SomaScan® is an aptamer-based technique which measures the abundance of 1,301 human proteins in biological samples simultaneously. The output signal is measured by custom DNA microarrays. The fluorescence of the labelled SOMAmer® is detected by a microarray reader and quantified into a numerical matrix with values proportional to the fluorescence detected and, therefore, to the abundance of the target proteins in the assay.

2.6 Quality control and normalisation

Once -omics datasets have been transformed from raw data to expression/abundance matrices, QC and normalisation steps are performed to identify and account for biases and errors introduced through technical factors [227]. The pre-processing steps usually vary from -omics to -omics since each of them is generated from different types of biological samples and different machines. Thus, the raw data are transformed into an analysis-ready format following different pipelines. Some techniques, such as RNA-Seq, need some extra pre-processing steps to convert the data into a format suitable for analysis. On the other hand, RT-qPCR, SomaScan®, ELISA and Luminex® generate the output data in a ready-to-analyse format easing the data analysis pipeline. A QC step is, however, always needed to detect outliers or any form of variation that needs to be accounted for in the analysis. This is usually assessed with the use of Principal Component Analysis (PCA). PCA is a dimensionality reduction method for analysing datasets with a high number of features per each observation (i.e., a large number of

genes or proteins for each sample) that enables the visualisation of multidimensional data [228]. PCA increases the interpretability of the data while preserving the maximum amount of information. Being p the original number of variables, PCA computes up to $p-1$ new variables that are linear functions of those in the original dataset. Each new linear combination is defined as Principal Component (PC). The first PC is a linear combination of all the p variables in the dataset and captures the most variance in the data. It is calculated as follows:

$$PC_1 = \sum_{j=1}^p a_{1j} X_j \quad (2.1)$$

where p is the number of variables, X_j are the p variables in the original dataset and a_{1j} represents the loadings for each variable across all samples. Loadings are important features of a specific PC_n since they describe the contribution of each variable (i.e., gene or protein) to the variance captured by PC_n . The subsequent $n-1$ PCs are calculated in the same way. Each PC is orthogonal to the previous one and captures the most variance in the data. PCA is a very powerful tool to visually explore and QC the data. It can reveal outliers and batch effects caused by known variables or unexpected clustering patterns, which have to be excluded and minimised respectively.

The pre-processing step is followed by normalisation, which is a technique depending on the data distribution used to eliminate unwanted bias. Normalisation makes samples more comparable and ensures that statistical methods measure differences in the expression of genes or abundance of proteins that are due to biological changes rather than biases or errors. Variations introduced by technical or experimental factors, such as batch effect, cannot be removed by normalisation and need to be minimised previously. Batch effects are due to variability introduced in the data by non-biological factors. They are a common problem in -omics data analysis since datasets are often generated in multiple batches due to logistical and practical limitations, including the size of assays and the number of samples that can be handled in a day [229, 230]. Batch effects can also be introduced by variations in experimental conditions, such as reagents used or the individual carrying out the experiments [230]. As these issues are quite common in -omics data generation and analysis, various methods are available for identifying and removing batch effects.

ComBat [231, 232] is a popular batch effect correction method that was designed for correcting batch effects in microarray datasets, but it has been successfully applied to various data modalities since its development [233, 234]. ComBat uses empirical Bayes (EB), which is a class of methods for statistical inference where the prior distribution is estimated from the data rather than fixed before the data is observed (as in standard Bayesian methods). EB methods are well suited to -omics due to their ability to handle high-dimensional datasets with small numbers of samples [231]. EB methods estimate the prior probability distribution from the data, resulting in more stable and reliable inferences.

ComBat-seq is a modified version of ComBat that is optimised for RNA-Seq count data [235]. ComBat assumes bell-shaped Gaussian distributions of the data, which is appropriate for microarrays but not for RNA-Seq count data. RNA-Seq data are typically skewed, over-dispersed counts and better fitted to a negative binomial distribution. ComBat-seq obtains estimated batch effects directly from a negative binomial regression model for each gene. Whilst the original ComBat batch effects are removed by subtracting the mean batch effect estimate and scaling by variance estimates, ComBat-seq maps the empirical distribution of count data to an expected distribution that is “batch-free” (i.e., if there were no batch effects in the data) retaining the integer counts that are necessary for downstream RNA-Seq

analysis tools [235]. Downstream tools for RNA-Seq data analysis expect raw count matrices, which are in integer format. ComBat-seq ensures that integers are returned following batch effect correction, unlike the original ComBat method [235]. It is, therefore, used before normalisation of the raw count matrix.

The `removeBatchEffect` function [236] is part of the `limma` R package and is another tool to minimise the systematic variation due to batch effect or other sources of technical variability in RNA-Seq data. This function fits a linear model for each gene using a blocking term for the batch. The coefficient for each blocking term is then set to zero and new expression values are calculated from the remaining residuals and terms. The new calculated expression matrix is without batch effects. `removeBatchEffect` requires logged expression values as input, thus it needs to be used after Differential Expression (DE) analysis, normalisation and log-transformation of the raw gene count matrix. To ensure that the DE analysis is not affected by the batch effect, this is usually added as an additive term in the statistical model. In this thesis, the batch effect for transcriptomic data has been accounted for in the DE analysis by adding the batch as a term in the additive statistical model and then minimised using `removeBatchEffect` before the feature selection step. This was done because existing tools for removing batch effect before normalisation and DE analysis might induce apparent batch differences when disease groups are not evenly distributed across the batches, and therefore lead to biases in differences between the groups of interest [237].

`normn.MA` is part of the `MDimNormn` R package [238] and was developed to minimise the technical variation due to multiple plates run in a single experiment in bead and antibody-based assays. It is based on the *MA* coordinates, with *A* being the average fluorescence intensity or concentration levels across all sample plates and *M* describing the differences between plates. Being *n* the number of sample plates, the *MA* coordinates are defined by a rotation in the *n*-dimensional Euclidean coordinate system. The i^{th} coordinate represents the mean or geometric mean of the fluorescence intensity or concentration levels in plate i^{th} . The *A* axis of the new *MA* coordinate system is chosen as the axis that captures the most variance. The *M* axes ($M_1, M_2, \dots, M_{n-1}$) are perpendicular to each other and to the *A* axis, capturing the deviation from the *A* axis. Together, the *A* and *M* axes capture the differences in the fluorescence intensity or concentration values between the sample plates. The plate effect for each analyte/protein in each plate is the distance between the *A* axis and the fluorescence intensity or concentration value in the new *MA* coordinate system. This is minimised by shifting the individual fluorescence intensity or concentration values by the distance to the *A* axis. This algorithm can be used only if the distribution of the measures of one single analyte/protein is invariant across plates. In fact, if the mean of the measures for each analyte is constant across plates, the divergence of the mean values can be identified as the technical variation due to the plate effect and minimised. This is the main reason why the randomisation of samples across plates is critical before plating and carrying out the experiment.

The specific QC and normalisation processes applied to RT-qPCR, RNA-Seq, Somascan®, ELISA and Luminex® are described in Sections 2.6.1, 2.6.2, 2.6.3, 2.6.4 respectively.

2.6.1 RT-qPCR data

C_t values are used to QC and normalise the data, to calculate the Fold Change (FC) and Benjamini-Hochberg (BH)-adjusted *p*-value for each gene. The most widely used method to calculate the relative expression between two comparative groups (i.e., the FC) in an RT-qPCR experiment is the $\Delta\Delta C_t$ model [239].

This is based on the definition of ΔC_t and $\Delta\Delta C_t$ values, which are respectively the difference in C_t values between the gene of interest and the reference gene and the difference between the ΔC_t of a particular gene for the group of interest and the ΔC_t of that same gene for the reference group. This can be described as follows:

$$\Delta C_t = C_{t_{goi}} - C_{t_{ref}} \quad (2.2)$$

$$\Delta\Delta C_t = \Delta C_{t_{goi}} - \Delta C_{t_{ref}} \quad (2.3)$$

where the *goi* is the gene of interest and the *ref* is the reference gene. $\Delta\Delta C_t$ is on a logarithmic scale with base 2, thus to find the linear fold change in gene expression between two samples, the following formula is used:

$$FC = 2^{-\Delta\Delta C_t} \quad (2.4)$$

The raw RT-qPCR were first imported into the R environment and a qPCRset object was created. A qPCRset is an R class containing the raw or normalised C_t values and related quality information. The data were normalised and analysed using the HTqPCR R package [240]. QC was performed to exclude the samples and genes that failed in more than 50% of the C_t values. PCA was performed to identify and exclude outliers, assess the presence of any plate effects and evaluate the association between PC scores and clinical variables. The ΔC_t normalisation of the C_t values was lastly performed. Following the MIQE guidelines [216], three reference genes were used to normalise the data: *GAPDH* (Glyceraldehyde-3-Phosphate Dehydrogenase), *ACTB* (Actin Beta) and *RNAseP* (ribonuclease P/MRP subunit p30). Their stability across the disease groups in the experiment was assessed before normalisation.

2.6.2 RNA-Seq data

The count matrix, which is a $p \times n$ matrix of integers with p RNA transcripts or genes and n samples, was imported into the R environment [215], and open-source software packages from the R Statistical Computing Language were used for the downstream analysis described below.

Samples with less than 10 million counts were excluded. 10 million counts are, in fact, insufficient to survey the reference genome fully and, hence, might suggest a technical issue or low quality of the RNA. Samples for which the sex at birth in the clinical database did not match the sex retrieved from the RNA-Seq count matrix were flagged. By looking at the expression of the *XIST* (X-inactive specific transcript) gene, it can be assessed whether the biological sample was taken from a male or a female patient. The *XIST* gene is a non-coding RNA on the X chromosome and is only expressed in cells containing at least two X chromosomes (i.e., female cells) [241]. Samples with *XIST* gene expressed but categorised as males in the clinical database, or samples with *XIST* gene not expressed but categorised as females in the clinical database were excluded if no errors in transcribing the clinical records were identified. This was done to exclude samples that might have been mislabeled or misplaced in the laboratory.

A raw count matrix contains over 50,000 genes, with some of them being lowly expressed. Genes with low counts can have a high level of technical noise and/or may be due to poor gene expression

in the tissue or condition being studied. Excluding genes with low counts is an important step in RNA-Seq analysis to ensure the quality and reliability of the data, increase the power of downstream analyses, and eliminate false positive results. It also increases the speed of normalisation, which can be computationally intensive. A pre-filtering step was used to exclude genes with very low counts across all samples [242]. The `filterByExpr` function in the `edgeR` package [243] provides an automatic way to determine which genes have sufficiently large counts to be retained in the statistical analysis. This function retains the genes that have an arbitrary minimum value of summed counts (10 reads counts by default) in a number of samples equal to the sample size of the smaller group. To avoid giving preference to samples with large library sizes, the `filterByExpr` function transforms the counts into Counts Per Million (CPM) values by normalising them by the library size (i.e., sequencing depth) and multiplying the results by a million.

PCA was performed to detect outliers and potential sources of unwanted variation. Outliers in the PCA plot were excluded and, if any batch or plate effect was identified, it was added as an additive term to the statistical model when performing the DE analysis. This was done to ensure that the list of SDE genes was not affected by this technical bias.

Once the quality of the data was assessed, and genes and samples that had failed to pass the QC were excluded, the RNA-Seq data were normalised. During the sample preparation or sequencing process, experimental and technical biases can change the range and distribution of the gene expression values across the samples in the experiment. Normalisation is required to make the expression distributions of each sample similar across the entire experiment [242], to ensure that technical bias has minimal impact on the results and to accurately estimate and detect the SDE genes [244]. `EdgeR` package [243] contains the `calcNormFactors` function, which uses the Trimmed Mean of M-values (TMM) [245] to estimate scale factors between samples that represent sample-specific biases. These factors are then multiplied by the library size to yield the effective library size (i.e., the library size that we would have gotten if those biases were not present). The normalised values were then used in the statistical methods for DE analysis.

2.6.3 ELISA and Luminex® data

The ELISA assay was used to only validate and assess the performance of one protein which had a good discriminating power between children with microbiologically confirmed TB and unlikely TB. Since Luminex® is a multi-plex assay which requires less quantity of biological samples, it was the preferred validation platform. The ELISA assay was used when the proteins were not available via Luminex®.

Since the format of the output data of ELISA and Luminex® assays is very similar, their statistical analysis is analogous. ELISA and Luminex® output data can be exported as “.CSV” files for each plate. Each file contains the values of fluorescence intensity and concentration for each analyte in each sample, along with the standard curves for each analyte in the plate and other values needed to assess the quality and reliability of the data. The fluorescence intensity or the concentration values out of the dynamic range, either under or over the limit of detection, cannot be quantified unless they are very close to the extreme points of the dynamic range, where the software is able to extrapolate their values. Extrapolated values are flagged with an asterisk (*) next to their numerical extrapolated values. The under- and over-the-limit-of-detection values are indicated respectively by OOR< and OOR>, and they do not have any numerical values associated. The over-the-limit-of-detection values were excluded from the analysis since they are between the highest value of the dynamic range and infinity, therefore

the numerical range they might fall into is unknown. Contrarily, the under-the-limit-of-detection values were kept in the analysis. These values fall between zero and either the lowest value of the dynamic range or the lowest extrapolated value, therefore the range they fall into is known. In this thesis, any under-the-limit-of-detection values in each sample plate and for each analyte were assigned either the lowest extrapolated value minus 10%, if any values lower than the lowest point on the standard curve were extrapolated, or half of the lowest value of the dynamic range.

Since the fluorescence intensity is detected by the ELISA or Luminex® reader and normalised against the standards present in each plate, there is no need for any additional data normalisation. However, the quality of the data was assessed by visual inspection of the standard curves and PCA was performed to detect and eventually exclude outliers, to correct for any plate/batch effect and to evaluate any association between the clinical variables and the PCs.

2.6.4 SomaScan® data

SomaScan® assays have 16 control aptamers in each run to pre-process and normalise the data. Normalisation and calibration of SomaScan® data [246] were performed by SomaLogic (Colorado, USA) and can be summarised in three steps:

1. Hybridization Control Normalisation - 12 hybridisation controls are added at incremental and known concentrations to give a series of Relative Fluorescence Unit (RFU) that span the whole dynamic range of the assay. The hybridisation control probes are used to measure the hybridisation efficiency and to remove potential biases introduced by different hybridization efficiencies for assay readout and quantification within and across assay runs. The median of all the RFUs in each plate is divided by the median of the RFUs of all the hybridisation controls giving a single scale factor for each sample. This is then applied to the measured signal of all the features within each biological sample.
2. Intraplate Median Signal Normalisation - This intends to remove the sample-to-sample variations due to differences in protein concentration, pipetting variation, variation in reagent concentrations, assay timing, and other sources of variability within a group of otherwise comparable samples. It is, in fact, performed between samples within the same dilution group (i.e., 0.005%, 1%, and 40%) and the same class (i.e., controls and biological samples). First, the median RFU is calculated for samples in the same sample class and in the same dilution group, and each individual RFU is divided by it. Each sample in each subgroup is median-normalised by multiplying the inverse of the correspondent scale factor to its RFU.
3. Calibration Normalisation - Calibration samples are used to detect and remove systematic variability between independent assay plates. They need to be of the same type as the samples that are being calibrated. The scale factor to be multiplied by each biological sample in the plate is calculated as the inverse of the median of the RFU values of all the calibration samples in a single plate.

The output file of a SomaScan® assay is an “.ADAT” file, which is a tab-delimited text file designed by SomaLogic to store SomaScan® study data. Once the experiment has been run and the output data exported from the machine, SomaLogic send both the unnormalised output data and the already normalised data. Following normalisation, protein abundances were $\log_2$ transformed and PCA was performed to identify and exclude outliers, and to evaluate the association between PC scores and

clinical variables.

2.7 Downstream data analysis

The output of QC and normalisation of raw data is a normalised matrix of features and samples. Statistical analysis methods, mainly machine learning approaches, are then used for downstream analysis. The statistical methods used to analyse the datasets presented in this thesis will be described below. All statistical analyses were performed in R Statistical Computing Language [215].

2.7.1 Differential expression and differential abundance analysis

Differential Expression (DE) and Differential Abundance (DA) analyses use statistical testing on normalised gene counts or protein abundance to decide whether, for a given feature, a quantitative change in expression or abundance levels between experimental groups is statistically significant, that is, whether the quantitative difference is greater than what would be expected just due to natural random variation. Although the most appropriate DE and DA methods may differ accordingly to the type of data that is being analysed, they are based on the same basic principle, which is to use Generalised Linear Models (GLM), a flexible generalisation of ordinary linear regression, to compare the levels of features between the groups of interest. The terms in a statistical model can be divided into the response or dependent variables, which are the quantity that we want to measure (i.e., gene expression), and the explanatory variables, which are the variables that influence the response variables (i.e., clinical or technical variables).

DE and DA analyses are performed on a $p \times n$ matrix where p represents genes or proteins and n represents samples. A linear model is fitted for each feature (row):

$$E[y_j] = X\alpha_j \quad (2.5)$$

where y_j is the expression data for feature j , X is the design matrix and α_j is the matrix of coefficients for feature j . Each row of the design matrix X corresponds to a sample j and the columns correspond to the covariates of interest, which include the covariates that could affect expression/abundance values (i.e., age, sex or sample batch) and the characteristics to be tested (i.e., disease groups that the model aims to compare). The design matrix has two main roles in the data analysis pipeline:

1. It defines the statistical model and relationships between the genes and the explanatory variables
2. It stores the values of the explanatory variables

Finally, the matrix of coefficients α contains the differences in expression/abundance according to the covariates included in X .

The gene expression or protein abundance matrix is analysed as an integrated whole rather than by considering single features in isolation through empirical Bayes (EB) methods, which borrow information between features in a dynamic way [236]. Since the design of linear models fitted across all features in the experiment is the same, the global variability is used to estimate the variance for each gene or protein, in addition to the variance estimated from the feature-specific linear model. This provides more accurate results than if the feature-specific models were used singularly [247].

DE and DA analyses are based on the null hypothesis that the means of all the features are not statistically different in the pairwise comparison of interest. To reject the null hypothesis and assess whether a feature is statistically different, the probability value or p -value has to be lower than an arbitrary threshold or significant level α . The p -value is the probability of obtaining the data if the null hypothesis was true. If the p -value is less than or equal to α , then the null hypothesis is rejected and the difference is statistically relevant. The standard value for the significance level used in this thesis was 0.05, which is translated into a 5% risk of concluding that a difference exists when there is no actual difference. Parametric or non-parametric tests are used to assess whether the means of the features or the mean of ranks of the feature of interest in two populations are different. If the data were normally distributed, parametric tests, like the t -test, were used. When the hypothesis of normality was not met, a non-parametric test, like the Mann–Whitney U test [248], was performed. The Mann–Whitney U test ranks all the dependent values and uses the sum of the ranks for each group to calculate the test statistic. The Shapiro-Wilk test was used to assess if either a parametric or non-parametric statistical test was more suitable. This method evaluates the goodness of fit of a normal distribution to a set of data and is based on the null hypothesis that the samples are part of a normally distributed population. If the resulting p -value was less than the chosen α level, then the null hypothesis was rejected and non-parametric tests were used. Since parametric tests are robust to minor violations of the assumption of normal populations in large datasets, it is acceptable to ignore minor violations with large sample sizes and proceed with the parametric tests. Many variables can also be mathematically transformed to a new scale on which the assumptions of parametric tests are met. The most common data transformation is the logarithmic transformation. When the sample size was large, the data were log-transformed and a parametric test was used to assess whether the difference between the two groups of interest was statistically significant. When the sample size was small, a non-parametric test was used.

Biological datasets might contain up to thousands of features (i.e., genes or proteins), and several statistical tests, one for each feature, need to be performed simultaneously. Since many features are tested simultaneously, it becomes more likely that the groups being compared will appear to differ in terms of at least one feature by chance. This would increase the false positive rate and the risk of rejecting a true null hypothesis (type 1 error). To overcome this problem, which is known as a multiple comparison problem, multiple testing correction approaches need to be used. The Benjamini-Hochberg (BH) correction [249] is a powerful tool that decreases the false discovery rate. It is defined as the expected ratio of false rejections to the number of total rejections. The BH correction method starts with ordering the individual uncorrected p -values in ascending order and assigning them a rank i , from the smallest value being rank 1 and the highest value being rank i . The following formula is then applied for calculating the adjusted p -value for the feature j :

$$p_{(j)}^{BH} = \frac{iQ}{n} \quad (2.6)$$

where i is the rank of the feature, n is the total number of features in the dataset and Q is the acceptable false discovery rate. Typically, a feature is considered to be statistically different across groups only if the p -value remains significant after the multiple testing correction.

There are many tools available for DE analysis, such as edgeR [243], DESeq2 [247], baySeq [250] and EBSeq [251], but the limma approach [242, 252, 253] was chosen for analysing the RNA-Seq data

described in this thesis because it was the most suitable for the study design model (more details are given further down in this section). The limma R package [242, 252, 253] was initially developed for microarray data analysis and then extended to RNA-Seq data analysis. The underlying assumption of the limma package is that the data are normally distributed, which is valid for microarray but not for RNA-Seq data. Since the read counts of a gene do not follow a normal distribution [254], they need to be transformed to converge quicker to normality and, hence, to use linear methods. This is why, in the limma approach, the RNA-Seq read counts are first converted to $\log_2\text{CPM}$. The method then fits a loess trend line to the scatterplot of variance versus mean (i.e., mean-variance relationship) to create weights that are then fed into a standard limma analysis. For the RNA-Seq data presented in this thesis, the voom function was used to transform the read counts to $\log_2\text{CPM}$, estimate the mean-variance relationship and use this to compute appropriate observation-level weights through linear modelling. This function was used along with the duplicateCorrelation function [255], which estimates the correlation between duplicate spots and avoids pseudoreplication by making the statistical model aware of samples that are not statistically independent (i.e., technical replicates). Since there were some bridging samples across the three RNA-Seq batches, the duplicateCorrelation was used to flag what samples were technical replicates. The duplicateCorrelation function was the main reason why the limma pipeline was chosen since there is not a similar function in other RNA-Seq analysis pipelines. After the voom transformation, functions originally written for microarray data analysis can be used on RNA-Seq data. Linear modelling and DE analysis were carried out, followed by empirical Bayes moderation to obtain more precise estimates of gene-wise variability by updating the estimates from information retrieved from all the genes [254]. At that point, the variance was no longer dependent on the mean expression level and linear statistical methods could be applied to identify the genes that are statistically significant across disease groups.

The limma approach has been also extended to other types of transcriptomic data analysis, such as RT-qPCR. The limmaCtData function [240], which is a wrapper for functions from the limma package [255], was used to run DE analysis on the RT-qPCR data described in this thesis. limmaCtData fits linear models for each gene and gives more flexibility than standard parametric or non-parametric tests.

Since no R tools are available for DA analysis in targeted proteomics data analysis, additive linear models accounting for the covariates of interest were manually built in R for larger datasets, where the underlying assumption that the data were normally distributed was valid, whilst BH-adjusted p -values were calculated using non-parametrical tests when the assumption of data normality was not valid and the datasets were relatively small.

The output of DE and DA analyses is a list of estimates of the change in expression/abundance between the two groups of interest, referred to as Fold Change (FC), and its statistical significance (i.e., BH-adjusted p -values). The FC is the ratio of two quantities at two different time points, or, in this case, the ratio between the expression of the same gene or the abundance of the same protein between two different disease groups. It is usually expressed in a logarithmic scale (base 2) since it is easier to interpret and visualise. In fact, the sign (+ or -) of the $\log_2\text{FC}$ reflects the increasing or decreasing levels of gene expression/protein abundance respectively. Arbitrary thresholds for $\log_2\text{FC}$ and BH-adjusted p -value are then chosen for generating the list of Significantly Differentially Expressed (SDE) genes and Significantly Differentially Abundant (SDA) proteins. Thresholds for $\log_2\text{FC}$ and BH-adjusted p -value used in this thesis were respectively $\log_2\text{FC} > 0.5/1$ or $\log_2\text{FC} < -0.5/-1$ and BH-adjusted p -value < 0.05 .

The resulting list of SDE or SDA features can then be used for clustering, pathway enrichment analysis

and/or feature selection.

2.7.2 Clustering

Clustering is an unsupervised approach for grouping data points according to their similarity. Clustering is a powerful tool for the analysis of complex -omics datasets as it uses all available information to detect hidden patterns and reveal natural groups of samples present in data. The term clustering does not refer to a single and unique algorithm, since there are various approaches to grouping the data based on their similarities or dissimilarities. Hierarchical clustering is an algorithm aiming at grouping the data into a hierarchy of clusters. It is less sensitive to outliers compared to other clustering algorithms and it has less stringent assumptions about cluster shape. Hierarchical clustering was used to cluster samples and features in heatmaps in Chapters 3 and 4.

2.7.3 Pathway enrichment analysis

Long lists of SDE/SDA features are often returned from DE/DA analysis, especially for high-throughput -omics assays. Although features of interest can be searched for in the literature, it can be impractical to manually explore their biological function. Pathway enrichment analysis tools detect groups of related features that are altered and provide information about the biological processes that these features are associated with, offering mechanistic insights [256].

A widely used statistical test in pathway enrichment analysis is the Fisher's exact test [257], which determines whether the proportion of features in the pathway (i.e., list of SDE genes) is greater than the proportion of the features outside the pathway (i.e., the background set). The background set is used to reduce the frequency of false positives. This analysis is known as over-representation analysis.

Whilst many pathway enrichment analysis tools exist, g:Profiler2 [258, 259] and clusterProfiler [260, 261] were used in this thesis because they can be run on both gene and protein lists. g:Profiler has both an online platform [<https://biit.cs.ut.ee/gprofiler/gost>; accessed February 11th, 2023] and an R package and it was used for performing pathway enrichment analysis in non-longitudinal data. clusterProfiler was used to perform pathway enrichment analysis in a longitudinal dataset since it is more suitable for time course analyses. Both tools use Gene Ontology (GO) terms, which are either molecular functions (e.g., transport or protein kinase activity), biological processes (e.g., DNA repair or angiogenesis) or cellular components (i.e., transport or cell projection), and other data sources such as Reactome or Kyoto Encyclopedia of Genes and Genomes (KEGG). Each term in the different data sources is associated with a selection of genes that are known to be involved with that specific term. g:Profiler2, clusterProfiler and other pathway enrichment analysis tools perform statistical analysis to determine whether the features in a given term are enriched in the input data at a higher rate than would normally be expected.

2.7.4 Feature selection algorithm

Feature selection algorithms develop a predictive model and reduce the number of input variables selecting the most relevant features. This is a critical step in the identification of biomarkers and signatures with the aim of developing a POCT. The number of biomarkers in a diagnostic signature is a tradeoff between diagnostic performance and cost and feasibility of the POCT itself. A high number of biomarkers can capture more biological variation between disease groups and achieve a better diagnostic

performance, but its translation into real-world applications would be challenging and expensive. Feature selection is a critical step since it needs to identify the minimum set of robust features with the highest diagnostic power. The feature selection method used in this thesis for biomarker identification was Forward Selection-Partial Least Squares (FS-PLS) [<https://github.com/lachlancoin/fspls>; accessed February 11th, 2023]. FS-PLS [262, 263] is an *in-house* feature selection algorithm that identifies small biosignatures with high predictive accuracy. FS-PLS combines the simplicity of Forward Selection (FS) algorithm and the dimensionality reduction strength of Partial Least Squares (PLS) regression. FS is a stepwise regression method that begins with just an intercept term (i.e., empty model) and adds the variables that give the single best improvement to the model one by one. Since FS regression tends to overfit and to add more variables than are actually needed into the model, the PLS component keeps into the final predictive model just uncorrelated and essential variables. In fact, the PLS regression transforms the data into orthogonal and non-correlated latent components, which are then used in a statistical model to predict the outcome of interest. The output of the algorithm is a small model with uncorrelated variables. Given a set of N features and an arbitrary p -value_{thres} (default p -value_{thres} = 0.05), the FS-PLS algorithm can be summarised as follow:

1. N univariate regression models are fitted, one for each feature
2. The coefficients of the regression models are estimated by a Maximum Likelihood Estimation (MLE) function and the goodness of fitting is assessed by a t-test and compared with the arbitrary p -value_{thres}
3. The first variable added to the model has the highest MLE and lowest p -value $< p$ -value_{thres}
4. The variance explained by the variable selected is projected out by using Single Value Decomposition (SVD)
5. The projected variation explained by the variable selected is subtracted from the remaining variables enabling the algorithm to add only uncorrelated variables to the final model
6. The algorithm fits $N-1$ models on the residual variation of each remaining variables
7. Iteratively the algorithm adds new variables one by one following the same criteria and terminates only when there is no remaining variable with MLE p -value $< p$ -value_{thres}

The FS-PLS algorithm is based on linear models to allow both binomial and multiclass regressions. Since the main aim of this thesis can be described as a binary problem (i.e., compare microbiologically confirmed TB to unlikely TB), a logistic regression model is more appropriate to make the most accurate prediction. Model retraining does not change the parameters and variables used in the model, but it adapts the model to the data with the aim of having the best predictor. Once the features in the signature had been selected by FS-PLS, a logistic regression model was used to retrain the predictive model and adapt the coefficients.

2.7.5 Signature performance evaluation

The last step of the analysis pipeline was the signature performance evaluation. The feature selection method returned a set of features with the best performance in discriminating between the groups of interest. Therefore, it was critical to quantify the diagnostic power of the candidate biomarkers and signatures. This is usually done by combining all the features selected by the feature selection algorithm into a Disease Risk Score (DRS) [137]. The DRS sums up the expression or abundance

of the up-regulated features and subtracts from it the sum of the expression or abundance of the down-regulated features for each individual i :

$$DRS^i = \sum_{k=0}^n value_k^i - \sum_{l=0}^m value_l^i \quad (2.7)$$

where n is the number of up-regulated features and m is the number of down-regulated features. The weighted Disease Risk Score (wDRS) is a variation of the DRS and consists in using the actual coefficients of each feature in the predictive model rather than -1 or $+1$. The wDRS for each individual i is calculated as:

$$wDRS^i = \sum_{j=0}^n \beta_j value_j^i \quad (2.8)$$

where n is the number of features included in the model and β are the coefficients of the model for each feature. Since the coefficients retrained by a logistic regression model were more suitable, the wDRS with the retrained β was used in this thesis.

The wDRS can then be used to evaluate the performance of the signatures through sensitivity, specificity, Receiver Operating Characteristic (ROC) curve and Area Under the ROC Curve (AUC). They are all based on four possible outcomes:

- True Positive (TP) - Outcome where the model correctly predicts the positive class
- True Negative (TN) - Outcome where the model correctly predicts the negative class
- False Positive (FP) - Outcome where the model incorrectly predicts the positive class
- False Negative (FN) - Outcome where the model incorrectly predicts the negative class

The sensitivity is the true positive rate. In diagnostics, it is the ability of a test to correctly identify patients with a positive test result. It is calculated as follows:

$$sensitivity = \frac{TP}{TP + FN} \quad (2.9)$$

The specificity is the true negative rate. In diagnostics, it is the ability of a test to correctly identify people with a negative test result. It is calculated as follows:

$$specificity = \frac{TN}{TN + FP} \quad (2.10)$$

Sensitivity and specificity describe the accuracy of a test and are visualised on a ROC curve. The ROC curve is a graphical representation of the relationship of the false positive rate (i.e., $1 - \text{specificity}$) as a function of the true positive rate (i.e., sensitivity) at various thresholds [Figure 2.9].

To generate a ROC curve, in fact, sensitivity and specificity are first tabulated for different values of a continuous test measure, and then the true positive rate (i.e., sensitivity) is plotted on the y-axis against the false positive rate (i.e., $1 - \text{specificity}$) on the x-axis for the various values tabulated. The ROC analysis is useful to:

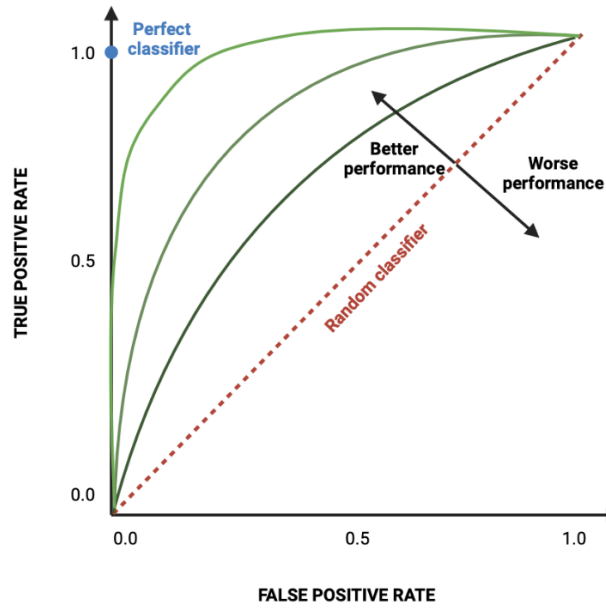


Figure 2.9: Example of a ROC curve.

A random classifier (red dashed line) represents a random guess. The closer the curve to the top left corner (blue dot), the better the diagnostic power. The test fit is usually in between the perfect and the random classifier. *ROC = Receiver Operating Characteristic*

- Evaluate the ability of a continuous biomarker, such as the DRS or wDRS, in assigning the samples into a binary classification correctly
- Find a cut-off point between the false and true positive rates to minimise the number of misclassifications
- Compare the accuracy of two or more biomarkers, signatures or diagnostic tests

Since the ROC curve is a combination of two measures (i.e., true and false positive rates), several summary measures have been proposed in the literature. The Area Under the ROC Curve (AUC) and Youden's J index are among the most widely used. The AUC is the area under the ROC curve and ranges in value from 0 to 1. The closer to 1, the better the diagnostic power. It is given by:

$$AUC = \int_0^1 ROC(u) du \quad (2.11)$$

An AUC of 0.5 represents a test with no ability to discriminate between two conditions (i.e., no better than chance), while an AUC of 1 represents the perfect test with the perfect discriminating ability.

Youden's J index [264] is used to detect the best threshold and, therefore, to capture the best performance in terms of sensitivity and specificity of a dichotomous diagnostic test. It is defined as follows:

$$J = \max_c \{ \text{sensitivity}(c) + \text{specificity}(c) - 1 \} \quad (2.12)$$

where c is the cut-off point that optimises the discriminating ability of the test.

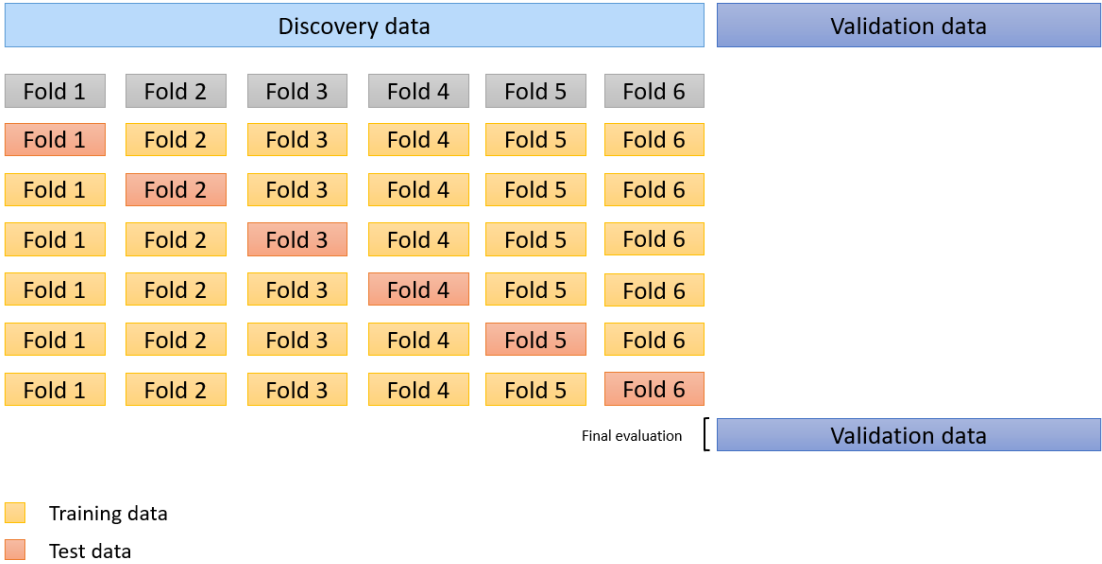


Figure 2.10: Cross-validation algorithm. The discovery data are split into k folds. The iterative algorithm trains the model on $k-1$ folds and tests it on the k^{th} remaining fold, changing the fold used to test the model at each iteration. The final model is then validated in one or more independent datasets to ensure the reproducibility of the results.

Youden’s J index was used in this thesis to identify the best threshold on the ROC curve and, hence, the values of sensitivity and specificity with the best diagnostic performance.

2.7.6 Cross validation

To avoid overfitting and identify biomarkers with the best diagnostic power and discriminating ability, a cross-validation step was added to the data analysis pipeline. Learning and testing the parameters of a model on the same data might, in fact, lead to biomarkers and coefficients that have good diagnostic power in the discovery data, but might fail to predict the right outcomes on yet-unseen data. This is defined as overfitting.

To overcome this problem, the cross-validation algorithm was used in this thesis in conjunction with the FS-PLS algorithm to avoid overfitting and obtain more reproducible results. An iterative process with 100 iterations was developed. At each iteration, 80% of the discovery data was randomly selected to be the training set, while the remaining 20% of the data was assigned to be the test set. FS-PLS was then run on the training set of each iteration and generated a model which was tested on the test set. The most robust signature (i.e., the signature selected more times across the 100 iterations) was chosen as the best signature to take forward to the validation step [Figure 2.10].

2.8 Multi-omic data integration

Single-omics data analysis may not always provide enough information to give a deep understanding of a biological system. Single-omics data give insights about the single layers analysed but do not account for the interactions about different -omics layers. In Sections 2.5, 2.6 and 2.7 of this chapter, the methods used to analyse the transcriptomic and proteomic datasets in this thesis are respectively described. Since there was enough overlap of samples with transcriptomic and proteomic data available, a multi-omics approach was used to identify a multi-omics diagnostic signature to differentiate between

TB disease and unlikely TB in children and assess whether a multi-omics signature performed better than a single-omics transcriptomic or proteomic signature. For this aim, the MixOmics R package [203] was used.

MixOmics [203] is an R toolkit developed to visualise and integrate omics datasets with a focus on biomarker selection. Graphical representation is a strong feature of this R package since the authors wanted to give the users a tool to better understand the relationship and the correlation between the different layers of -omics at both sample and feature levels. This package is based on multivariate methods, which are well-suited for large datasets with a number of features much larger than the number of samples. Multivariate methods reduce the dimension of the data by combining the features into latent variables defined as components. These components are then used to plot graphical outputs which are helpful to the user in understanding the relationships and correlations between the different integrated -omics datasets. MixOmics offers a wide range of sparse multivariate methods to identify the key features that are highly correlated between the combined single-omics datasets and explain the biological outcome of interest. In addition, MixOmics can handle missing data. Hence, entire rows or columns with null values in the integrated multi-omics dataset are not excluded. This R package has been introduced in the context of supervised analysis since its aim is to find the most relevant biological features able to discriminate and classify already categorised sample groups and predict the class of new samples.

MixOmics is one of the few available tools to perform both N- and P-integration. N-integration, or vertical integration, is the integration of the same N biological samples run on different single-omics platforms. This is performed by the Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO) framework in MixOmics [265]. P-integration, or horizontal integration, is the integration of independent datasets or studies with the same P features. This is performed by the Multivariate INTEgrative method (MINT) framework in MixOmics [266]. Whilst N-integration is usually used to integrate multiple single-omics datasets available for the same biological samples or combine the molecular signatures identified from separate analyses of each -omics on the same subset of samples, the P-integration is usually used to increase the sample size and the statistical power. The multivariate methods in MixOmics can handle features on a continuous scale (such as RT-qPCR, Luminex®, SomaScan® and ELISA data) or count data which become continuous after normalisation (i.e., RNA-Seq data).

The following pre-processing steps are needed for the data to be correctly handled by the MixOmics package:

- Since the pre-processing and normalisation vary from -omics to -omics, MixOmics assumes that the input data have been pre-processed and normalised when applicable. Therefore, single-omics datasets need to be pre-processed and normalised prior to the analysis
- MixOmics can handle high-throughput data. However, it is recommended to pre-filter the data to less than 10k features per dataset to lessen the computational time
- Since MixOmics uses matrix decomposition techniques, the input datasets need to have n observations or samples in rows and p predictors or variables (e.g. genes, proteins) in columns
- Covariates that might confound the analysis need to be accounted for before the analysis in MixOmics, since its pipeline does not currently include them in the methods

2.8.1 DIABLO pipeline

The single-omics datasets need to be pre-processed and normalised before the integration with MixOmics. DIABLO [265] is a supervised MixOmics integrative method that identifies the key and most correlated biomarkers among the single-omics datasets whilst discriminating the disease groups of interest. DIABLO is an extension of the sparse Generalized Canonical Correlation Analysis (sGCCA) [267], a multivariate dimension reduction method that maximises the covariance between linear combinations of the features and projects out the data into a reduced dimensional space. The selection of the correlated variables across single-omics levels is performed with l_1 penalisation on the variable coefficients of the linear combinations. Having Q pre-processed and normalised single-omics dataset $[X^{(1)}(N \times P_1), X^{(2)}(N \times P_2), \dots, X^{(Q)}(N \times P_Q)]$ with $P_1, P_2, \dots, P_Q$ omics variables and N same samples, the sGCCA solves the following optimisation function for each dimension h :

$$\max_{a_h^{(1)}, \dots, a_h^{(Q)}} \sum_{i,j=1, i \neq j}^Q c_{i,j} \text{cov}(X_h^{(i)} a_h^{(i)}, X_h^{(j)} a_h^{(j)}) \quad (2.13)$$

$$\text{such that } \|a_h^{(q)}\|_2 = 1 \text{ and } \|a_h^{(q)}\|_1 \leq \lambda^{(q)} \text{ for all } 1 \leq q \leq Q$$

where $a_h^{(q)}$ is the variable coefficient (loading) vector on the dimension h associated with the residual matrix $X_h^{(Q)}$ of the single-omics dataset $X^{(Q)}$. $\lambda^{(q)}$ is a non-negative parameter that controls the number of zeros in the loading vector ($a_h^{(q)}$) and thus the amount of shrinkage. The subset of variables with non-zero coefficients are highly correlated between and within the single-omics datasets. The design matrix ($C = \{c_{i,j}\}$) accounts for the correlation between the single-omics datasets. Elements in the design matrix can be set between 0 and 1 by the user. Zero reflects no correlation, 1 full correlation.

This method is iterative and maximises Equation 2.13 for each dimension h . It starts with maximising Equation 2.13 for $h=1$ and residual matrix $X_1^{(q)} = X^{(q)}$. The residual matrix, loading vector and component score of the previous iteration step are then used for $h=2$ ($X_2^{(q)} = X_1^{(q)} - t_1^{(q)} a_1^{(1)}, 1 \leq q \leq Q$). This iterative process continues until a sufficient number of dimensions is obtained. The major sources of biological variation among the single-omics datasets are captured by the first set of component scores $t_1^{(q)}, \dots, t_h^{(Q)}$, whilst variation due to heterogeneity across the single-omics datasets $X^{(q)}$ does not impact the statistical model.

DIABLO extends sGCCA by adding to the single-omics datasets $X^{(q)}$ in Equation 2.13 the disease categorisation of each N biological samples. It also replaces the l_1 penalty parameter $\lambda^{(q)}$ with an arbitrary number of variables that the user wants to select in each single-omics dataset.

Since DIABLO is based on Projection to Latent Structures (PLS), the single variables are scaled and centred internally, therefore it does not assume particular data distributions of the input data. The `block.splsda` function takes in input the single-omics datasets and the disease categorisation. The output is a list of the most correlated features among the single-omics datasets that are also the most discriminant among the disease groups of interest. The multi-omics biomarkers can then be combined into a signature and its performance evaluated in terms of AUC, sensitivity and specificity.

CHAPTER 3

DISCOVERY AND VALIDATION OF A SIX-MARKER TRANSCRIPTOMIC SIGNATURE FOR DIAGNOSIS OF CHILDHOOD TUBERCULOSIS IN AN AFRICAN MULTI-COUNTRY COHORT

Despite the identification of TB's causative agent over a century ago, TB is nowadays one of the top ten causes of death worldwide [2]. The majority of the disease burden occurs in LMIC, where most of the undiagnosed and untreated cases reside.

There is still no test for TB that is affordable, rapid, easy-to-use and appropriate for resource-limited settings. Most current diagnostic tests for TB require high-infrastructure laboratories and skilled technicians, are costly, time-consuming and lack sensitivity and/or specificity. The diagnosis of TB is even more difficult in children, where usually invasive procedures are needed to collect biological samples, clinical diagnosis is challenging for symptomatology similar to other paediatric diseases, and the detection of *Mtb*, the causative agent of TB, is more difficult due to its paucibacillary nature.

Recent years have seen a shift towards the detection of changes in the host immune system in biological samples caused by the pathogen rather than detecting the pathogen itself. New high-throughput quantification techniques, such as microarrays and RNA-Seq, have enabled the identification of sparse sets of gene biomarkers - gene signatures - by comparing the blood RNA expression between different disease groups or disease states and/or healthy controls. For the identification of diagnostic biomarkers, the disease group contrasted to TB is either healthy individuals or made up of patients with diseases with symptomatology similar to those with TB.

In the past few decades, many blood gene expression studies aiming at identifying a diagnostic signature for TB have been published. Most signatures were discovered in adult cohorts and underperformed in children, whilst the few published paediatric studies focused on specific countries or included a large number of genes, making their development into a POCT difficult and expensive.

3.1 Overview and chapter aims

The diagnosis of paediatric TB poses many challenges and has not always been given high priority by international TB control programmes. Childhood TB is, in fact, difficult to diagnose because it mimics other common childhood infections, such as pneumonia, and often requires invasive procedures to collect biological samples, with current diagnostic tests being less accurate in children compared to adults. More recently, there has been a renewed interest and awareness of the importance of controlling paediatric TB worldwide, which has led to an interest in developing new diagnostic tests, therapeutic formulations, surveillance and reporting systems [268]. Since invasive procedures, such as induced sputum or gastric aspiration, are often needed to collect samples from children, the WHO has endorsed the development of new non-sputum and biomarker-based diagnostic tests.

Chapter 3 of this thesis describes the discovery by RNA-Seq and validation by RT-qPCR of a parsimonious gene signature for the diagnosis of childhood TB. The main aims of this chapter are:

1. To explore the whole blood host transcriptomic profiles of children with TB disease and contrast them to the whole blood host transcriptomic profiles from children with unlikely TB
2. To identify a sparse host gene diagnostic signature that can be used to distinguish TB disease from unlikely TB in children
3. To validate the candidate biomarkers identified in Aim 2 using RT-qPCR in the same cohort of samples (cross-platform validation) and independent cohorts (independent validation)

The structure of this chapter is as follows: Section 3.2 describes the clinical cohorts and the data generation pipelines. Results and some biological insights are discussed in Section 3.3. The chapter ends with a summary and discussion of the results in Section 3.4.

3.2 Cohorts and study design

The study comprised three phases: the discovery phase, the cross-platform validation phase and the independent validation phase. Candidate biomarker genes to distinguish TB disease from unlikely TB were identified in the discovery phase using host whole blood transcriptomic profiles from RNA-Seq (work led by Dr Sandra Newton and Dr Victoria Wright, Imperial College London, UK), a transcriptomic high-throughput method which quantifies and characterises known and unknown RNA molecules present in the biological sample described in Section 1.10.3. The expression of genes identified as candidate biomarkers in the discovery phase was then measured by RT-qPCR (work led by Dr Jesus Rogriguez Manzano, Imperial College London, UK), a transcriptomic technique that quantifies a small set of target genes described in Section 1.10.1, in the same set of samples (cross-platform validation) and in independent cohorts (independent validation).

The discovery and validation cohorts included children (<15 years of age) with microbiologically confirmed TB disease, unconfirmed TB disease and unlikely TB. Children were recruited in 4 African countries with a high burden of TB (i.e., South Africa, Malawi, Kenya and The Gambia) between 2008 and 2017 and across different studies funded by the EU, Wellcome Trust, MRC and NIH [Section 2.4]. Clinical data were anonymised and biological samples were assigned a unique study number. The common main aims of the recruiting studies were to identify transcriptomic and proteomic biomarkers for childhood TB and assess their diagnostic performance in distinguishing TB disease from other diseases with similar symptomatology (i.e., unlikely TB). Children were recruited in the studies if they were presenting to hospitals or healthcare centres with suspected TB disease, or if they had documented TB exposure and were diagnosed with TB infection.

The clinical features of the children included in the discovery RNA-Seq, cross-platform validation RT-qPCR and independent validation RT-qPCR cohorts are described in Table 3.1. The discovery cohort included 783 samples collected from patients with microbiologically confirmed TB ($n=279$), unconfirmed TB ($n=185$) and unlikely TB ($n=319$). Children in the discovery cohort categorised as microbiologically confirmed TB and unlikely TB, after excluding the samples there was no RNA left for ($n=163$), were included in the cross-platform validation phase. Contrarily, the samples used in the independent validation phase had not been used previously. Children with microbiologically confirmed TB were slightly older compared to children with either unconfirmed TB or unlikely TB in the discovery and validation datasets. Females and males were balanced across the disease groups, with a slightly lower percentage of females in the unlikely TB group. The majority of children in the three cohorts

were HIV-uninfected. Most of the microbiologically confirmed TB and unconfirmed TB cases had PTB and most of the unlikely TB cases had a respiratory tract infection.

Discovery RNA-Seq dataset			
	Confirmed TB (n=279)	Unconfirmed TB (n=185)	Unlikely TB (n=319)
Age (months), median (IQR) <i>missing</i>	56.0 (22.5-99.5) 0	31.0 (15.0-60.5) 0	40.0 (17.0-86.5) 0
Female, no. (%) <i>missing</i>	138.0 (49.5%) 0	84.0 (45.4%) 0	146.0 (45.8%) 0
HIV-infected, no. (%) <i>missing</i>	62.0 (22.2%) 0	44.0 (23.8%) 0	102.0 (32.0%) 0
PTB, no. (%) <i>missing</i>	251.0 (90.0%) 0	167.0 (90.3%) 1 (0.5%)	
Respiratory tract infection, no. (%) <i>missing</i>			141.0 (44.2%) 5.0 (1.6%)
Cross-platform validation RT-qPCR dataset			
	Confirmed TB (n=204)		Unlikely TB (n=231)
Age (months), median (IQR) <i>missing</i>	78.5 (37.8-122.0) 0		58.0 (23.0-109.0) 0
Female, no. (%) <i>missing</i>	108.0 (52.9%) 0		100.0 (43.3%) 0
HIV-infected, no. (%) <i>missing</i>	50.0 (24.5%) 0		79.0 (34.2%) 0
PTB, no. (%) <i>missing</i>	183.0 (89.7%) 0		
Respiratory tract infection, no. (%) <i>missing</i>			112.0 (48.5%) 5.0 (2.2%)
Independent validation RT-qPCR dataset			
	Confirmed TB (n=52)	Unconfirmed TB (n=63)	Unlikely TB (n=57)
Age (months), median (IQR) <i>missing</i>	97.5 (49.0-144.8) 0	76.0 (36.5-130.0) 0	79.0 (41.0-110.0) 0
Female, no. (%) <i>missing</i>	24.0 (46.2%) 0	27.0 (42.9%) 0	19.0 (33.3%) 0
HIV-infected, no. (%) <i>missing</i>	13.0 (25.0%) 0	7.0 (11.1%) 0	12.0 (21.1%) 0
PTB, no. (%) <i>missing</i>	52.0 (100.0%) 0	58.0 (92.1%) 0	
Respiratory tract infection, no. (%) <i>missing</i>			30.0 (52.6%) 0

Table 3.1: Clinical features of patients included in the discovery RNA-Seq, cross-platform validation RT-qPCR and independent validation RT-qPCR datasets of novel gene biomarkers for distinguishing between TB disease and unlikely TB in children.

RNA-Seq: RNA Sequencing; RT-qPCR: Reverse Transcription - quantitative Polymerase Chain Reaction; IQR: Interquartile range; HIV: Human Immunodeficiency Virus; PTB: Pulmonary Tuberculosis

3.2.1 Case definitions

The diagnostic algorithm to classify recruited children into microbiologically confirmed TB, unconfirmed TB and unlikely TB is described in detail in Section 2.3. In brief, microbiologically confirmed TB diagnosis was based on the bacteriological confirmation either by culture or NAATs from at least 1 specimen obtained from the site of TB. A child was diagnosed with unconfirmed TB when the bacteriological confirmation was not obtained and the patient had at least 2 of the following: symptoms or signs suggestive of TB; close TB exposure or immunological evidence of *Mtb*; positive response to TB treatment; chest radiography or other evidence suggestive of TB. When the bacteriological confirmation was not obtained and the criteria for “unconfirmed TB” were not met, the symptomatic child was classified as unlikely TB. A healthy child with no clinical features of TB disease and a positive IGRA or

TST result was classified as having *Mtb* infection.

Although the clinical algorithm describes the classification of patients into 4 diagnostic groups, only microbiologically confirmed TB, unconfirmed TB and unlikely TB patients were included in the analysis described in this chapter.

3.2.2 Data generation

Discovery RNA-Seq dataset The discovery dataset was generated using RNA-Seq on whole blood. Whole blood was collected into PAXgene Blood RNA Tubes (PreAnalytiX) from patients at the earliest time point after recruitment and prior to any treatment. PAXgene tubes were frozen within 6 hours after collection and stored at -80°C until extraction. Total RNA was isolated using the PAXgene Blood RNA Kits following the recommended methodology and sent for sequencing at The Wellcome Centre for Human Genetics in Oxford, United Kingdom.

The protocol followed to generate the RNA-Seq data was as follows. Genomic material was quantified using RiboGreen (Invitrogen) on the FLUOstar OPTIMA plate reader (BMG Labtech) and the size profile and integrity were analysed on the 2200 TapeStation (Agilent, RNA ScreenTape). Input material was normalised, and strand-specific library preparation was completed using NEBNext® Ultra™ II mRNA kit (NEB) and NEB rRNA/globin depletion probes following the manufacturer's instructions. Libraries were prepared on a Tetrad (Bio-Rad) using *in-house* unique dual indexing primers [269]. Individual libraries were normalised using Qubit and pooled together. The pooled library was diluted to ~10 nM for storage, denatured and further diluted prior to loading on the sequencer. Samples were randomised across sequencing plates. Sequencing was performed at the Oxford Genomics Centre using a Novaseq6000 platform at 150 base-pair paired-end configuration, generating a raw read count of 30 million reads per sample. Following sequencing, the raw RNA-Seq reads were processed as outlined in Section 2.5.2.

Validation RT-qPCR datasets Genes identified as candidate biomarkers were measured by RT-qPCR on the same set of samples used in the discovery phase (i.e., cross-platform validation) and in an independent set of samples (i.e., independent validation). Three reference genes - *GAPDH* (Glyceraldehyde 3-Phosphate Dehydrogenase), *ACTB* (Actin Beta) and *RNAseP* (Ribonuclease P/MRP Subunit P30) - were also quantified for normalisation purposes, following the MIQE guidelines [216] [Section 2.6.1]. Each gene was run in quadruplicates (i.e., it was quantified using 4 replicates). Primers for the candidate diagnostic biomarkers and the reference genes were designed *in-house*. Gene sequences were downloaded from Ensembl Genome Browser (<https://www.ensembl.org/index.html>) using the assembly Genome Reference Consortium Human Build 38 patch release 14 (GRCh38.p14) as reference human genome. Primer design and *in silico* analysis were conducted using GENEious Prime 2022.0.1 (<https://www.geneious.com/>). Primer characteristics were evaluated through IDT OligoAnalyzer software (<https://eu.idtdna.com>) using the J. SantaLucia thermodynamic table for melting temperature calculation, hairpin, self-dimer, and cross-primer formation (<https://www.thermofisher.com>). Intercalating dye-based PCR assays were designed, developed and purchased from IDT Integrated DNA Technologies (Coralville, IA). Samples were randomised across RT-qPCR plates and the target genes were quantified by using a Biomark HD high-throughput qPCR platform and 192.24 Dynamic Array integrated fluidic circuit (IFC) following the manufacturer's instructions. The experimental workflow consisted of 4 steps: reverse transcription (RT), pre-amplification, on-chip

gene expression, and data analysis.

The RT-qPCR data were pre-processed by performing baseline correction (i.e., the noise level in early cycles where there is no detectable increase in fluorescence) and C_t values were extracted using a global threshold, as determined by Fluidigm's Real-Time qPCR Software (version 4.1.2). The principles of RT-qPCR are described in more detail in Section 1.10.1. In brief, RT-qPCR measures the actual amount of DNA produced by the PCR reaction at each cycle by measuring the fluorescence intensity of fluorescent reporter molecules added to the reaction. The cycle in which the fluorescence intensity is greater than the background fluorescence is used as a measure of the target DNA's starting concentration. Lower C_t values indicate that fewer PCR cycles were required for the fluorescence intensity to exceed the background signal and a greater quantity of target DNA was present in the biological sample at the start of the PCR reaction (i.e., the gene was more expressed). Contrarily, higher C_t values indicate less quantity of target DNA at the start of the PCR reaction with the target gene being less expressed. A 40-cycle RT-qPCR experiment was carried out to measure the expression of the candidate biomarker genes and the reference genes.

3.3 Results

The aim of this study was to discover and validate host gene expression biomarkers for distinguishing between TB disease and unlikely TB in children. Host whole blood RNA-Seq was used to identify the candidate genes and RT-qPCR was used to validate the host biomarkers in the samples used in the discovery phase (i.e., cross-platform validation) and independent cohorts (i.e., independent validation). A 6-gene signature was identified by FS-PLS, a feature selection algorithm described in Section 2.7. In brief, FS-PLS starts with an empty model and adds, one at a time, non-correlated and highly predictive features using SVD to project out the variance of the residual variance captured by the features already included. The result is a small signature composed of uncorrelated features. The 6 candidate gene biomarkers were then taken forward for validation and their performance was evaluated in independent cohorts.

3.3.1 Discovery phase

Data pre-processing and normalisation Whole blood RNA-Seq of children with microbiologically confirmed TB disease ($n=279$), unconfirmed TB ($n=185$) and unlikely TB ($n=319$) were included in the discovery cohort. The discovery cohort included HIV-infected ($n=208$) and HIV-uninfected ($n=575$) children. The discovery set was composed of three sequencing batches known as batch A, batch B and batch C, with 72, 78 and 633 samples respectively. Prior to merging the three datasets, pre-processing was performed on each dataset separately. Genes with low counts in at least one dataset were excluded with the `filterByExpr` function [243]. This function retains the genes that have an arbitrary minimum value of summed counts (10 reads counts by default) in a number of samples equal to the minimum group sample size [Section 2.6.2]. Principal Component Analysis (PCA) was performed to identify and remove the outliers, followed by the identification and exclusion of samples with *XIST* gene expressed but categorised as males in the clinical database, or samples with *XIST* gene not expressed but categorised as females in the clinical database. As described in Section 2.6, the *XIST* (X-inactive specific transcript) gene is a non-coding RNA on the X chromosome and is only expressed in cells containing at least two X chromosomes (i.e., female cells) [241]. If no errors in transcribing the

clinical records were identified, samples were excluded due to possible mislabeling or misplacing in the laboratory.

PCA was performed on the RNA-Seq dataset to identify outliers and any source of unwanted variation. The primary source of variation captured by the first 4 PCs (i.e., PC1, PC2, PC3 and PC4) in the pre-processed raw RNA-Seq was the sample batch [Figure 3.1].

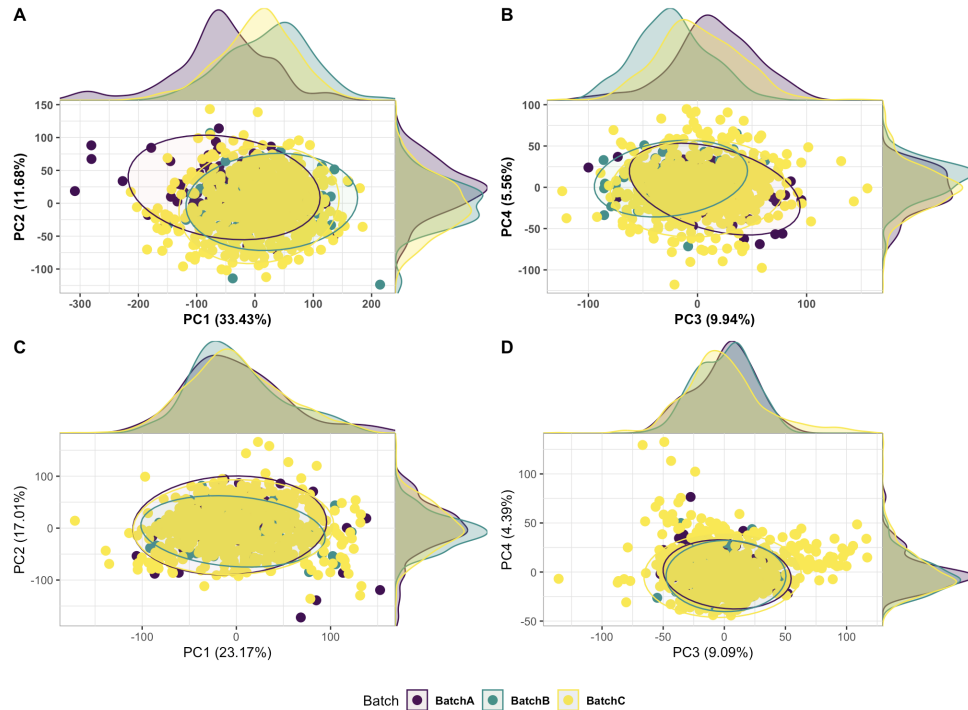


Figure 3.1: PCA biplots from PCA performed on the discovery RNA-Seq dataset coloured according to sequencing batch.

PC 1-2 and PC 3-4 are shown before [Panel A and Panel B] and after batch correction [Panel C and Panel D]. Points represent samples. The distributions of PC1, PC2, PC3 and PC4 are shown in the density plots. PCA: Principal Component Analysis; RNA-Seq = RNA Sequencing; PC: Principal Component

The discovery RNA-Seq experiment was in fact run across three different batches, and this technical variation could be identified when plotting PC1 against PC2 and PC3 against PC4 [Figure 3.1; Panels A, B]. The batch effect was accounted for in the DE analysis and corrected with the use of the `removeBatchEffect` function [236] [Figure 3.1; Panels C, D].

PCA was performed anew after normalisation and batch correction of the data. No other source of technical variability was captured by the first PCs [Figure 3.2]. The disease group was captured by PC2 [Figure 3.2; Panel A] with the microbiologically confirmed TB group towards the bottom right of the plot and the unlikely TB group towards the top left of the plot. The unconfirmed TB patients lay in between the confirmed and unlikely cases. HIV status was captured by PC2 with the HIV-infected group towards the top of the plot [Figure 3.2; Panel C]. The recruitment site was captured by PC2, with the ILULU Cape Town and ILULU Malawi cohorts towards the top of the plot, the Gambian cohort towards the bottom and the NIH Cape Town and Kenyan cohorts at the centre of the plot [Figure 3.2; Panel E]. PC3 was also capturing the recruitment site, with the Gambian samples clustering towards the right of PC3 [Figure 3.2; Panel F]. No differences were observed in terms of sex or age [Figure 3.2; Panels G, H, I, J].

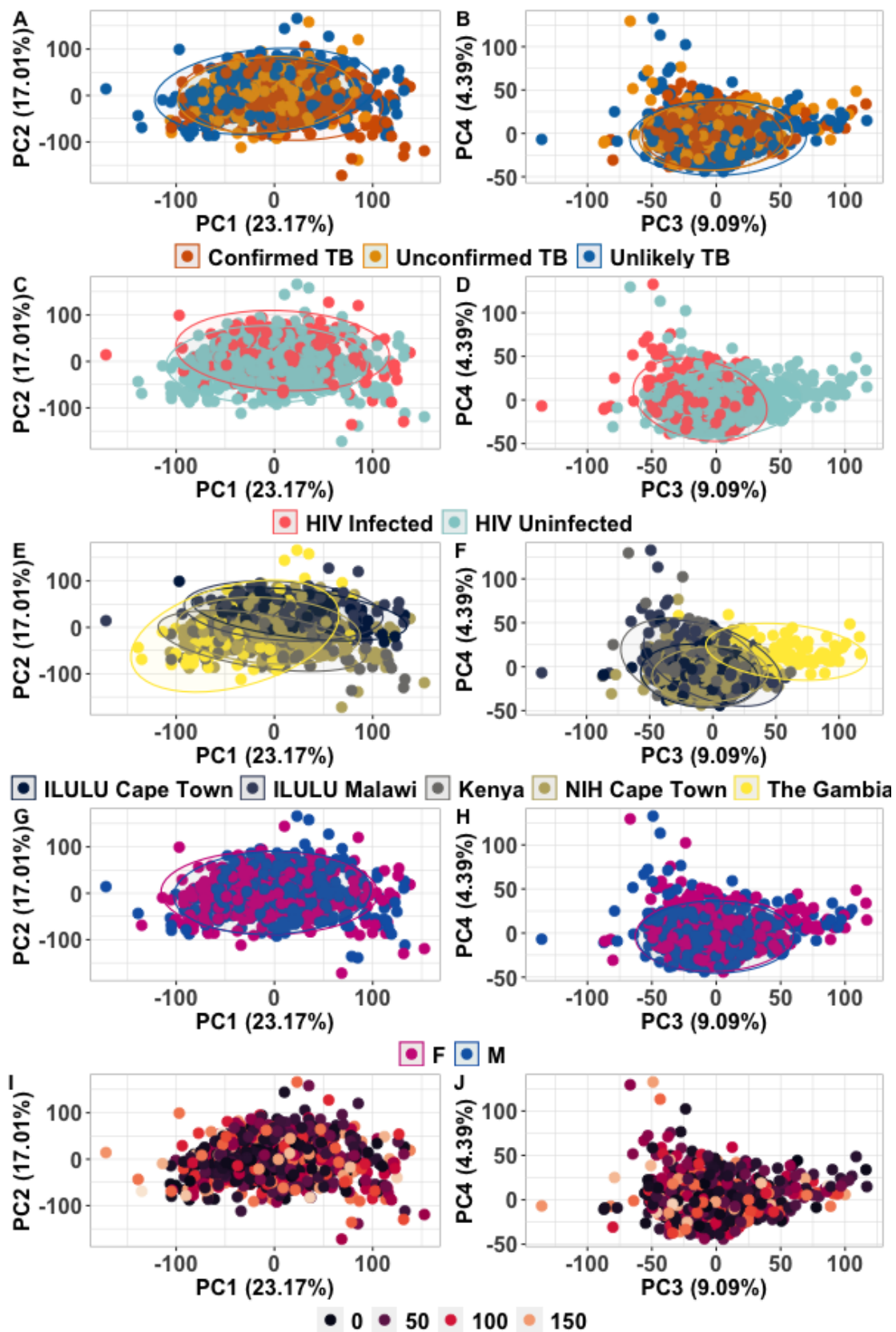


Figure 3.2: PCA biplots from PCA performed on the discovery RNA-Seq dataset coloured according to clinical variables.

PC 1-2 and PC 3-4 are shown. Points represent samples and are coloured according to disease group [Panel A and Panel B], HIV status [Panel C and Panel D], recruitment site [Panel E and Panel F], sex [Panel G and Panel H] and age [Panel I and Panel J]. *PCA*: Principal Component Analysis; *PC*: Principal Component; *HIV* = Human Immunodeficiency Virus; *F* = female; *M* = male

The counts from the pre-processed datasets were merged using only genes present in the three datasets

and normalised using the edgeR package [243] [Section 2.6.2]. DE analysis was performed using the limma R package [236] with age and sample batch accounted for and included in the additive statistical model. The duplicateCorrelation function [255] was used to account for the technical replicates in the statistical model. Sex was not added to the model since it was not significantly different (uncorrected p -value >0.05) between the disease groups of interest (i.e., microbiologically confirmed TB and unlikely TB). Genes with BH-adjusted p -value <0.05 and $\log_2FC >0.5$ or $\log_2FC <-0.5$ were considered SDE. 178 genes (7 under- and 171 over-expressed) [Appendix B] were SDE between microbiologically confirmed TB and unlikely TB [Figure 3.3].

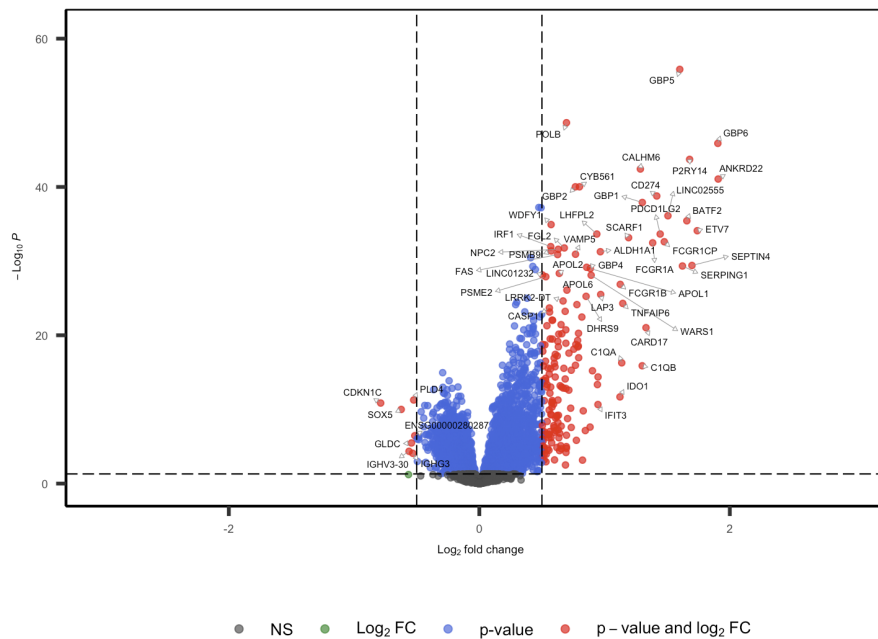


Figure 3.3: Volcano plot showing $\log_2FC$ and BH-adjusted p -values from DE analysis comparing microbiologically confirmed TB vs. unlikely TB in the discovery RNA-Seq data.

Points represent genes and are coloured as follows: red - genes with BH-adjusted p -values <0.05 and absolute $\log_2FC >0.5$; blue - genes with BH-adjusted p -values <0.05 and absolute $\log_2FC <0.5$; green - genes with BH-adjusted p -values >0.05 and absolute $\log_2FC >0.5$; grey - genes with BH-adjusted p -values >0.05 and absolute $\log_2FC <0.5$. Genes to the right of zero are over-expressed in TB. FC = Fold Change; BH = Benjamini-Hochberg; DE = Differential Expression; RNA -Seq = RNA Sequencing

Feature selection The technical replicates were averaged after normalisation and before the feature selection step. FS-PLS with cross-validation [Section 2.7] was used on the SDE genes to identify a small diagnostic signature to distinguish microbiologically confirmed TB from unlikely TB. The feature selection algorithm was run across 100 iterations with a randomly selected training set (80% of the data) at each time. The signature was then tested on the test set of that specific iteration (the remaining 20% of the data). The results of the cross-validation are shown in Figure 3.4. *GBP5* was the most discriminating gene when comparing microbiologically confirmed TB vs. unlikely TB, it was in fact selected in each of the 100 signatures identified across the 100 iterations. The signature taken forward for validation was a 6-gene signature composed of *GBP5* (Guanylate Binding Protein 5), *CDKN1C* (Cyclin Dependent Kinase Inhibitor 1C), *CPVL* (Carboxypeptidase Vitellogenic Like), *CYB561* (Cytochrome B561), *RSAD2* (Radical S-Adenosyl Methionine Domain Containing 2) and *EFCAB2* (EF-Hand Calcium Binding Domain 2). This was the signature selected the most across the 100 iterations (first column in Figure 3.4) and thus the most robust and less prone to overfitting.

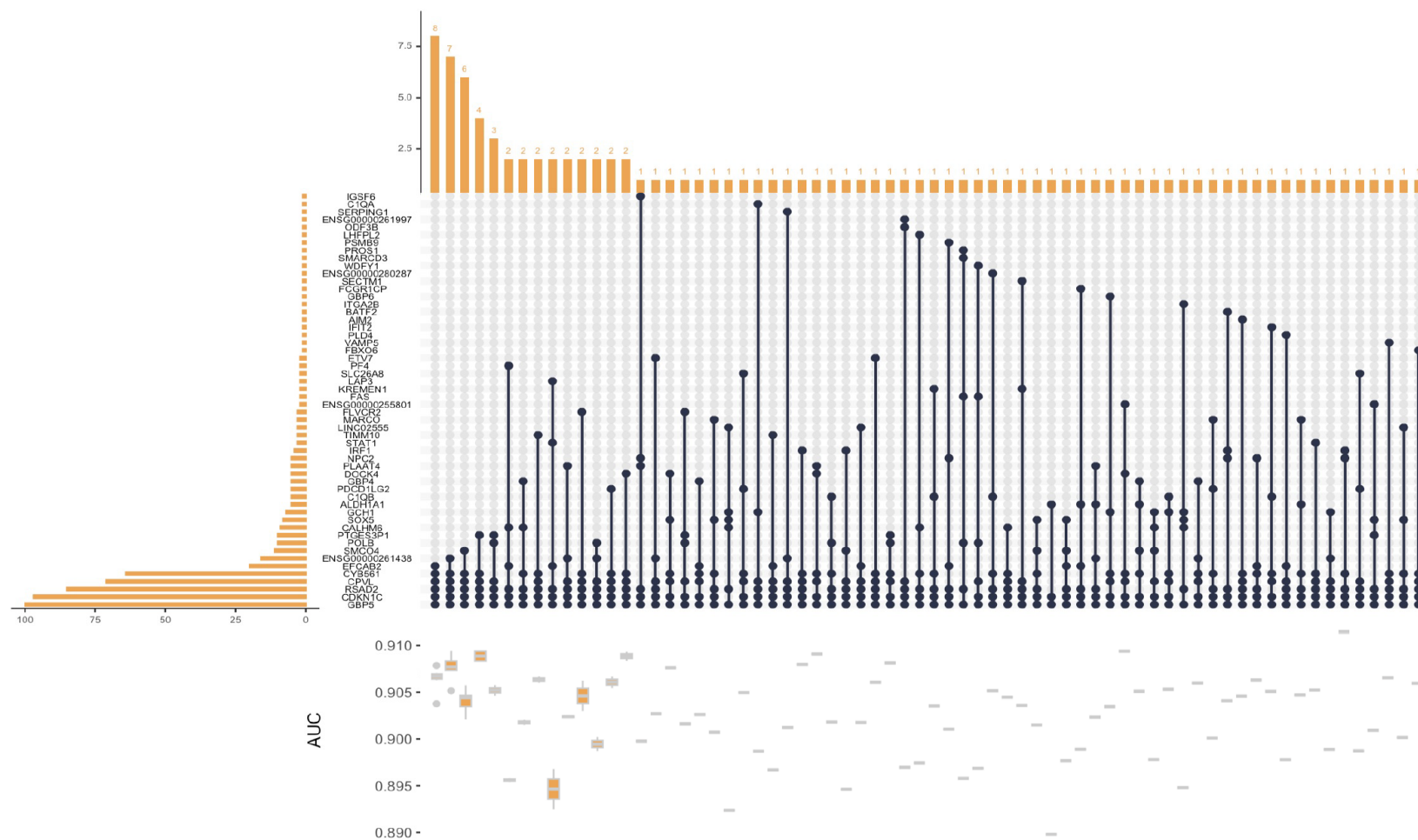


Figure 3.4: Results of FS-PLS with cross-validation in the discovery RNA-Seq dataset.

Each column corresponds to a signature composed of the genes indicated as a blue dot. The barplots at the top correspond to the times each signature was selected across the 100 iterations. The barplots at the left-hand side correspond to the times each gene was selected in any of the signatures identified across the 100 iterations. The boxplots at the bottom represent the range of the AUC for each signature in the whole dataset (i.e., training and test sets). *FS-PLS* = *Forward Selection-Partial Least Squares*; *RNA-Seq* = *RNA Sequencing*; *AUC* = *Area Under the ROC Curve*.

Performance evaluation $\log_2$ FC, BH-adjusted p -value, AUC, sensitivity and specificity in the discovery RNA-Seq dataset of the 6 candidate biomarker genes are listed in Table 3.2, and their boxplots are shown in Figure 3.5.

Single candidate biomarker genes						
Gene Name	Ensembl ID	$\log_2$ FC	BH-adjusted p -value	AUC	Sensitivity	Specificity
GBP5	ENSG00000154451	1.60	1.4×10^{-56}	86.4%	75.6%	81.8%
CDKN1C	ENSG00000129757	-0.79	1.4×10^{-11}	66.9%	71.5%	57.6%
CPVL	ENSG00000106066	0.53	6.3×10^{-15}	69.8%	69.3%	62.1%
CYB561	ENSG00000008283	0.80	9.6×10^{-41}	82.7%	80.0%	69.9%
RSAD2	ENSG00000134321	0.69	6.3×10^{-05}	58.4%	91.9%	29.8%
EFCAB2	ENSG00000203666	0.57	3.2×10^{-16}	70.1%	65.6%	67.4%
Combined 6-gene signature						
GBP5, CDKN1C, CPVL, CYB561, RSAD2, EFCAB2				91.2%	80.0%	92.2%

Table 3.2: $\log_2$ FC, BH-adjusted p -value, AUC, sensitivity and specificity of the 6 candidate gene biomarkers (individually and combined into a signature) identified in the discovery RNA-Seq dataset when contrasting microbiologically confirmed TB vs. unlikely TB.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve; RNA-Seq = RNA Sequencing

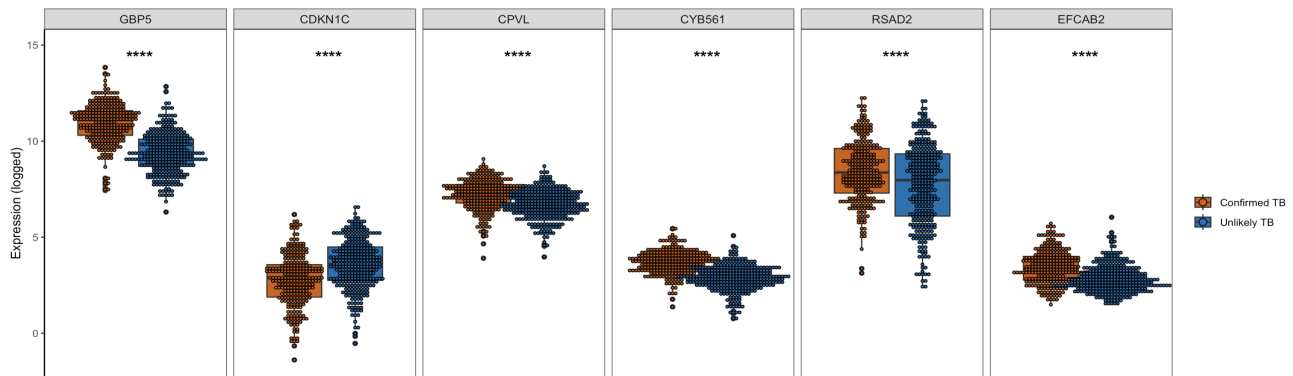


Figure 3.5: Logged expression values of the 6 candidate biomarker genes identified in the discovery RNA-Seq dataset coloured according to disease group.

ns: p -value > 0.05 ; *: p -value < 0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . All p -values are adjusted for multiple testing using Benjamini-Hochberg. RNA-Seq = RNA Sequencing

The output model of FS-PLS was retrained using logistic regression for two main reasons: logistic regression is more suitable for a binary outcome; the cross-validation step selected the final signature 8 times with different weights/coefficients for each of the 8 models, retraining was therefore needed to have final weights/coefficients. The performance of the final signature was evaluated in terms of AUC, using a wDRS for each patient [Section 2.7], sensitivity and specificity. The wDRS was calculated by multiplying the logged (base 2) gene expression of each of the 6 genes for each patient in the final signature by their weights/coefficients obtained from the logistic regression model. The wDRS was then used to calculate the performance of the signature. The 6-gene signature achieved an AUC of 91.2% [CI_{95%} 88.7%-93.6%], sensitivity of 80.0% [CI_{95%} 75.2%-84.8%] and specificity of 92.2% [CI_{95%} 89.0%-95.0%] in the discovery RNA-Seq dataset [Figure 3.6].

Table 3.3 describes the performance of the signature when stratifying by HIV status, site of TB and cohort. The signature achieved a better performance in the HIV-uninfected patients and patients with any forms of EPTB, achieving the best performance in children diagnosed with PTB and EPTB together. The lowest performance was achieved in the Malawian and Gambian cohorts.

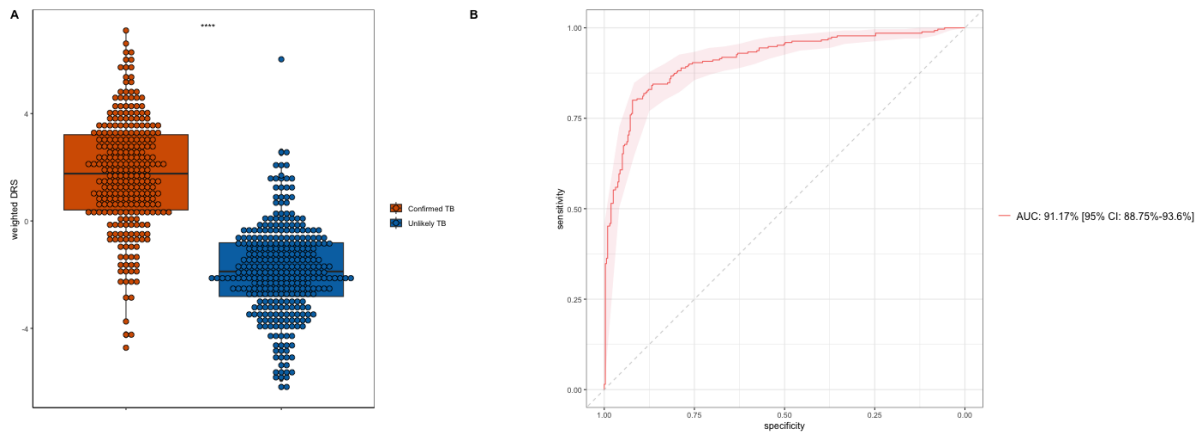


Figure 3.6: Boxplot of the wDRS of the 6-gene signature and its ROC curve in the discovery RNA-Seq cohort. ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . wDRS = weighted Disease Risk Score; ROC = Receiver Operating Characteristic; RNA-Seq = RNA Sequencing; AUC = Area Under the ROC Curve.

wDRS			
	AUC [CI _{95%}]	Sensitivity [CI _{95%}]	Specificity [CI _{95%}]
HIV-uninfected	92.1% [CI _{95%} 89.4%-94.9%]	86.1% [CI _{95%} 81.3%-90.9%]	89.9% [CI _{95%} 85.7%-93.6%]
HIV-infected	87.1% [CI _{95%} 81.4%-92.8%]	83.9% [CI _{95%} 74.2%-91.9%]	77.5% [CI _{95%} 69.6%-85.3%]
PTB	88.8% [CI _{95%} 85.0%-92.5%]	75.8% [CI _{95%} 68.0%-83.6%]	92.2% [CI _{95%} 89.0%-95.0%]
EPTB	92.0% [CI _{95%} 87.7%-96.4%]	87.0% [CI _{95%} 73.1%-100.0%]	87.8% [CI _{95%} 84.0%-91.2%]
PTB+EPTB	94.9% [CI _{95%} 92.7%-97.1%]	88.2% [CI _{95%} 80.3%-94.7%]	88.4% [CI _{95%} 84.6%-91.9%]
South Africa ILULU	95.0% [CI _{95%} 90.1%-100.0%]	89.7% [CI _{95%} 75.9%-100.0%]	95.4% [CI _{95%} 88.4%-100.0%]
Malawi ILULU	78.9% [CI _{95%} 69.3%-88.4%]	71.9% [CI _{95%} 56.3%-87.5%]	82.1% [CI _{95%} 73.8%-89.3%]
South Africa NIH	93.1% [CI _{95%} 89.8%-96.4%]	84.4% [CI _{95%} 78.8%-89.4%]	91.7% [CI _{95%} 85.4%-96.9%]
Kenya	95.4% [CI _{95%} 90.5%-100.0%]	91.2% [CI _{95%} 79.4%-100.0%]	92.9% [CI _{95%} 85.7%-98.2%]
The Gambia	82.2 % [CI _{95%} 66.4%-97.9%]	73.3% [CI _{95%} 46.7%-93.3%]	92.5% [CI _{95%} 82.5%-100.0%]

Table 3.3: Performance of the transcriptomic signature in the discovery RNA-Seq cohort broken down by clinical subgroups.

CI = Confidence Interval; HIV = Human Immunodeficiency Virus; PTB = Pulmonary Tuberculosis; EPTB = Extra-Pulmonary Tuberculosis

The 6 candidate biomarker genes were taken forward to cross-platform validation and, subsequently, to independent validation.

3.3.2 Exon analysis and RT-qPCR assay optimisation

RNA-Seq captures the expression of a given gene using its introns (i.e., noncoding sections of an RNA transcript) and exons (i.e., coding sections of an RNA transcript). Exons are transformed into cDNA, which is the starting material for a PCR reaction. Thus, it was critical to explore the coding regions of the candidate biomarker genes for identifying the assay targeting the gene region that translated best across platforms. Two optimal exons for each candidate biomarker were selected based on the following parameters in the RNA-Seq data: $\log_2$ FC; BH-adjusted p -value; AUC; mean exon counts in the two groups of interest (i.e., microbiologically confirmed TB and unlikely TB); exon length; and Guanine-Cytosine (GC) content. The GC content is the percentage of guanine and cytosine in a DNA or RNA molecule. GC base pairs are held together by three hydrogen bonds, compared to adenine and thymine/uracil which are held together by two hydrogen bonds. The stronger hydrogen bonding between guanine and cytosine helps with the stability of the RT-qPCR primers. A good value of GC content for stable RT-qPCR primers is 40%-60%.

48 samples, previously included in the discovery phase, were used to quantify the two optimal exons for each candidate biomarker and three reference genes. The following disease groups were included in the assay optimisation phase: microbiologically confirmed TB ($n=18$), unlikely TB ($n=17$) and *Mtb* infection ($n=13$). Patients were HIV-uninfected or HIV-infected and included children recruited in all the cohorts used in the discovery RNA-Seq dataset. The final exons, listed in Table 3.4, were finally selected based on their correlation with the discovery RNA-Seq data.

Gene Name	Exon Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value	AUC	Mean counts confirmed TB	Mean counts unlikely TB	GC content	Length
<i>GBP5</i>	ENSE00003681108	1.66	1.7×10^{-60}	87.1%	1607.4	498.5	51.2%	281
<i>CDKN1C</i>	ENSE00001486037	-0.32	3.1×10^{-03}	56.1%	0.6	2.6	72.4%	116
<i>CPVL</i>	ENSE00000832149	0.59	2.6×10^{-14}	69.6%	178.2	106.3	49.2%	183
<i>CYB561</i>	ENSE00001513794	0.52	1.3×10^{-06}	63.8%	1.3	0.79	63.8%	113
<i>RSAD2</i>	ENSE00001956769	0.57	8.8×10^{-10}	62.3%	106.7	82.2	36.3%	529
<i>EFCAB2</i>	ENSE00003640672	0.93	5.0×10^{-14}	71.1%	10.6	5.2	45.2%	73

Table 3.4: log₂FC, BH-adjusted *p*-value, AUC, mean exon counts in the confirmed TB and unlikely TB groups, GC content and exon length of the 6 exons selected for each of the candidate biomarker genes identified in the discovery RNA-Seq dataset when contrasting microbiologically confirmed TB vs. unlikely TB.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve; GC = Guanine-Cytosine

Each of the selected assays was normalised using the 3 reference genes [Section 2.6], and log₂FC and BH-adjusted *p*-value were calculated when contrasting microbiologically confirmed TB vs. unlikely TB using a linear model accounting for sex and age. Table 3.5 outlines log₂FC, BH-adjusted *p*-value, AUC, sensitivity and specificity of the assays selected for validation of the 6 candidate biomarker genes using RT-qPCR. 2 genes (*CDKN1C* and *CPVL*) out of 6 did not have a BH-adjusted *p*-value <0.05 probably because of the small sample size ($n=48$).

Gene Name	log ₂ FC	BH-adjusted <i>p</i> -value	AUC	Sensitivity	Specificity
<i>GBP5</i>	2.17	8.5×10^{-04}	85.8%	76.5%	88.2%
<i>CDKN1C</i>	-0.68	0.37	60.6%	82.4%	52.9%
<i>CPVL</i>	0.84	0.13	66.4%	41.2%	100%
<i>CYB561</i>	1.57	6.6×10^{-03}	79.6%	76.5%	76.5%
<i>RSAD2</i>	1.14	6.3×10^{-02}	69.6%	58.8%	94.1%
<i>EFCAB2</i>	1.19	6.3×10^{-02}	71.3%	64.7%	76.5%

Table 3.5: log₂FC, BH-adjusted *p*-value, AUC, sensitivity and specificity of the 6 candidate biomarker genes measured using RT-qPCR in the assay optimisation phase.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve; RT-qPCR = Reverse Transcription - quantitative Polymerase Chain Reaction

The assays were finally used to quantify the 6 candidate gene biomarkers in the discovery cohort by RT-qPCR.

3.3.3 Cross-platform validation phase

Cross-platform validation is needed to assess the transferability of the findings across platforms. The aim of this study was to identify candidate biomarker genes for childhood TB to be translated into a rapid and non-sputum-based POCT. Rapid transcriptomic POCT are usually PCR based, therefore it is critical to assure that any signature identified in an RNA-Seq dataset has a similar performance in a PCR based platform. Some of the genes previously identified as candidate biomarkers for childhood TB failed to translate across platforms and could not be used to develop a POCT. This might occur

when the discovery dataset contains latent factors, such as batch effect or other technical sources of variability that drive the selection of biomarkers or might be due to lowly expressed genes which make them difficult to target in low-throughput quantification methods. For these reasons the cross-platform validation step is critical, as well as the validation in independent cohorts to ensure that the model in the discovery cohort has not been overfitted.

Six candidate biomarker genes (*GBP5*, *CDKN1C*, *CPVL*, *CYB561*, *RSAD2* and *EFCAB2*) in addition to three reference genes (*GAPDH*, *ACTB* and *RNAseP*) were taken forward to cross-platform validation using RT-qPCR.

Data pre-processing and normalisation The assays selected for each candidate biomarker in the optimisation phase were measured in the whole discovery cohort after excluding the samples there was no RNA available for [Table 3.1]. The 6 candidate biomarker genes were run in quadruplicates (i.e., 4 replicates) using RT-qPCR and were measured on samples from patients with microbiologically confirmed TB ($n=204$) and unlikely TB ($n=231$) across 6 RT-qPCR plates.

Since a 40-cycle RT-qPCR experiment was carried out for validating the candidate gene biomarkers, C_t values of 40 (i.e., no or very low gene expression) were assigned a *NULL* value. For each sample, the standard deviation of each gene across replicates was calculated. Genes with standard deviation ≤ 2 across replicates were averaged for each sample. If for a given sample and a given gene, the standard deviation across replicates was >2 , the C_t values were manually examined and either set to *NULL* or the mean across a subset of replicates was calculated if they had a standard deviation ≤ 2 . Within each plate, each assay was normalised using the 3 reference genes following the ΔC_t normalisation [Section 2.6]. Normalised data from the 6 RT-qPCR plates were then merged and PCA performed to identify any sources of technical variability. There was no plate effect captured by the first 4 PCs after normalisation [Figure 3.7].

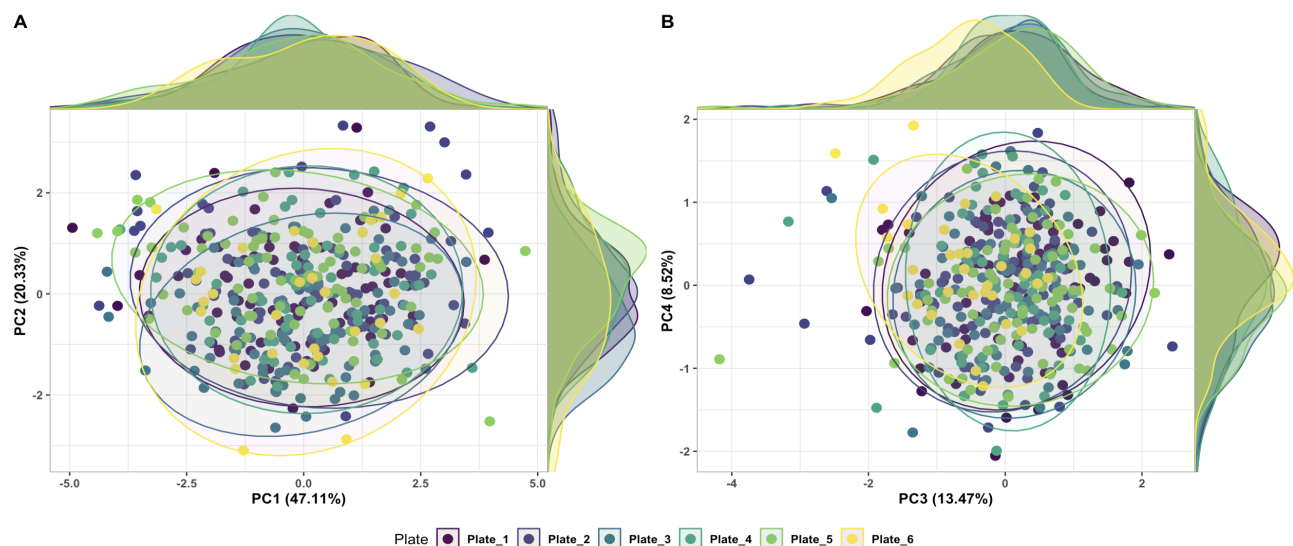


Figure 3.7: PCA biplots from PCA performed on the cross-platform validation RT-qPCR dataset coloured according to sample plate.

PC 1-2 [Panel A] and PC 3-4 [Panel B] are shown. Points represent samples and are coloured according to the sample plate. The distributions of PC1, PC2, PC3 and PC4 are shown in the density plots. *PCA* = Principal Component Analysis; *RT-qPCR* = Reverse Transcription - quantitative Polymerase Chain Reaction; *PC* = Principal Component

No other technical variability was captured by the first 4 PCs [Figure 3.8]. The disease group was

captured by PC1 [Figure 3.8; Panel A] with the microbiologically confirmed TB group towards the right of PC1 and the unlikely TB group towards the top left of the plot. HIV status was captured by PC1 and PC2 with the HIV-uninfected group towards the right of PC1 [Figure 3.8; Panel C]. The recruitment site was captured by PC1 and PC2 with the ILULU Cape Town and ILULU Malawi cohorts towards the bottom left of the PC1 vs. PC2 biplot and Kenya, NIH Cape Town and The Gambia cohorts towards the top right of the plot [Figure 3.8; Panels E, F]. No differences were observed in terms of sex or age [Figure 3.8; Panels G, H, I, J].

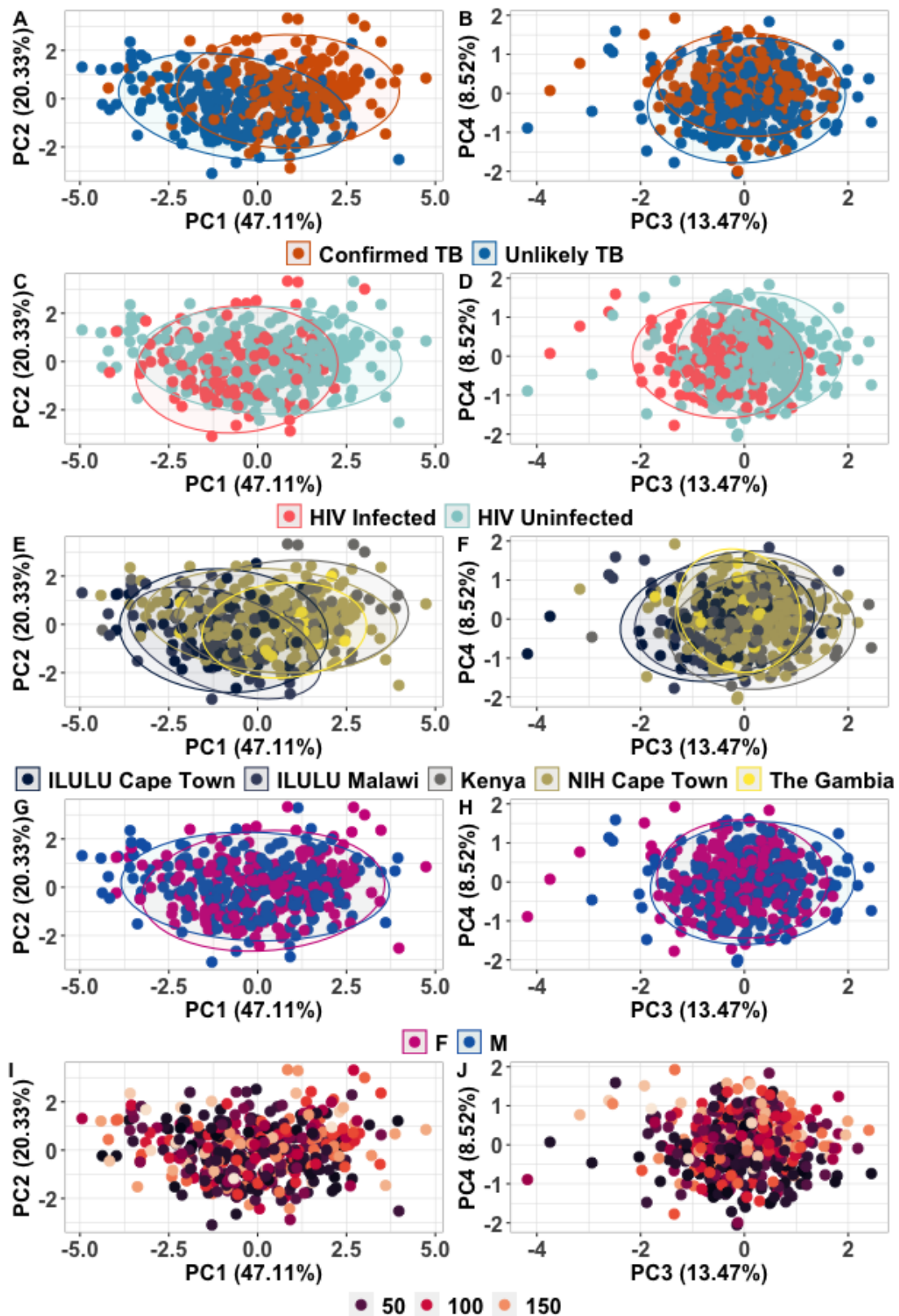


Figure 3.8: PCA biplots from PCA performed on the cross-platform validation RT-qPCR dataset coloured according to clinical variables.

PC 1-2 and PC 3-4 are shown. Points represent samples and are coloured according to disease group [Panel A and Panel B], HIV status [Panel C and Panel D], recruitment site [Panel E and Panel F], sex [Panel G and Panel H] and age [Panel I and Panel J]. PCA = Principal Component Analysis; RT-qPCR = Reverse Transcription - quantitative Polymerase Chain Reaction; PC = Principal Component; HIV = Human Immunodeficiency Virus; F = female; M = male

Performance evaluation $\log_2FC$ and BH-adjusted p -value were calculated on the normalised data using a linear model accounting for sex, age and sample plate when contrasting microbiologically confirmed TB and unlikely TB [Table 3.6].

Single candidate biomarker genes					
Gene Name	$\log_2FC$	BH-adjusted p -value	AUC	Sensitivity	Specificity
<i>GBP5</i>	1.46	1.2×10^{-29}	82.5%	81.4%	72.5%
<i>CDKN1C</i>	-0.99	1.4×10^{-09}	67.7%	66.8%	63.2%
<i>CPVL</i>	0.63	2.8×10^{-09}	65.9%	70.1%	57.8%
<i>CYB561</i>	0.57	7.0×10^{-08}	67.5%	55.4%	73.5%
<i>RSAD2</i>	0.53	9.7×10^{-05}	60.1%	39.8%	79.4%
<i>EFCAB2</i>	0.68	1.7×10^{-13}	71.8%	70.6%	64.7%
Combined 6-gene signature - Retrained wDRS					
<i>GBP5, CDKN1C, CPVL, CYB561, RSAD2, EFCAB2</i>			87.5%	81.9%	80.5%

Table 3.6: $\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity of the 6 candidate biomarker genes (individually and combined into a signature) measured using RT-qPCR in the cross-platform validation phase.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve; $RT-qPCR$ = Reverse Transcription - quantitative Polymerase Chain Reaction

Figure 3.9 shows the boxplots of the 6 candidate biomarker genes measured using RT-qPCR in the discovery cohort. All the genes were SDE between the two disease groups of interest.

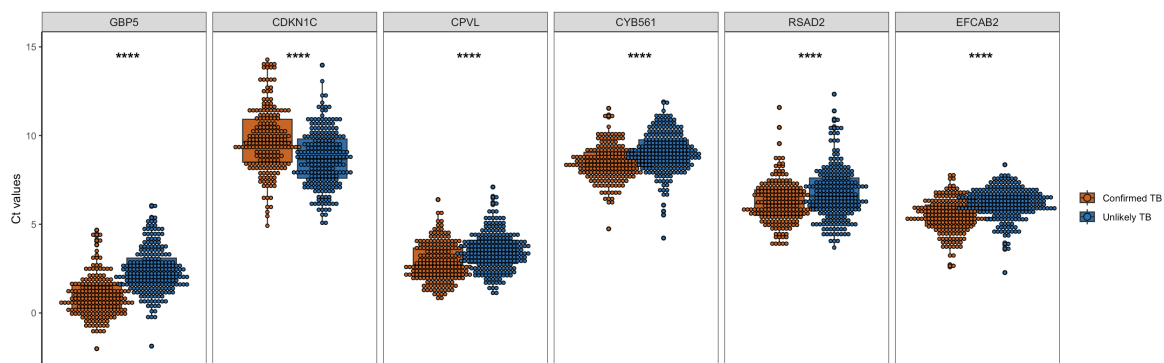


Figure 3.9: C_t values of the 6 candidate biomarker genes measured in the cross-platform RT-qPCR dataset coloured according to disease group.

All p -values are adjusted for multiple testing using Benjamini-Hochberg. ns: p -value > 0.05 ; *: p -value < 0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . C_t = Cycle Threshold; $RT-qPCR$ = Reverse Transcription - quantitative Polymerase Chain Reaction

The 6 genes were then combined into a wDRS using the weights/coefficients from the RNA-Seq statistical model, the weights/coefficients from a new logistic regression model retrained on the RT-qPCR dataset and into a simple DRS [Section 2.7]. Since the gene counts are inversely proportional to C_t values, the coefficients of the wDRS identified in the RNA-Seq (i.e., original wDRS) and the coefficients of the wDRS optimised for a RT-qPCR based platform (i.e., retrained wDRS), have opposite direction when comparing microbiologically confirmed TB vs. unlikely TB [Figure 3.10, Panel A]. An inverse DRS was also used (i.e., assigning -1 to the over-expressed genes and +1 to the under-expressed genes). The performance of the signature in terms of AUC, sensitivity and specificity using the weights/coefficients from the RNA-Seq statistical model were respectively 85.6% [$CI_{95\%}$ 82.0%-89.1%], 71.1% [$CI_{95\%}$ 65.2%-77.0%] and 86.2% [$CI_{95\%}$ 81.4%-90.5%]. When logistic regression was used to retrain the model in the RT-qPCR data and the new weights/coefficients were used, the performance of the signature in terms of AUC, sensitivity and specificity were respectively 87.5% [$CI_{95\%}$ 84.1%-90.8%],

81.9% [CI_{95%} 76.5%-87.3%] and 80.5% [CI_{95%} 75.3%-85.3%]. After combining the genes based on a simple DRS, the performance in terms of AUC, sensitivity and specificity were respectively 80.9% [CI_{95%} 76.9%-85.0%], 80.4% [CI_{95%} 75.0%-85.8%] and 68.8% [CI_{95%} 62.3%-74.9%]. The boxplots and ROC curves of the three models are shown in Figure 3.10.

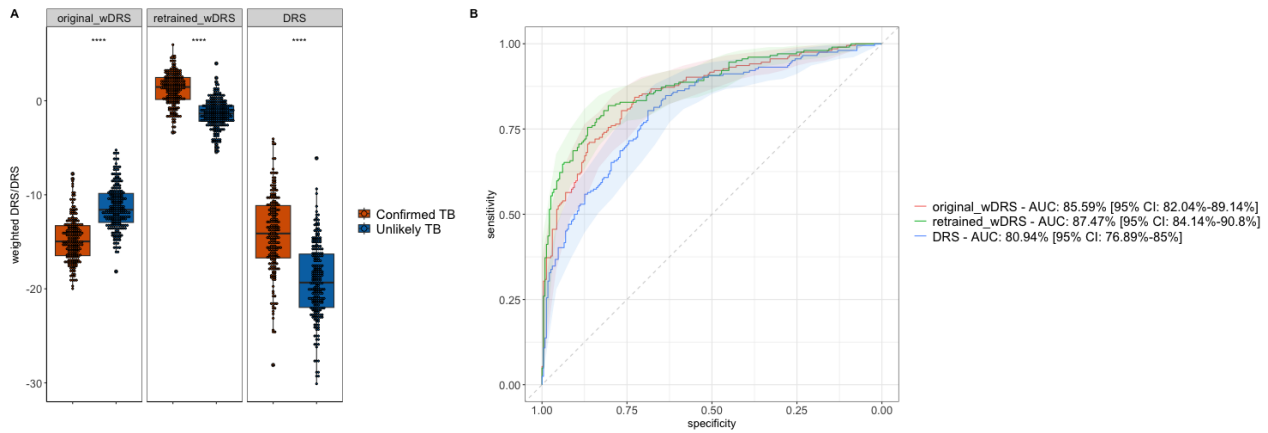


Figure 3.10: Boxplots of the original wDRS (i.e., statistical model trained on RNA-Seq data), retrained wDRS (i.e., statistical model retrained on RT-qPCR data) and DRS of the 6-gene signature and the correspondent ROC curves in the cross-platform RT-qPCR dataset.

ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . wDRS = *weighted Disease Risk Score*; DRS = *Disease Risk Score*; ROC = *Receiver Operating Characteristic*; RT-qPCR = *Reverse Transcription - quantitative Polymerase Chain Reaction*; AUC = *Area Under the ROC curve*

The 6-gene signature translated across platforms with an excellent performance in distinguishing microbiologically confirmed TB disease vs. unlikely TB in children when RT-qPCR was used. The cross-platform validation phase was used in this study not only to assess the transferability of the findings across platforms but also to retrain the model containing the 6 candidate biomarker genes and calculate new weights/coefficients ideal for a PCR based platform. C_t values and gene counts are, in fact, generated following different pipelines respectively in RT-qPCR and RNA-Seq, therefore weights/coefficients calculated on an RNA-Seq dataset might not be ideal for RT-qPCR. The new weights/coefficients calculated in the cross-platform validation phase were then used in the independent validation phase.

3.3.4 Independent validation phase

Validation in independent cohorts is needed to assess the transferability of the findings, the performance of the signature in samples not used in the discovery phase and ensure that the model was not overfitted.

Data pre-processing and normalisation The pipelines for the RT-qPCR laboratory experiment and the data analysis were identical to the ones followed in the cross-platform validation [Section 3.3.3]. In brief, the 6 candidate biomarker genes were run in quadruplicates and measured on samples taken from patients with microbiologically confirmed TB ($n=52$), unconfirmed TB ($n=63$) and unlikely TB ($n=57$) across 5 RT-qPCR plates.

For each sample and each gene, the quadruplicates were averaged if their standard deviation was ≤ 2 . If the standard deviation across replicates was >2 , the C_t values were manually examined and either set to *NULL* or the mean across a subset of replicates was calculated if they had a standard deviation ≤ 2 . Within each plate, each assay was normalised using the 3 reference genes following the ΔC_t

normalisation [Section 2.6]. Normalised data from the 5 RT-qPCR plates were finally merged and PCA was performed to identify any source of technical variability. There was no plate effect captured by the first 4 PCs [Figure 3.11].

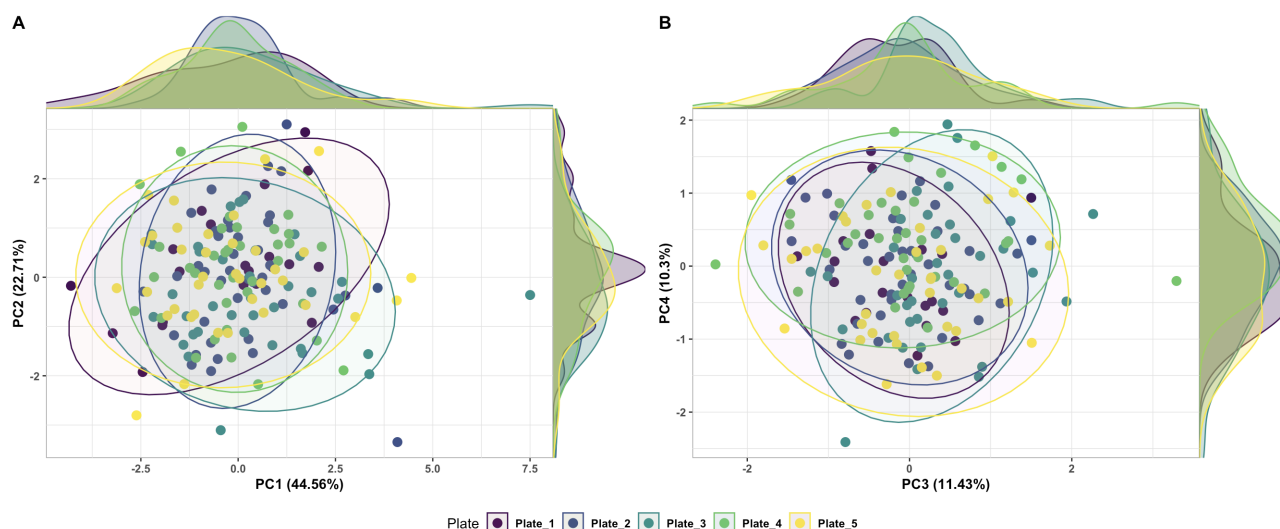


Figure 3.11: PCA biplots from PCA performed on the independent validation RT-qPCR dataset coloured according to sample plate.

PC 1-2 [Panel A] and PC 3-4 [Panel B] are shown. Points represent samples and are coloured according to the sample plate. The distributions of PC1, PC2, PC3 and PC4 are shown in the density plots. *PCA* = *Principal Component Analysis*; *RT-qPCR* = *Reverse Transcription - quantitative Polymerase Chain Reaction*; *PC* = *Principal Component*

No other sources of technical variation were captured by the first PCs [Figure 3.12]. The disease group was captured by PC1 [Figure 3.12; Panel A] with the microbiologically confirmed TB group towards the bottom left of the plot and the unlikely TB group towards the top right of the plot. The unconfirmed TB group was in between the microbiologically confirmed TB and unlikely TB groups. No difference was captured by the first 4 PCs in terms of HIV status, recruitment site, sex or age [Figure 3.12; Panels C, D, E, F, G, H, I, J].

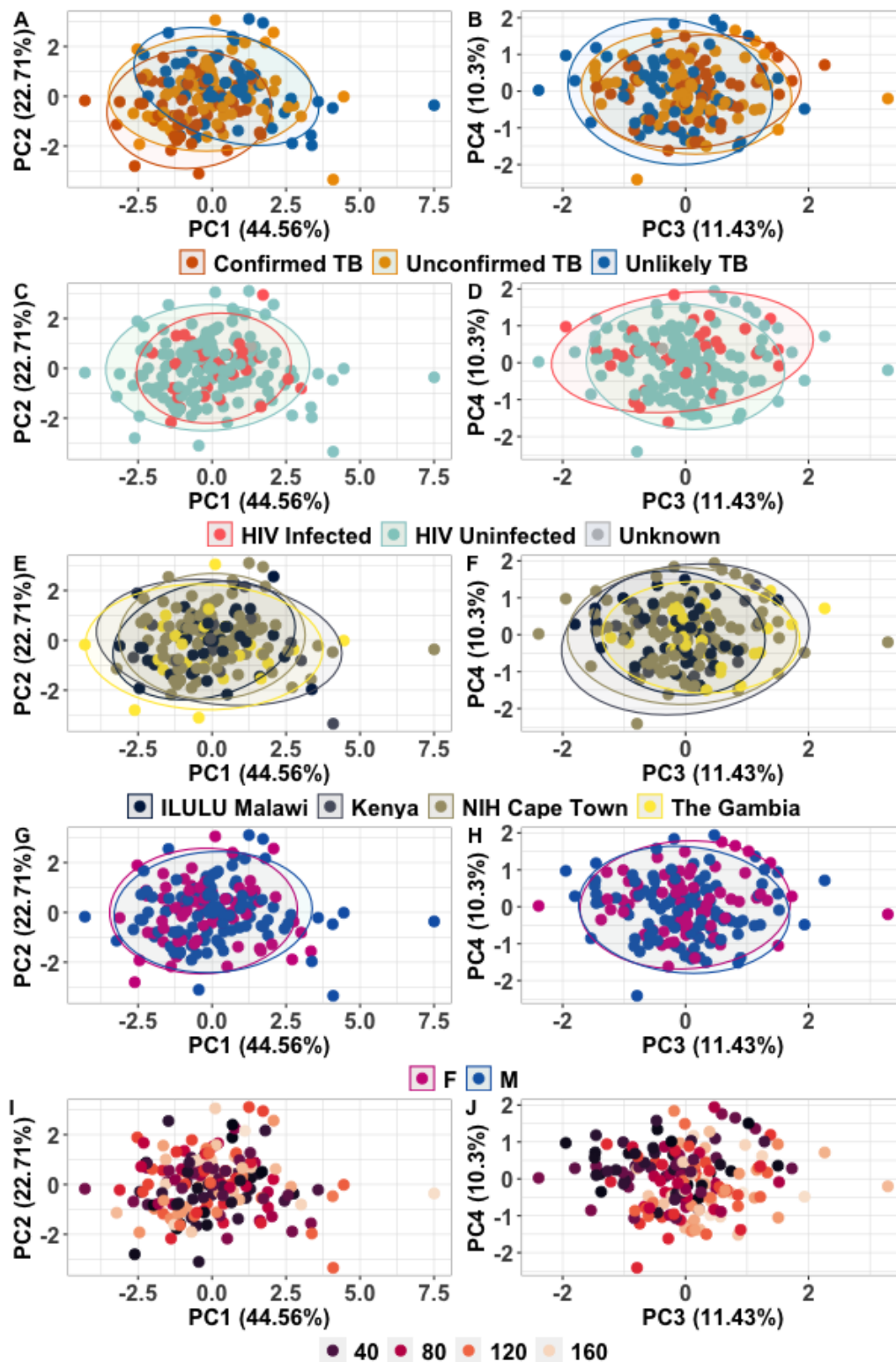


Figure 3.12: PCA biplots from PCA performed on the independent validation RT-qPCR dataset coloured according to clinical variables.

PC 1-2 and PC 3-4 are shown. Points represent samples and are coloured according to disease group [Panel A and Panel B], HIV status [Panel C and Panel D], recruitment site [Panel E and Panel F], sex [Panel G and Panel H] and age [Panel I and Panel J]. PCA = Principal Component Analysis; RT-qPCR = Reverse Transcription - quantitative Polymerase Chain Reaction; PC = Principal Component; HIV = Human Immunodeficiency Virus; F = female; M = male

Single candidate biomarker genes					
Gene Name	$\log_2FC$	BH-adjusted p -value	AUC	Sensitivity	Specificity
<i>GBP5</i>	1.57	5.2×10^{-09}	87.9%	85.9%	80.8%
<i>CDKN1C</i>	-0.80	1.8×10^{-02}	65.0%	75.4%	57.7%
<i>CPVL</i>	0.49	1.9×10^{-02}	61.6%	35.1%	94.2%
<i>CYB561</i>	0.06	0.76	53.1%	21.1%	94.2%
<i>RSAD2</i>	0.73	7.1×10^{-03}	64.3%	66.7%	63.5%
<i>EFCAB2</i>	0.99	4.5×10^{-08}	78.8%	64.9%	78.8%
Combined 6-gene signature - wDRS					
<i>GBP5, CDKN1C, CPVL, CYB561, RSAD2, EFCAB2</i>			89.3%	88.5%	86.0%

Table 3.7: $\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity of the 6 candidate biomarker genes (individually and combined into a signature) measured using RT-qPCR in the independent validation phase. FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve

Performance evaluation $\log_2FC$ and BH-adjusted p -value were calculated using a linear model accounting for sex and age when contrasting microbiologically confirmed TB vs. unlikely TB [Table 3.7].

Figure 3.13 shows the boxplots of the 6 candidate biomarker genes measured using RT-qPCR in the independent validation cohort.

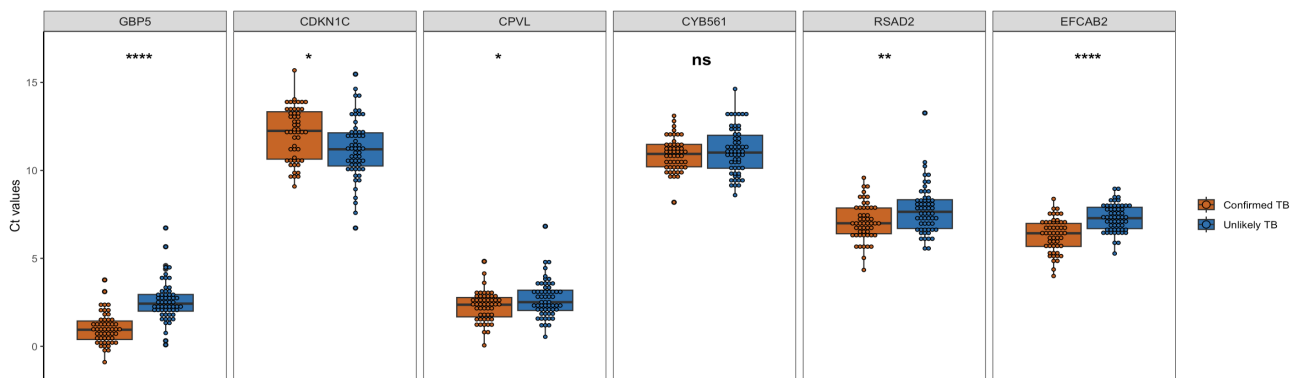


Figure 3.13: C_t values of the 6 candidate biomarker genes measured in the independent RT-qPCR dataset coloured according to disease group.

All p -values are adjusted for multiple testing using BH. ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . C_t = Cycle Threshold; RT-qPCR = Reverse Transcription - quantitative Polymerase Chain Reaction; BH = Benjamini-Hochberg

5 candidate biomarker genes were SDE when measured using RT-qPCR in a new set of samples. *CYB561* was the only gene to be not different between the two disease groups of interest in terms of $\log_2FC$ and BH-adjusted p -value. The 6 genes were then combined into a wDRS using the weights/coefficients from the logistic regression model previously retrained on the cross-platform RT-qPCR dataset and into a simple inverse DRS (i.e., assigning -1 to the over-expressed genes and +1 to the under-expressed genes) [Section 2.7]. The performance of the signature in terms of AUC, sensitivity and specificity using the weights/coefficients optimised for RT-qPCR data were 89.3% [$CI_{95\%}$ 82.5%-96.1%], 88.5% [$CI_{95\%}$ 78.9%-96.2%] and 86.0% [$CI_{95\%}$ 75.4%-94.7%] respectively. When the specificity was fixed at 70%, the sensitivity achieved 90.4% [$CI_{95\%}$ 80.8%-96.2%], meeting the minimal WHO requirements for sensitivity and specificity for a biomarker-based triage test for TB in children [44]. After combining the genes based on a simple inverse DRS, the performance in terms of AUC, sensitivity and specificity were 84.1% [$CI_{95\%}$ 76.9%-91.4%], 75.0% [$CI_{95\%}$ 63.5%-86.5%] and 80.7% [$CI_{95\%}$ 70.2%-91.2%] respectively. The boxplots and ROC curves of the two models are shown in Figure 3.14. When wDRS

and DRS were calculated in the unconfirmed TB group, the median was in between the microbiologically confirmed TB and unlikely TB groups [Figure 3.14 Panel A].

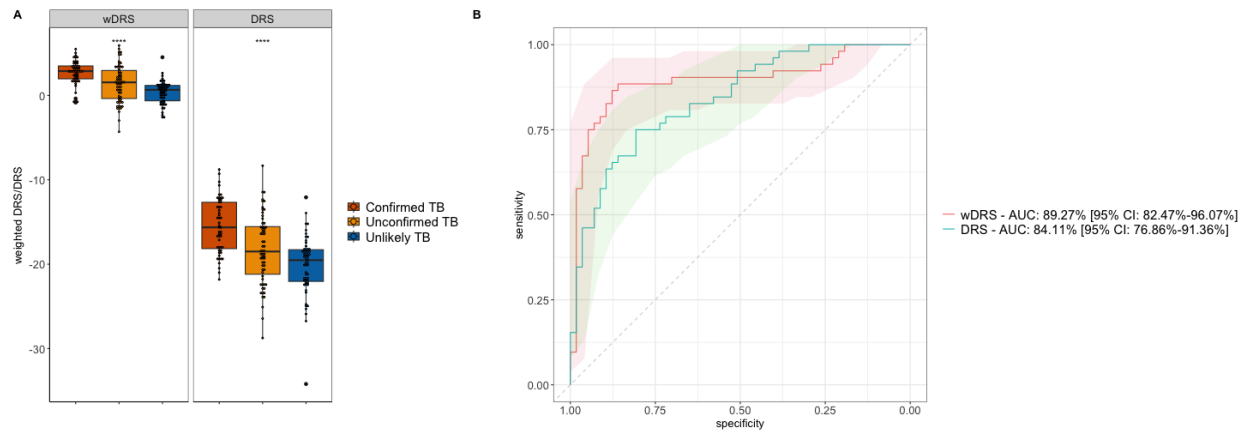


Figure 3.14: Boxplots of the wDRS and DRS of the 6-gene signature and the correspondent ROC curves in the independent RT-qPCR dataset.

ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . wDRS = weighted Disease Risk Score; DRS = Disease Risk Score; RT-qPCR = Reverse Transcription - quantitative Polymerase Chain Reaction; AUC = Area Under the ROC Curve

Table 3.8 shows the performance of the signature when stratifying by HIV status, site of TB infection and cohort. The signature achieved a higher performance in the HIV-uninfected and EPTB cases. Kenyan and Gambian cohorts achieved lower performance compared to the South African and Malawian cohorts. The lower performance in the Gambian cohort reflects the results of the discovery RNA-Seq cohort. The lower performance in the Kenyan cohort is probably due to its small sample size (South Africa: $n=103$; Malawi: $n=46$; Kenya: $n=18$; The Gambia: $n=38$), where the only one misclassification causes a drop in terms of AUC and sensitivity.

Original wDRS			
	AUC [CI _{95%}]	Sensitivity [CI _{95%}]	Specificity [CI _{95%}]
HIV-uninfected	89.1% [CI _{95%} 81.4%-96.8%]	87.2% [CI _{95%} 76.9%-97.4%]	86.4% [CI _{95%} 75.0%-95.5%]
HIV-infected	87.2% [CI _{95%} 69.8%-100.0%]	92.3% [CI _{95%} 76.9%-100.0%]	91.7% [CI _{95%} 75.0%-100.0%]
PTB	87.2% [CI _{95%} 78.6%-95.8%]	85.3% [CI _{95%} 73.5%-97.1%]	86.0% [CI _{95%} 77.2%-94.7%]
PTB+EPTB	98.1% [CI _{95%} 95.2%-100.0%]	100.0%	94.7% [CI _{95%} 89.4%-100.0%]
South Africa	94.3% [CI _{95%} 87.8%-100.0%]	96.9% [CI _{95%} 90.6%-100.0%]	87.7% [CI _{95%} 75.0%-96.9%]
Malawi	95.0% [CI _{95%} 81.1%-100.0%]	100.0%	80.0% [CI _{95%} 40.0%-100.0%]
Kenya	75.0% [CI _{95%} 26.0%-100.0%]	75.0% [CI _{95%} 25.0%-100.0%]	100.0%
The Gambia	74.4% [CI _{95%} 51.3%-97.4%]	75.0% [CI _{95%} 50.0%-100.0%]	84.6% [CI _{95%} 61.5%-100.0%]
DRS			
	AUC [CI _{95%}]	Sensitivity [CI _{95%}]	Specificity [CI _{95%}]
HIV-uninfected	85.6% [CI _{95%} 77.7%-93.5%]	71.8% [CI _{95%} 56.4%-84.6%]	86.4% [CI _{95%} 75.0%-95.5%]
HIV-infected	77.6% [CI _{95%} 58.0%-97.1%]	69.2% [CI _{95%} 46.0%-92.3%]	83.3% [CI _{95%} 58.3%-100.0%]
PTB	83.5% [CI _{95%} 75.2%-91.9%]	73.5% [CI _{95%} 58.8%-88.2%]	80.7% [CI _{95%} 70.2%-89.5%]
PTB+EPTB	85.3% [CI _{95%} 74.1%-96.6%]	78.6% [CI _{95%} 57.1%-100.0%]	80.7% [CI _{95%} 70.2%-89.5%]
South Africa	88.6% [CI _{95%} 80.4%-96.7%]	75.0% [CI _{95%} 59.4%-90.6%]	90.6% [CI _{95%} 78.1%-100.0%]
Malawi	100.0%	100.0%	100.0%
Kenya	78.6% [CI _{95%} 36.0%-100.0%]	75.0% [CI _{95%} 25.0%-100.0%]	100.0%
The Gambia	80.8% [CI _{95%} 61.4%-100.0%]	83.3% [CI _{95%} 58.3%-100.0%]	76.9% [CI _{95%} 53.9%-100.0%]

Table 3.8: Performance of the transcriptomic signature in clinical subgroups.

CI = Confidence Interval; HIV = Human Immunodeficiency Virus; PTB = Pulmonary Tuberculosis; EPTB = Extra-Pulmonary Tuberculosis

The results of the independent validation phase and the performance of the 6-gene signature are

very encouraging. They suggest that a biomarker-based POCT might display good performance in discriminating TB disease from unlikely TB in children by measuring the expression of only 6 genes.

3.3.5 Biology

Biological functions of the 6 genes The 6-gene signature identified in the RNA-Seq dataset and validated by RT-qPCR is composed of *GBP5* (Guanylate Binding Protein 5), *CDKN1C* (Cyclin Dependent Kinase Inhibitor 1C), *CPVL* (Carboxypeptidase Vitellogenic Like), *CYB561* (Cytochrome B561), *RSAD2* (Radical S-Adenosyl Methionine Domain Containing 2) and *EFCA2* (EF-Hand Calcium Binding Domain 2). These 6 genes have biological functions that might provide insights into the pathogenesis of childhood TB.

GBP5 is a protein-coding gene belonging to the TRAFAC class dynamin-like GTPase superfamily. The encoded protein is an activator of both AIM2 and NLRP3 inflammasome assembly and, therefore, has a role in innate immunity and inflammation. The NLRP3 inflammasome is a protein complex implicated in a wide range of diseases, including some infectious diseases, that triggers the release of proinflammatory cytokines [270]. The AIM2 inflammasome detects DNA in the cytosol and has a key role in the innate immune response against bacterial and viral infections [271]. Both AIM2 and NLRP3 inflammasome have a role in response to pathogenic bacteria [272, 273]. *GBP5* was identified as a potential biomarker for TB disease in previously published studies, for both adults [151, 154, 162, 274, 275] and children [137].

CDKN1C is a protein-coding gene belonging to cyclin-CDK inhibitors of the Cip/Kip family. The encoded protein is a negative regulator of cell proliferation, and it has been shown that, with the other CDK inhibitors, it plays an important role in cell transcription, apoptosis and migration [276]. Although the role of *CDKN1C* in TB pathogenesis was not extensively studied before, it was previously identified as a potential biomarker for childhood TB [137].

CPVL is a protein-coding gene. The encoded protein was first characterised in human macrophages [277] but its exact functions have not been characterised yet. It has been shown to be over-expressed during the maturation of monocytes to macrophages [278]. A study using immunohistochemical techniques on lung and lymph node tissues from South African HIV-uninfected children and adults showed that the abundance of *CPVL* was higher in children with TB disease compared to adults with TB and to the pediatric comparative group [279].

CYB561 is a protein-coding gene. The encoded protein is a transmembrane reductase involved in ascorbate (i.e., vitamin C) regeneration. Many studies have shown a possible inverse correlation between *Mtb* load and vitamin C in the host suggesting that vitamin C might help kill the bacteria [280, 281]. *CYB561* was previously identified as a potential biomarker for childhood TB [137].

RSAD2 is a protein-coding gene. The encoded protein is an interferon-inducible antiviral protein and is involved in cellular antiviral response and innate immune signalling. It has been shown that the encoded protein inhibits both DNA and RNA viruses, including HIV. It has been previously reported that *Mtb* might alter the expression of *RSAD2* [282].

EFCA2 is a protein-coding gene. The encoded protein contains two EF-hand calcium-binding domains, but its biological function is still unknown. Calcium (Ca^{2+}) plays a crucial role when an intracellular pathogen enters a host cell - its intracellular concentration changes as an indicator of manipulative

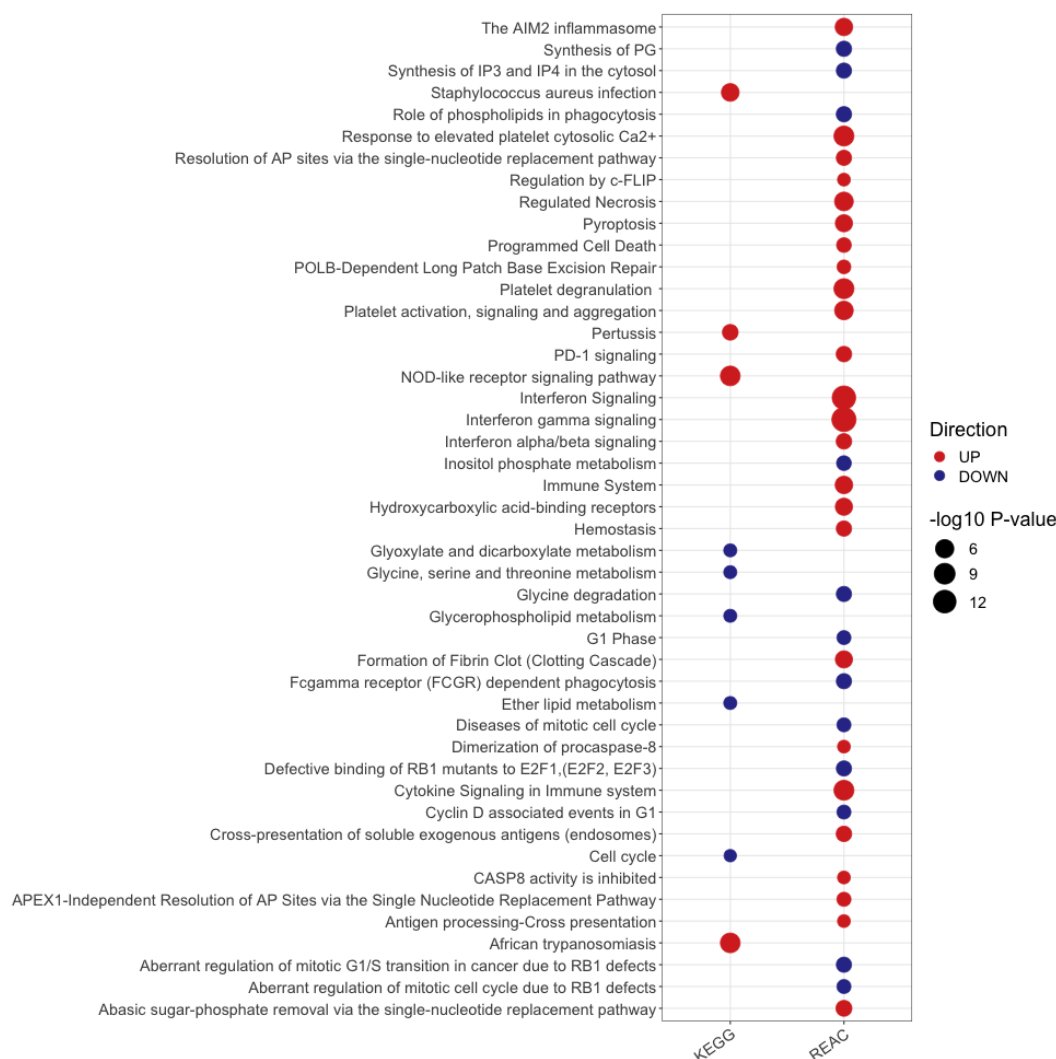


Figure 3.15: Pathways identified as significantly enriched from the list of SDE genes between microbiologically confirmed TB and unlikely TB in the discovery RNA-Seq dataset when querying the KEGG and Reactome databases.

Pathways were identified by g:Profiler2. Points represent pathways with colour indicating directionality and size representing BH-adjusted p -value (larger = lower p -value). *SDE* = Significantly Differentially Expressed; *RNA-Seq* = RNA Sequencing; *KEGG* = Kyoto Encyclopedia of Genes and Genomes

events. These alterations are generated by various calcium-binding molecules [283].

Pathway enrichment analysis Pathway enrichment analysis tools are usually used to give biological meaning to the results of high-throughput assays. Pathway enrichment analysis was performed on the 178 genes (7 under- and 171 over-expressed in TB disease) [Appendix B] identified as SDE (absolute $\log_2FC > 0.5$ and BH-adjusted p -value < 0.05) between microbiologically confirmed TB and unlikely TB in the discovery RNA-Seq dataset.

As outlined in Section 2.7, pathway analysis tools detect groups of related features that are altered and provide information about the biological processes that these features are associated with. gProfiler2 [258, 259] was used to query the GO, KEGG and Reactome databases using as input the 178 SDE genes. These 3 different databases were queried to have information in terms of biological processes (GO database, subontology Biological Process), gene functions (KEGG database) and reactions/proteins (Reactome database).

In the pathway enrichment analysis querying the KEGG and Reactome databases, pathways with p -value <0.05 were considered significantly enriched [Figure 3.15].

9 pathways (5 under- and 4 over-represented in TB disease) were identified as significantly enriched from the list of SDE genes between microbiologically confirmed TB and unlikely TB when querying the KEGG database. One of the over-represented and significantly enriched pathways was the NOD-like receptor signaling pathway (KEGG:04621; p -value = 5.2×10^{-04}). This pathway contains specific families of pattern recognition receptors which play a key role in recognising pathogens and mounting innate immune responses. Guanylate binding protein (GPB) genes, including one of the 6 genes identified as candidate biomarkers (*GBP5*) for childhood TB in this thesis, are part of this pathway, along with some interleukin, chemokine, defensin and stimulator of interferon response genes, all of which play a key role in the host immune response. The Cell cycle pathway (KEGG:04110; p -value = 4.9×10^{-02}) is under-represented in TB disease compared to unlikely TB. This pathway contains cyclin dependent kinase inhibitor genes, along with cell division cycle and transforming growth factor genes, which are involved in mitotic cell cycle progression, suggesting that DNA replication and cell mitosis are inhibited during the disease.

37 pathways (12 under- and 25 over-represented in TB disease) were identified as significantly enriched from the list of SDE genes between microbiologically confirmed TB and unlikely TB when querying the Reactome database. Pathways involving the host immune response, such as Cytokine Signaling in Immune system (R-HSA-1280215; p -value = 3.9×10^{-04}), Immune System (R-HSA-168256; p -value = 5.4×10^{-03}), Interferon alpha/beta signaling (R-HSA-909733; p -value = 2.1×10^{-02}), Interferon gamma signaling (R-HSA-877300; p -value = 4.3×10^{-07}), Interferon Signaling (R-HSA-913531; p -value = 1.6×10^{-06}), were all significantly over-represented in TB disease. The AIM2 inflammasome pathway (R-HSA-844615; p -value = 5.5×10^{-03}), which is activated by the encoded protein of *GBP5*, is also significantly enriched in TB disease.

In the pathway enrichment analysis querying the GO database (subontology: Biological Process), pathways with p -value <0.0001 were considered significantly enriched [Figure 3.16]. 71 pathways, all with GO terms over-represented in TB disease compared to unlikely TB, were significantly enriched.

Many GO terms significantly over-represented in TB disease were related to the host immune response, such as activation of immune response (GO:0002253; p -value = 4.4×10^{-05}), biological process involved in interspecies interaction between organisms (GO:0044419; p -value = 1.5×10^{-16}), cellular response to cytokine stimulus (GO:0071345; p -value = 1.0×10^{-08}), cellular response to interferon-gamma (GO:0071346; p -value = 1.4×10^{-09}), cellular response to tumor necrosis factor (GO:0071356; p -value = 4.7×10^{-06}), cytokine production (GO:0001816; p -value = 2.3×10^{-06}), defence response (GO:0006952; p -value = 4.3×10^{-15}), inflammatory response (GO:0006954; p -value = 1.1×10^{-07}), response to bacterium (GO:0009617; p -value = 6.4×10^{-08}), response to molecule of bacterial origin (GO:0002237; p -value = 3.7×10^{-06}), T cell activation (GO:0042110; p -value = 6.3×10^{-05}).

Although three different databases were queried to perform pathway enrichment analysis, all the results suggested that the gene expression profiles are very different between microbiologically confirmed TB and unlikely TB. During TB disease, there is an important host immune response, mainly involving IFN- γ and TNF- α , which play pivotal roles in protective immunity against intracellular pathogens including *Mtb*. The differences in gene expression profiles between the two groups of interest are highlighted in Figure 3.17, with the unlikely TB patients clustering towards the left-hand side of the

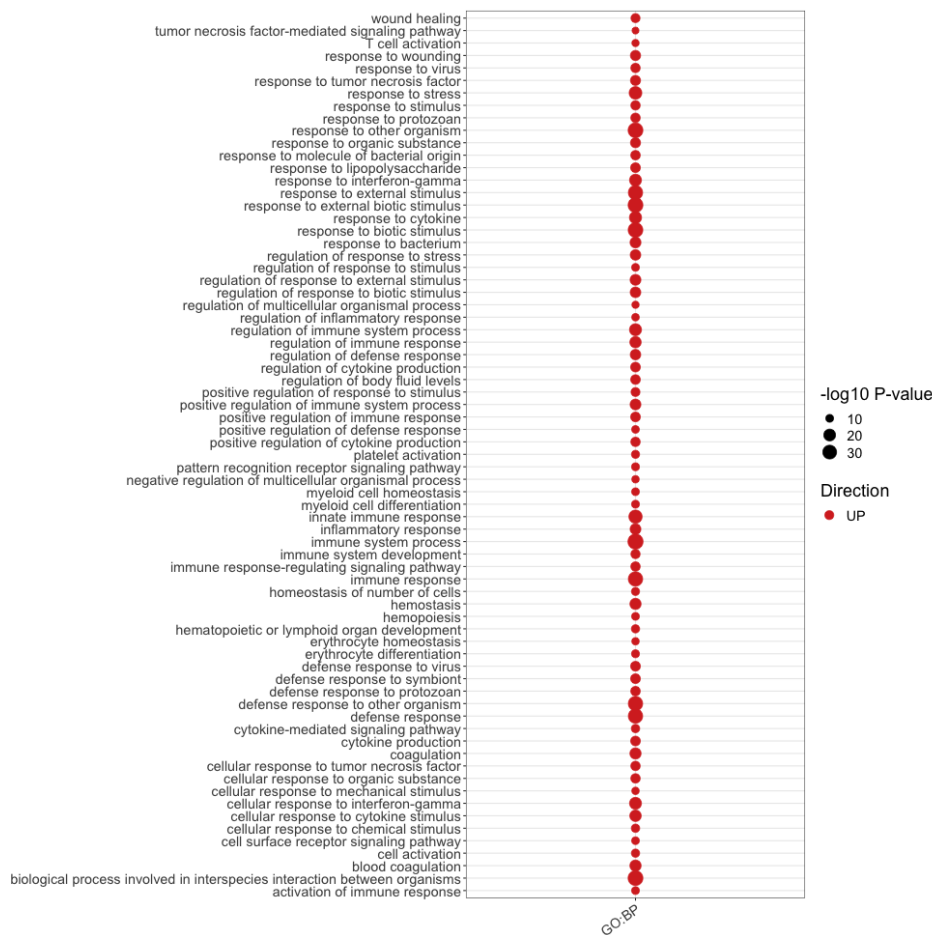


Figure 3.16: Pathways identified as significantly enriched from the list of SDE genes between microbiologically confirmed TB and unlikely TB in the discovery RNA-Seq dataset when querying the GO database (subontology = biological process; GO:BP). Pathways were identified by g:Profiler2. Points represent pathways with colour indicating directionality and size representing BH-adjusted p -value (larger = lower p -value) *SDE* = Significantly Differentially Expressed; *RNA-Seq* = RNA Sequencing; *GO:BP* = Gene Ontology with Biological Process as subontology.

heatmap and the microbiologically confirmed TB patients clustering towards the right-hand side of the heatmap.

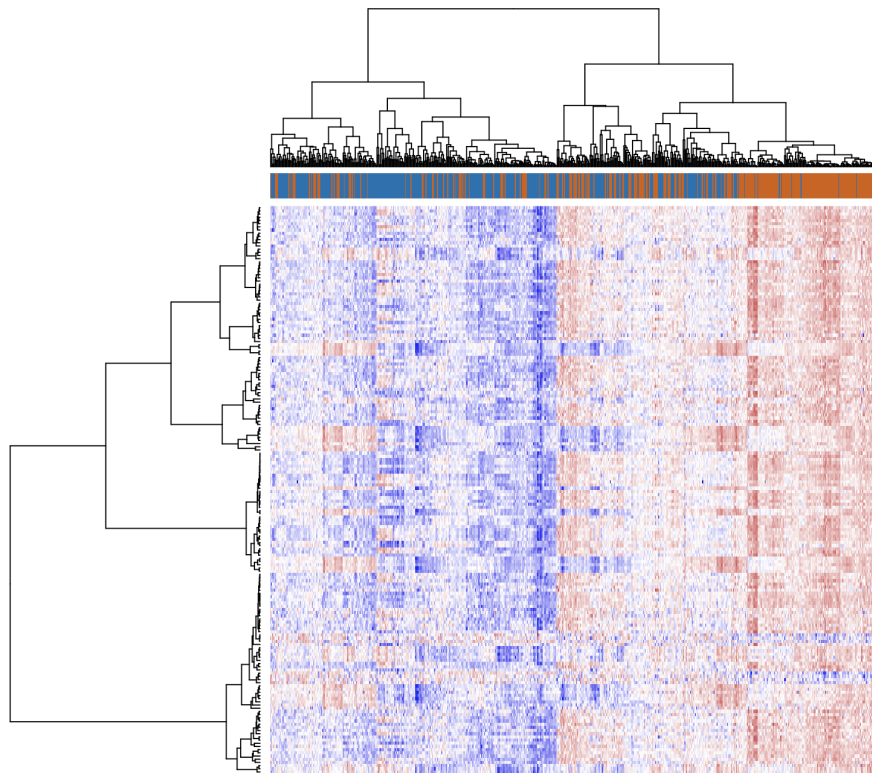


Figure 3.17: Heatmap showing clustering of genes (rows) and samples (columns) of the SDE genes between microbiologically confirmed TB and unlikely TB in the discovery RNA-Seq dataset.

Each pixel corresponds to one gene and one patient. The colours range from blue (gene under-expressed in microbiologically confirmed TB) to red (gene over-expressed in microbiologically confirmed TB). The stronger the blue the more under-expressed, the stronger the red the more over-expressed. The disease group of each patient is described in the top bar (red = microbiologically confirmed TB; blue = unlikely TB). Hierarchical clustering (Unweighted Pair Group Method with Arithmetic Mean or UPGMA method) was used to group the patients and the genes in clusters. *SDE* = Significantly Differentially Expressed; *RNA-Seq* = RNA Sequencing

3.4 Summary

In this chapter, I have used whole blood RNA-Seq data to compare the gene expression profile of children with microbiologically confirmed TB disease to the gene expression profile of children with unlikely TB. Most of the TB disease cases in our cohorts had PTB, whilst most of the unlikely TB cases had other respiratory tract infections. PTB and non-TB pulmonary infections have very similar symptomatology and clinical signs in children, making difficult their diagnosis for healthcare workers and clinicians. The blood transcriptome of these two disease groups presented clear differences, with 178 SDE genes. I then used a feature selection algorithm to identify a minimal gene signature for distinguishing between microbiologically confirmed TB vs. unlikely TB in children. Cross-platform validation of the candidate biomarkers using RT-qPCR, a methodology more suited for clinical use of the signature as a diagnostic test, was performed on the samples used in the discovery phase. The same genes were then measured using RT-qPCR in independent cohorts to ensure the generalisation of the findings. The signature achieved an AUC of 89.3% [CI_{95%} 82.5%-96.1%], sensitivity of 88.5% [CI_{95%} 78.9%-96.2%] and specificity of 86.0% [CI_{95%} 75.4%-94.7%] in the independent validation cohort. When the specificity was fixed at 70%, the sensitivity achieved 90.4% [CI_{95%} 80.8%-96.2%], meeting the minimal WHO requirements for sensitivity and specificity for a biomarker-based triage test for TB in children [44]. The signature achieved better performance in the EPTB cases, which are the patients

where the current tests for diagnosing TB disease underperform.

The translation of target genes from untargeted RNA-Seq to targeted RT-qPCR assays is challenging, resulting in genes failing to translate. Genes fail to translate for several reasons including sources of technical variability not accounted for in the analysis of the discovery RNA-Seq data, overfitting of the statistical model in the discovery phase, and inappropriate RT-qPCR primer design. The identified candidate biomarkers included in the 6-gene signature have successfully translated across platforms, suggesting that the development of a rapid and biomarker-based POCT could be possible. Such a test would be extremely useful in LMIC, where resources are limited and where most of the paediatric TB cases reside. A rapid, easy-to-use and cheap POCT for childhood TB would overcome most of the limitations that currently used diagnostic/triage tests have.

The work presented in this chapter is not without limitations. Although this was a multi-country study that included patients recruited in 4 African countries, independent validation cohorts recruited in other countries and continents are required to assess the generability and robustness of the signature. Transcriptomic signatures have clear advantages over sputum-based techniques in diagnosing TB, especially in children. However, there are several technical challenges in detecting mRNA from whole blood and measuring the expression of the target gene biomarkers, which can make more difficult the development of a POCT. Moreover, it is difficult to evaluate the signature's potential and applicability without a clinical trial. After the development of the POCT, in fact, a clinical trial will be needed to assess the performance of the signature in a real-world clinical setting. Considerable progress was made in the field in the past decade, the Cepheid GeneXpert MTB-HR prototype showed that a small signature can be translated into a fingerstick blood test to diagnose TB [155].

In conclusion, this is the first time that a small gene signature successfully translated across platforms and met the WHO TPPs criteria of sensitivity and specificity for a biomarker-based triage test for childhood TB. The 6-gene signature presented in this chapter represents a very promising step forward towards the end of the global TB epidemic by 2035 [74] since it can be developed in a rapid POCT that could revolutionise the diagnosis of childhood TB.

CHAPTER 4

CHANGE IN GENE EXPRESSION PROFILES DURING ANTI-TUBERCULOSIS TREATMENT IN CHILDREN

Host gene expression profiles have been explored not only as a diagnostic tool but also to monitor treatment response. Sparse gene signatures can, in fact, capture changes in the host immune system caused by infection of external pathogens or by response to treatments, which reduces antigen load and thus reduces inflammation.

Whilst studies linking gene expression profiles to treatment response are very limited in children, changes in gene expression distinguishing between early and late responders to TB treatment [284] or returning to healthy level on completion of TB treatment [147, 285] have been explored in adults. One of the very few paediatric studies was published in 2020 [148]. In this work, the gene expression profiles of 16 Indian children with TB disease before TB treatment were compared to the gene expression profiles of the same 16 children after TB treatment. 25 genes were identified as SDE and therefore as potential biomarkers to monitor the treatment response.

Monitoring TB treatment response is critical, especially in case of treatment failure. There are no fast and reliable methods to monitor treatment response in patients with TB. Conversion of sputum smear and culture from positive to negative after 2 months of TB treatment is currently used to predict treatment success. These tests achieve a sensitivity of less than 40% and specificity of 85% in adults [160], and their performances are even lower in children [148].

Since optimal tests to monitor treatment response in patients with TB are not yet available, biomarkers could be used to develop a tool to monitor or shorten treatment, identify treatment failure and/or predict long-term outcomes.

4.1 Overview and chapter aims

TB treatment response is believed to be biphasic, with the majority of the *Mtb* bacilli being rapidly killed in the first phase, followed by the more persisting bacilli being eradicated to reduce the risk of relapse in the second phase [286]. The treatment of paediatric TB currently varies in duration between 4 months, in drug-sensitive and not severe TB cases, and up to 18 months, in drug-resistant TB cases [287]. The WHO guidelines [288] have been recently updated following the publication of the SHINE trial [16]. This multi-country, randomised and controlled study showed that the symptomatic and non-severe cases of paediatric TB, which are two-thirds of the global paediatric TB cases, can be treated with a regimen shorter than the previously WHO-endorsed 6 months with very similar treatment success.

Response to treatment is currently monitored through physical examination, chest radiography and laboratory monitoring [289]. The symptomatology of TB disease, such as fever, cough, weight loss and normal growth rate, generally improves within the first few weeks of treatment. A non-improvement in symptoms might be seen as a treatment failure. For PTB, chest radiography returned to a healthy state

can also be used as a sign of recovery from the disease. However, chest radiography might take months to show any signs of improvement. The most important evidence of improvement is sputum conversion from positive to negative, which predicts treatment success with sensitivity and specificity lower than 40% and 85% respectively [160].

The low accuracy of the methods currently used to monitor TB treatment response and the difficulties of collecting biological samples, especially in children, highlight the need for new, rapid and non-invasive approaches. In Chapter 4 of this thesis, I present results from a paediatric RNA-Seq longitudinal dataset generated from samples collected before and during treatment in children with either TB disease or unlikely TB. The aims of this chapter are:

1. To compare the SDE genes between children with microbiologically confirmed TB and unlikely TB before treatment to the SDE genes between children with microbiologically confirmed TB and unlikely TB during treatment
2. To describe the whole blood host transcriptomic profiles of children with microbiologically confirmed TB disease pre anti-TB treatment and after 3 months of anti-TB treatment and their changes over time
3. To assess whether the expression levels of the 6 candidate biomarkers identified in Chapter 3 change during treatment

The structure of this chapter is as follows. Section 4.2 describes the clinical cohorts, the study design and the data generation pipeline. Results are introduced in Section 4.3, followed by the summary and discussion in Section 4.4.

4.2 Cohorts and study design

The study included children recruited at Red Cross War Memorial Children's Hospital (Cape Town, South Africa) between September 2011 and December 2014. Children were eligible for enrollment in the study if they were 15 years old or younger and had features suggestive of PTB. Children were diagnosed with PTB if they had a cough of any duration and one of the following: weight loss, a positive TST, chest radiography consistent with PTB or household contact with TB in the preceding 3 months. Children were excluded if they had TB treatment before enrollment, informed consent could not be obtained, they could not attend any follow-up visit or if a pulmonary specimen could not be collected. Written informed consent was obtained from the parents prior to enrollment in the study. Biological samples were collected at enrollment (i.e., before treatment) and after 3 months (i.e., during treatment). This study was approved by the Human Research Ethics Committee, Faculty of Health Sciences, University of Cape Town (HREC 827/2019).

Children diagnosed with TB disease were initiated on TB treatment by the treating clinician, who prescribed TB drug regimens using national TB protocols and based on drug-resistant patterns [<http://www.kznhealth.gov.za/family/National-Childhood-TB-Guidelines-2013-ZA.pdf>; accessed February 23rd, 2023]. Children with unlikely TB received treatments other than any anti-TB treatments based on their initial diagnosis.

119 children were included in the study and categorised into two groups: microbiologically confirmed TB ($n=70$) and unlikely TB ($n=49$). For each child, biological samples were collected at enrollment and

at 3 months. The clinical features at enrollment of the children in the cohort are described in Table 4.1. Children with microbiologically confirmed TB were older compared to children with unlikely TB with a median age in months of 63.2 [23.0-99.0] and 26.0 [17.0-68.0] respectively. The percentage of females in the microbiologically confirmed TB group was higher compared to the percentage of females in the unlikely TB group (62.9% and 49.0% respectively). All the microbiologically confirmed TB cases had PTB, according to the design study. Children in the unlikely TB group had symptomatology similar to TB. 67.3% had a respiratory tract infection. Other diagnoses were chronic lung disease (4.1%), asthma (2%), allergic rhinitis (4.1%), failure to thrive (2%), and ventricular septal defect (2%).

Clinical metadata		
	Confirmed TB (n=70)	Unlikely TB (n=49)
Age (months), median (IQR) missing	63.2 (23.0-99.8) 0	26.0 (17.0-68.0) 0
Female, no. (%) missing	44.0 (62.9%) 0	24.0 (49.0%) 0
HIV-infected, no. (%) missing	10.0 (14.3%) 0	8.0 (16.3%) 0
PTB, no. (%) missing	70.0 (100.0%) 0	
Respiratory tract infection, no. (%) missing		33.0 (67.3%) 0

Table 4.1: Clinical features at enrollment of patients included in the longitudinal study to assess the changes of gene expression profile during treatment in children with TB disease or unlikely TB.

IQR: Interquartile range; HIV: Human Immunodeficiency Virus; PTB: Pulmonary Tuberculosis

4.2.1 Case definitions

Children were evaluated at enrollment through clinical, microbiological and radiological investigations, and categorised into two groups: microbiologically confirmed TB and unlikely TB. The categorisation of children was done following the algorithm described in Section 2.3. In brief, microbiological confirmation of TB disease was obtained with a positive Xpert MTB/RIF (Cepheid, USA) on induced sputum. When children had clinical signs of TB disease but microbiological confirmation was not achieved, they were categorised as having unconfirmed TB disease. When the diagnosis of TB was ruled out but the children were symptomatic, they were categorised as unlikely TB. The unlikely TB group was assumed to have lower respiratory tract infection caused by pathogens different from *Mtb*.

Although the clinical algorithm describes the classification of patients into 4 diagnostic groups, only microbiologically confirmed TB and unlikely TB patients were included in the analysis described in this chapter.

4.2.2 Data generation

The dataset was generated using whole blood RNA-Seq on whole blood. Whole blood was collected into PAXgene Blood RNA Tubes (PreAnalytiX) from patients at the earliest time point after recruitment and prior to any treatment initiation. PAXgene tubes were frozen within 6 hours after collection and stored at -80°C until extraction. Total RNA was isolated using the PAXgene Blood RNA Kits following the recommended methodology and sent for sequencing at The Wellcome Centre for Human Genetics in Oxford, United Kingdom.

The protocol followed to generate the RNA-Seq data was as follows. Genomic material was quantified

using RiboGreen (Invitrogen) on the FLUOstar OPTIMA plate reader (BMG Labtech) and the size profile and integrity were analysed on the 2200 TapeStation (Agilent, RNA ScreenTape). Input material was normalised, and strand-specific library preparation was completed using NEBNext® Ultra™ II mRNA kit (NEB) and NEB rRNA/globin depletion probes following the manufacturer's instructions. Libraries were prepared on a Tetrad (Bio-Rad) using *in-house* unique dual indexing primers [269]. Individual libraries were normalised using Qubit and pooled together. The pooled library was diluted to ~10 nM for storage, denatured and further diluted prior to loading on the sequencer. Samples were randomised across sequencing plates. Sequencing was performed at the Oxford Genomics Centre using a Novaseq6000 platform at 150 base-pair paired-end configuration, generating a raw read count of 30 million reads per sample. Following sequencing, the raw RNA-Seq reads were processed as outlined in Section 2.5.2.

4.3 Results

The aim of this study was to identify the changes in gene expression after 3-month treatment in children with TB disease and unlikely TB. Whole blood RNA-Seq of children with microbiologically confirmed TB ($n=70$) and unlikely TB ($n=49$) were included in the study. Biological samples were collected at enrollment (before treatment) and 3 months after enrollment (during treatment). The cohort included HIV-infected ($n=18$) and HIV-uninfected ($n=101$) children.

4.3.1 Data pre-processing

The dataset was composed of two sequencing batches known as batch A and batch C, with 34 and 204 samples respectively. Prior to merging the two datasets, pre-processing was performed on each dataset separately. Genes with low counts in at least one dataset were excluded with the `filterByExpr` function [243]. This function retains the genes that have an arbitrary minimum number of counts (10 counts by default) in a number of samples equal to the minimum group sample size [Section 2.6.2]. PCA was performed to identify and remove the outliers ($n=1$), followed by the identification and exclusion of samples with *XIST* gene expressed but categorised as males in the clinical database ($n=2$), or samples with *XIST* gene not expressed but categorised as females in the clinical database. As described in Section 2.6, the *XIST* (X-inactive specific transcript) gene is a non-coding RNA on the X chromosome and is only expressed in cells containing at least two X chromosomes (i.e., female cells) [241]. If no errors in transcribing the clinical records were identified, samples were excluded because mislabeling or misplacing could have taken place in the laboratory.

The two sequencing batches were merged keeping only the genes in common. PCA was performed again on the unnormalised counts to assess the presence of any source of technical variability [Figure 4.1]. A batch effect was captured by PC1 and PC4, with the samples in the RNA-Seq batch C clustering towards the higher values of PC1 [Figure 4.1; Panel A] and PC4 [Figure 4.1; Panel B]. The sample batch was accounted for in the DE analysis and corrected with the `removeBatchEffect` function [236] [Section 2.6] after normalisation. PCA performed on the normalised and batch-corrected data showed complete correction of the batch effect [Figure 4.1; Panels C, D].

The count matrix was normalised using the edgeR package [243] [Section 2.6]. edgeR uses Trimmed Mean of M-values (TMM) to estimate scale factors between samples that represent sample-specific biases. These factors are then multiplied by the library size to yield the effective library size (i.e.,

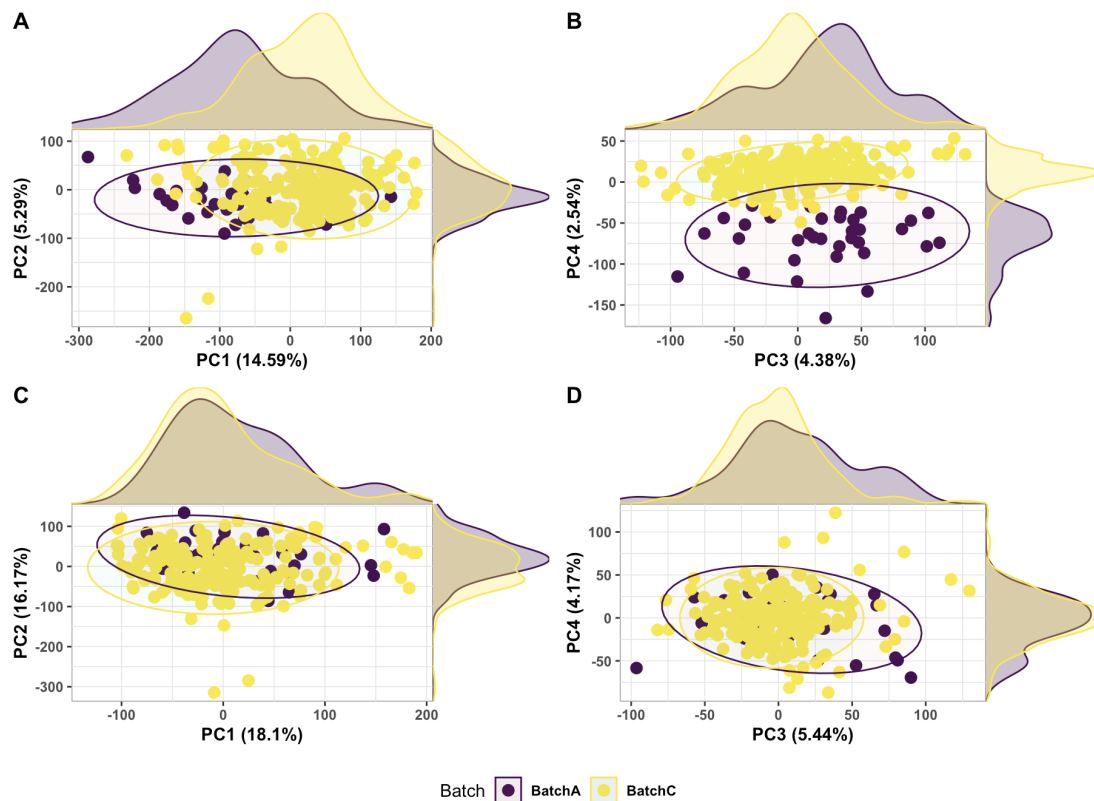


Figure 4.1: PCA biplots from PCA performed on the RNA-Seq dataset coloured according to sample batch. PC 1-2 and PC 3-4 are shown. Points represent samples and are coloured according to sample batch before [Panel A and Panel B] and after batch correction [Panel C and Panel D]. The distributions of PC1, PC2, PC3 and PC4 are shown in the density plots. *PCA* = Principal Component Analysis; *RNA-Seq* = RNA Sequencing; *PC* = Principal Component

the library size that we would have gotten if those biases were not present). Age and sequencing batch were included in the statistical model for DE analysis. The `duplicateCorrelation` function [255] was also used to account for the patient ID in the statistical model and avoid pseudoreplication in the time course analysis. Sex was not included in the statistical model since it was not significantly different (uncorrected p -value >0.05) between the disease groups of interest. PCA was performed anew after normalisation and batch correction to assess the presence of any other sources of technical variability or association of the PCs with any clinical variables. HIV status was captured by PC2 and PC4 [Figure 4.2; Panels C, D]. The time point (i.e., baseline or 3-month follow-up visit) was captured by PC1, PC2 and PC3 [Figure 4.2; Panels I, J]. Biplots from PCA coloured according to the concatenation of the disease group and the visit/time point [Figure 4.2; Panels E, F] show that the greatest difference between baseline and follow-up visits was due to the microbiologically confirmed TB groups. This suggests that PCA captures a bigger difference due to treatment in the microbiological confirmed TB patients compared to the unlikely TB patients. Age is captured by PC1 with the older children clustering towards the highest values of PC1 [Figure 4.2; Panels K, L]. No differences were observed in terms of sex [Figure 4.2; Panels G, H].

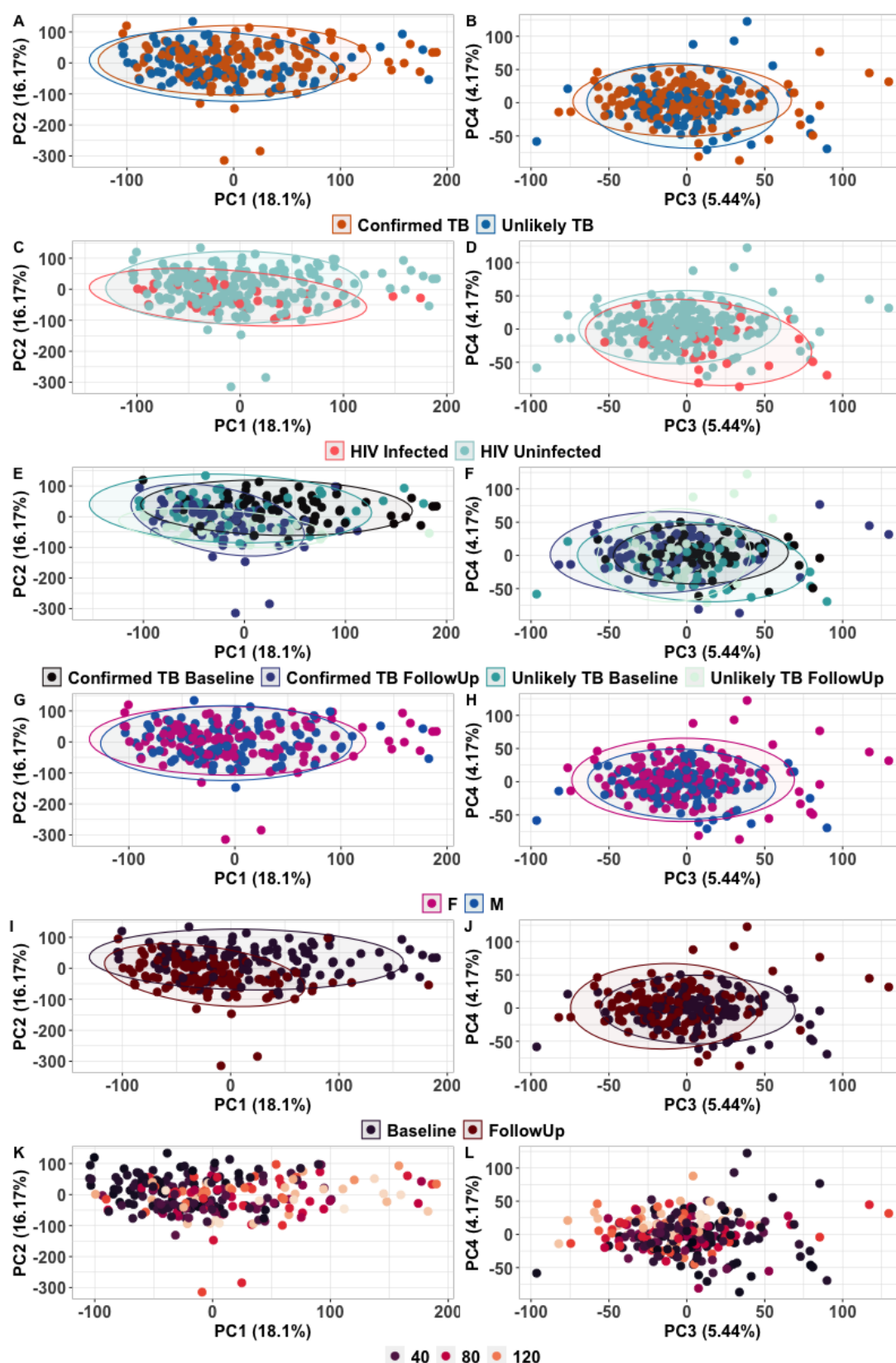


Figure 4.2: PCA biplots from PCA performed on the RNA-Seq dataset coloured according to clinical variables. PC 1-2 and PC 3-4 are shown. Points represent samples and are coloured according to disease group [Panel A and Panel B], HIV status [Panel C and Panel D], disease group and visit [Panel E and Panel F], sex [Panel G and Panel H], visit [Panel I and Panel J] and age [Panel K and Panel L]. PCA = Principal Component Analysis; RNA-Seq = RNA Sequencing; PC = Principal Component; HIV = Human Immunodeficiency Virus; F = female; M = male

4.3.2 Comparison between confirmed TB versus unlikely TB

The gene expression profile of children with microbiologically confirmed TB ($n=70$) was contrasted to the gene expression profile of children with unlikely TB ($n=49$) before treatment. The two disease groups were contrasted again after 3 months of treatment to assess whether the gene expression profiles had changed since baseline. Genes with BH-adjusted p -value <0.05 and $\log_2FC >1$ or $\log_2FC <-1$ were considered SDE. 170 genes (11 under- and 159 over-expressed in TB disease) [Appendix C] were SDE when the microbiological confirmed TB group was contrasted to the unlikely TB group at baseline [Figure 4.3; Panel A]. Most of the top SDE genes are known to be involved in the host immune response in TB and/or identified as biomarkers for TB disease in previous studies [274]; such as *GBP5*, *CYB561*, *P2RY14*, and *GBP6*. The number of SDE genes decreased when the microbiologically confirmed TB group was contrasted to the unlikely TB group during treatment [Figure 4.3; Panel B]. 26 genes (18 under- and 8 over-expressed in TB disease) [Appendix D] were SDE in this comparison. 3 out of the 18 genes under-expressed in TB disease were part of the immunoglobulin (IG) family. IG genes are produced by B cells, which are involved in the host immune response against *Mtb*. Under-expression of IG genes in patients with TB disease after 3 months of anti-TB treatment compared to patients with unlikely TB after 3 months of treatment might suggest improvement in the TB patients during treatment.

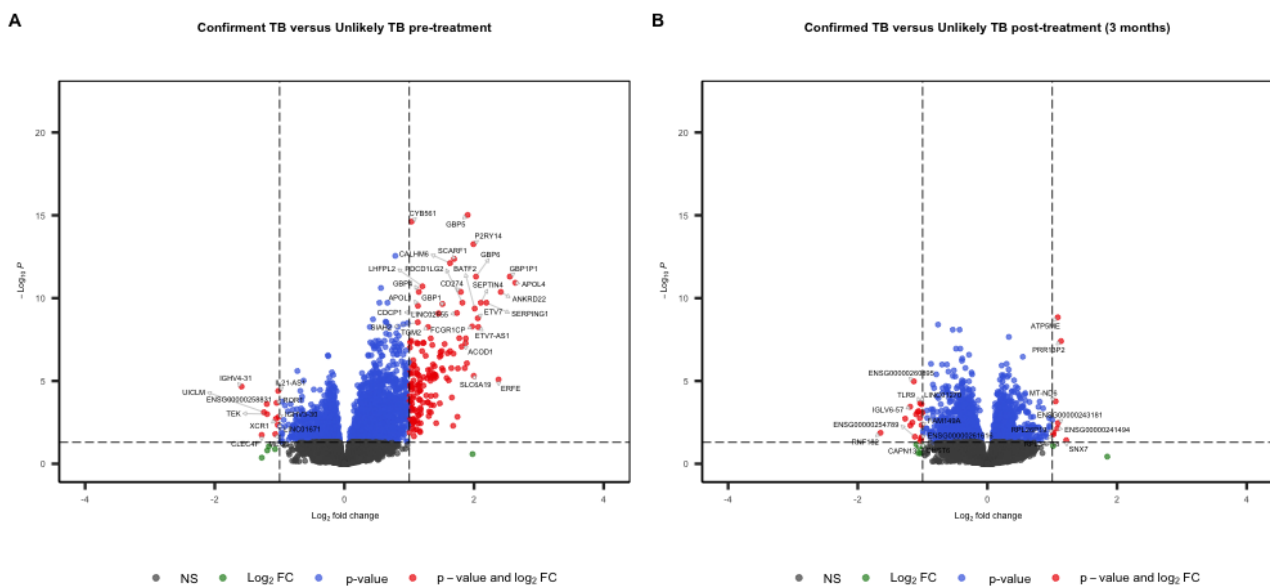


Figure 4.3: Volcano plots showing $\log_2FC$ and BH-adjusted p -values from Differential Expression analysis comparing microbiologically confirmed TB vs. unlikely TB in the RNA-Seq data before and during treatment. Points represent genes and are coloured as follows: red - genes with BH-adjusted p -values <0.05 and absolute $\log_2FC >1$; blue - genes with BH-adjusted p -values <0.05 and absolute $\log_2FC <1$; green - genes with BH-adjusted p -values >0.05 and absolute $\log_2FC >1$; grey - genes with BH-adjusted p -values >0.05 and absolute $\log_2FC <1$. Genes to the right of zero are over-expressed in TB. FC = Fold Change; BH = Benjamini-Hochberg; DE = Differential Expression; $RNA\text{-}Seq$ = RNA Sequencing

To investigate the relationship between the genes with BH-adjusted p -values <0.05 before and during treatment (i.e., microbiologically confirmed TB vs. unlikely TB before treatment and microbiologically confirmed TB vs. unlikely TB during treatment) a disco score was calculated [290]. The disco score is a numerical value that takes into account the magnitude of the gene expression change (i.e., the absolute value of $\log_2FC$), the direction of the gene expression change (i.e., the sign of $\log_2FC$) and

significance (i.e., BH-adjusted p -value). A positive disco score indicates a concordant gene (i.e., same direction in both comparisons). Contrarily, a negative disco score indicates a discordant gene (i.e., different direction in the two comparisons). The higher the absolute value of the disco score, the more concordant or discordant the gene in the two comparisons. Figure 4.4 shows the concordance and discordance of the genes with BH-adjusted p -value <0.05 in both comparisons of interested.

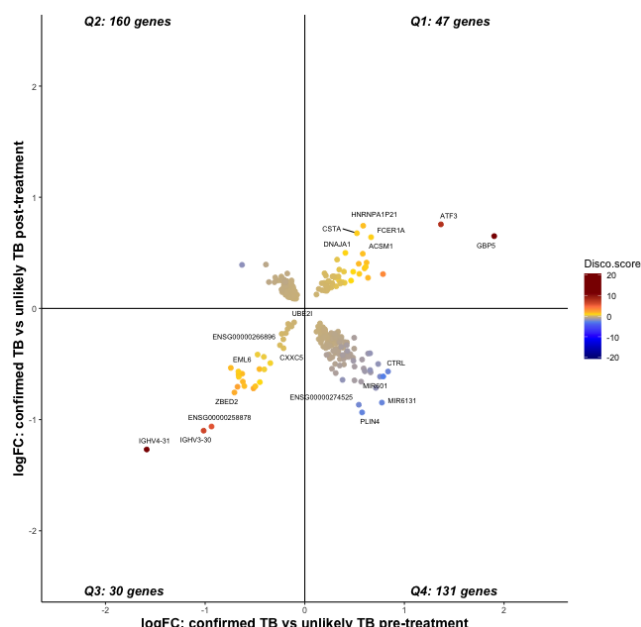


Figure 4.4: Concordance and discordance of the $\log_2FC$ of genes with BH-adjusted p -value <0.05 in microbiologically confirmed TB vs. unlikely TB before treatment against the $\log_2FC$ of genes with BH-adjusted p -value <0.05 in microbiologically confirmed TB vs. unlikely TB during treatment.

Points represent genes and are coloured according to their disco score. The stronger the red colour the more concordantly regulated the gene pair; the stronger the blue colour, the more discordantly regulated the gene pair. FC: Fold Change; BH: Benjamini-Hochberg; Q: Quadrant

The top genes with the highest positive disco score in Figure 4.4 are *GBP5* (Guanylate Binding Protein 5) and *ATF3* (Activating Transcription Factor 3), which were over-expressed in the microbiologically confirmed TB group compared to the unlikely TB group before and during treatment. Both genes play a pivotal role in the host's immune response to *Mtb*. *GBP5* is an activator of both AIM2 and NLRP3 inflammasome assembly, which have a role in response to pathogenic bacteria [272, 273]. *ATF3* is a pivotal regulator of the innate immune response, and it has been shown to be over-expressed during the early phases of *Mtb* infection and that depletion of this gene enhances *Mtb* survival in the macrophages [291]. Although they had a BH-adjusted p -value <0.05 in both comparisons, their significance in terms of $\log_2FC$ and BH-adjusted p -value decreased during treatment (*GBP5* before treatment: $\log_2FC=1.9$, BH-adjusted p -value= 9.5×10^{-16} ; *GBP5* during treatment: $\log_2FC=0.7$, BH-adjusted p -value= 1.7×10^{-02} ; *ATF3* before treatment: $\log_2FC=1.4$, BH-adjusted p -value= 1.0×10^{-06} ; *ATF3* during treatment: $\log_2FC=0.8$, BH-adjusted p -value= 2.0×10^{-02}). This might suggest that the two genes were returning to healthy expression levels.

4.3.3 Time course analysis: Comparison between confirmed TB pre-TB treatment versus confirmed TB post-TB treatment

To assess how gene expression in children with microbiologically confirmed TB changes after anti-TB treatment, a time course analysis was carried out. Gene expression profiles of microbiologically

confirmed TB patients during treatment ($n=70$) were compared to gene expression profiles of microbiologically confirmed TB patients before treatment ($n=70$).

Genes with BH-adjusted p -value <0.05 and $\log_2FC >1$ or $\log_2FC <-1$ were considered SDE. 616 genes (483 genes under-expressed and 133 genes over-expressed during treatment) [Appendix E] were SDE between the two time points [Figure 4.5]. Most of the genes whose expression was reduced after anti-TB treatment are known to be involved in the immune response to TB (i.e., *C1QA*, *C1QB*, *GBP6*, *GBP5*, *FCGR1A*, *FCGR1B*, *FCGR1CP*).

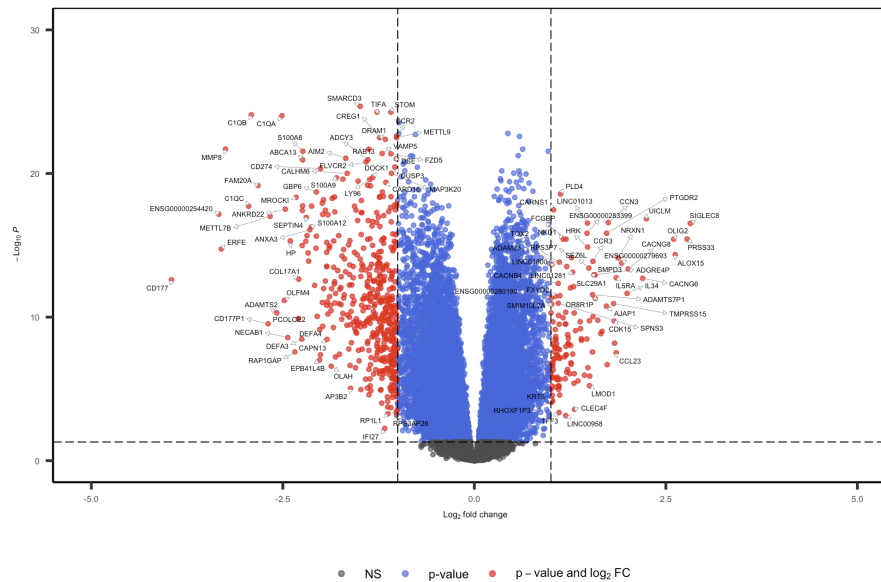


Figure 4.5: Volcano plot showing $\log_2FC$ and BH-adjusted p -values from DE analysis comparing microbiologically confirmed TB during treatment vs. microbiologically confirmed TB before treatment.

Points represent genes and are coloured as follows: red - genes with BH-adjusted p -values <0.05 and absolute $\log_2FC >1$; blue - genes with BH-adjusted p -values <0.05 and absolute $\log_2FC <1$; green - genes with BH-adjusted p -values >0.05 and absolute $\log_2FC >1$; grey - genes with BH-adjusted p -values >0.05 and absolute $\log_2FC <1$. Genes to the right of zero are over-expressed in TB during treatment. FC = Fold Change; DE = Differential Expression; BH = Benjamini-Hochberg

The differences in the gene expression profile caused by anti-TB treatment in the microbiologically confirmed TB group are highlighted by the heatmap in Figure 4.6. An unsupervised clustering algorithm was used to group the samples based on their similarity, obtaining a good separation between the two time points.

Pathway enrichment analysis was performed on the lists of SDE genes between microbiologically confirmed TB after TB-treatment vs. microbiologically confirmed TB before TB-treatment [Figure 4.7] using clusterProfiler [260], a pathway enrichment analysis tool designed to perform over-representation analysis [292] to compare functional profiles of various conditions (e.g., different treatment groups). clusterProfiler tested whether the genes were enriched in GO terms (Biological Process ontology) with a BH-adjusted p -value <0.05 . The directionality of the pathways (i.e., under-expressed and over-expressed) was identified from the list of genes that increased or decreased in the microbiologically confirmed TB after TB-treatment vs. microbiologically confirmed TB before TB-treatment.

777 pathways were identified as significantly enriched (696 under-expressed and 81 over-expressed) from the list of SDE genes between microbiologically confirmed TB after anti-TB treatment vs. microbiologically confirmed TB before anti-TB treatment.

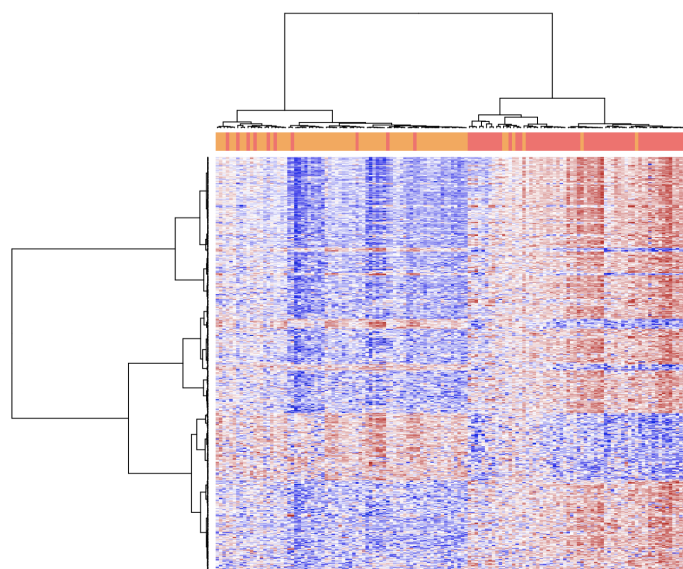


Figure 4.6: Heatmap showing clustering of genes (rows) and samples (columns) of the SDE genes between microbiologically confirmed TB during treatment and microbiologically confirmed TB before treatment. Each pixel corresponds to one gene and one patient. The colours range from blue (gene under-expressed) to red (gene over-expressed). The stronger the blue the more under-expressed, the stronger the red the more over-expressed. The disease group of each patient is described in the top bar (red = microbiologically confirmed TB before treatment; orange = microbiologically confirmed TB during treatment). Hierarchical clustering (Ward's clustering criterion) was used to group the patients and the genes in clusters. *SDE* = *Significantly Differentially Expressed*

Figure 4.7 shows the top 20 under-expressed and top 20 over-expressed pathways in the microbiologically confirmed TB after TB-treatment. Among the top 20 pathways under-expressed in microbiologically confirmed TB after TB-treatment vs. microbiologically confirmed TB before TB-treatment, many were involved in the immune response and/or response to viruses or bacteria, such as “response to molecule of bacterial origin” (GO:0002237), “response to virus” (GO:0009615), “regulation of inflammatory response” (GO:0050727). Many GO terms linked to leukocyte activation and migration, such as “myeloid leukocyte activation” (GO:0002274), “phagocytosis” (GO:0006909), “leukocyte chemotaxis” (GO:0030595), and “leukocyte migration” (GO:0050900), were also under-expressed in microbiologically confirmed TB after TB-treatment, showing a higher host immune response before anti-TB treatment compared to their levels after anti-TB treatment. The under-expression of the “phagocytosis” (GO:0006909) term, which was over-expressed before commencing of anti-TB treatment, suggests, as already reported [293], the involvement of macrophages into phagocytosis of *Mtb* bacilli after infection and before treatment.

Some GO terms involving the Wnt pathway, such as “Wnt signaling pathway” (GO:0016055), “cell-cell signaling by wnt” (GO:0198738) and “regulation of Wnt signaling pathway” (GO:0030111), were over-expressed after anti-TB treatment. The Wnt pathway, a pathway involved in cellular proliferation, differentiation, apoptosis, motility, and polarization of cells [294], has recently been involved in immunoregulatory mechanisms in infectious and inflammatory diseases, such as TB [294]. *Mtb* bacilli might impair the Wnt pathway, especially in severe cases [295]. This suggests that anti-TB treatment might return the expression levels of genes involved in this pathway to a healthy level.

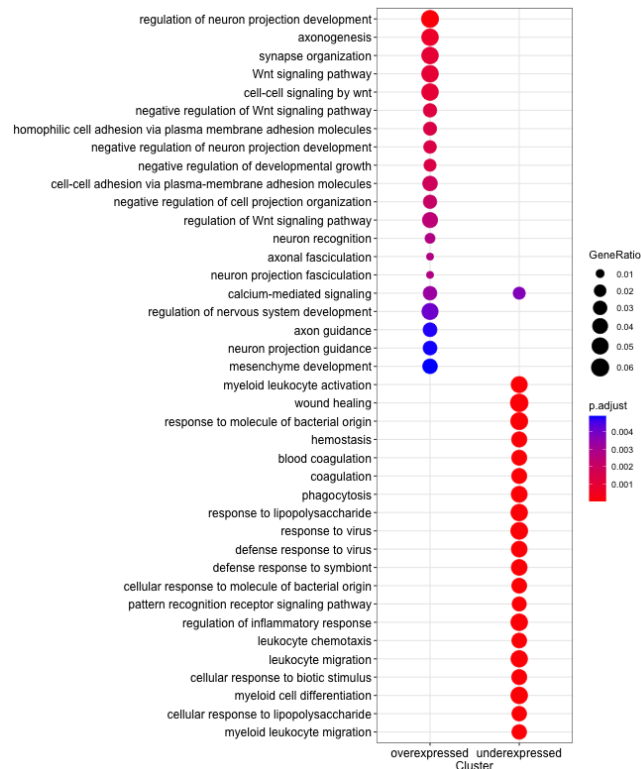


Figure 4.7: Pathways over-expressed and under-expressed between microbiologically confirmed TB patients after and before anti-TB treatment.

Points represent pathways with colour indicating their statistical significance in terms of BH-adjusted p -values, and size indicating the percentage of total SDE genes in the given GO term. BH-adjusted p -values < 0.05 were considered significant. *BH* = Benjamini-Hochberg

4.3.4 Performance of the 6-gene diagnostic signature pre- and post-TB treatment

In Chapter 3, I identified and validated a 6-gene diagnostic signature for paediatric TB disease composed of *GBP5* (Guanylate Binding Protein 5), *CDKN1C* (Cyclin Dependent Kinase Inhibitor 1C), *CPVL* (Carboxypeptidase Vitellogenic Like), *CYB561* (Cytochrome B561), *RSAD2* (Radical S-Adenosyl Methionine Domain Containing 2). Biological samples collected from children with microbiologically confirmed TB and unlikely TB at enrollment were used to identify and validate the candidate biomarkers. The signature achieved an AUC, sensitivity and specificity of 89.3% [CI_{95%} 82.5%-96.1%], 88.5% [CI_{95%} 78.9%-96.2%] and 86.0% [CI_{95%} 75.4%-94.7%] respectively in the validation cohort [Section 3.3.4], showing optimal performance in discriminating children with microbiologically confirmed TB from children with unlikely TB before treatment. To assess whether the signature performance changes during treatment, log₂FC, BH-adjusted p -value, AUC, sensitivity and specificity of the 6 individual candidate biomarker genes and the combined 6-gene signature were evaluated in the subset of samples described in Table 4.1. Table 4.2 shows log₂FC, BH-adjusted p -value, AUC, sensitivity and specificity of the single genes for three comparisons:

1. Gene expression profiles from samples collected at enrollment from children with microbiologically confirmed TB vs. gene expression profiles from samples collected at enrollment from children with unlikely TB
2. Gene expression profiles from samples collected after 3 months of treatment from children with microbiologically confirmed TB vs. gene expression profiles from samples collected after 3 months

of treatment from children with unlikely TB

- Gene expression profiles from samples collected after 3 months of treatment from children with microbiologically confirmed TB vs. gene expression profiles from samples collected at enrollment from the same children

Confirmed TB vs. Unlikely TB pre-treatment						
Gene Name	Ensembl ID	log ₂ FC	BH-adjusted p-value	AUC	Sensitivity	Specificity
<i>GBP5</i>	ENSG00000154451	1.90	9.5×10^{-16}	88.8%	85.7%	78.6%
<i>CDKN1C</i>	ENSG00000129757	-0.78	4.7×10^{-03}	68.0%	57.1%	74.3%
<i>CPVL</i>	ENSG00000106066	0.53	7.5×10^{-04}	74.4%	75.5%	70.0%
<i>CYB561</i>	ENSG00000008283	1.03	2.4×10^{-14}	88.4%	91.8%	72.9%
<i>RSAD2</i>	ENSG00000134321	0.78	0.08	59.2%	28.6%	91.4%
<i>EFCAB2</i>	ENSG00000203666	0.76	1.1×10^{-06}	76.4%	73.5%	72.9%
Confirmed TB vs. Unlikely TB post-treatment						
<i>GBP5</i>	ENSG00000154451	0.7	1.7×10^{-02}	70.4%	67.3%	67.1%
<i>CDKN1C</i>	ENSG00000129757	-4.7×10^{-03}	9.9×10^{-01}	48.8%	12.2%	95.7%
<i>CPVL</i>	ENSG00000106066	0.17	4.3×10^{-01}	63.3%	59.2%	68.6%
<i>CYB561</i>	ENSG00000008283	0.23	1.6×10^{-01}	65.0%	95.9%	28.6%
<i>RSAD2</i>	ENSG00000134321	0.5	0.3	58.0%	73.5%	47.1%
<i>EFCAB2</i>	ENSG00000203666	3.0×10^{-02}	9.2×10^{-01}	55.3%	42.9%	74.3%
Confirmed TB post-treatment vs. Confirmed TB pre-treatment						
<i>GBP5</i>	ENSG00000154451	-1.25	1.1×10^{-12}	77.8%	72.9%	71.4%
<i>CDKN1C</i>	ENSG00000129757	0.96	3.2×10^{-07}	72.4%	57.1%	78.6%
<i>CPVL</i>	ENSG00000106066	-0.81	9.8×10^{-13}	79.4%	81.4%	68.6%
<i>CYB561</i>	ENSG00000008283	-0.40	2.8×10^{-05}	65.2%	55.7%	71.4%
<i>RSAD2</i>	ENSG00000134321	-1.09	2.0×10^{-04}	68.4%	87.1%	41.4%
<i>EFCAB2</i>	ENSG00000203666	-1.04	1.6×10^{-18}	82.7%	78.6%	80.0%

Table 4.2: log₂FC, BH-adjusted *p*-value, AUC, sensitivity and specificity of the 6 candidate gene identified in Chapter 3 when contrasting microbiologically confirmed TB vs. unlikely TB before treatment, microbiologically confirmed TB vs. unlikely TB during treatment, and microbiologically confirmed TB during treatment vs. microbiologically confirmed TB before treatment.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve.

Five candidate biomarker genes had a significant BH-adjusted *p*-value (i.e., <0.05) when the microbiologically confirmed TB group was contrasted to the unlikely TB group at enrollment (Comparison 1). The AUC of the single genes ranged from 59.2% to 88.8%, with *GBP5* and *CYB561* being the top performing genes in this comparison. The log₂FC values highlight the difference in expression of the 6 genes between the two disease groups at enrollment. When the two disease groups were contrasted during treatment (Comparison 2), the BH-adjusted *p*-value and the log₂FC were respectively much higher and lower compared to the values in Comparison 1. The only gene that remained significant was *GBP5*, with a BH-adjusted *p*-value very close to the threshold of significance (i.e., 0.05). These results were subsequently confirmed by contrasting microbiologically confirmed TB during treatment to microbiologically confirmed TB before treatment. All 6 genes were SDE, with 5 genes (i.e., *GBP5*, *CPVL*, *CYB561*, *RSAD2* and *EFCAB2*) being over-expressed at baseline and *CDKN1C* being under-expressed at baseline, which reflected the direction of the genes when TB disease group was contrasted vs. the unlikely TB group at baseline.

Figure 4.8 shows the boxplots of the expression (logged) of 6 genes in the three comparisons.

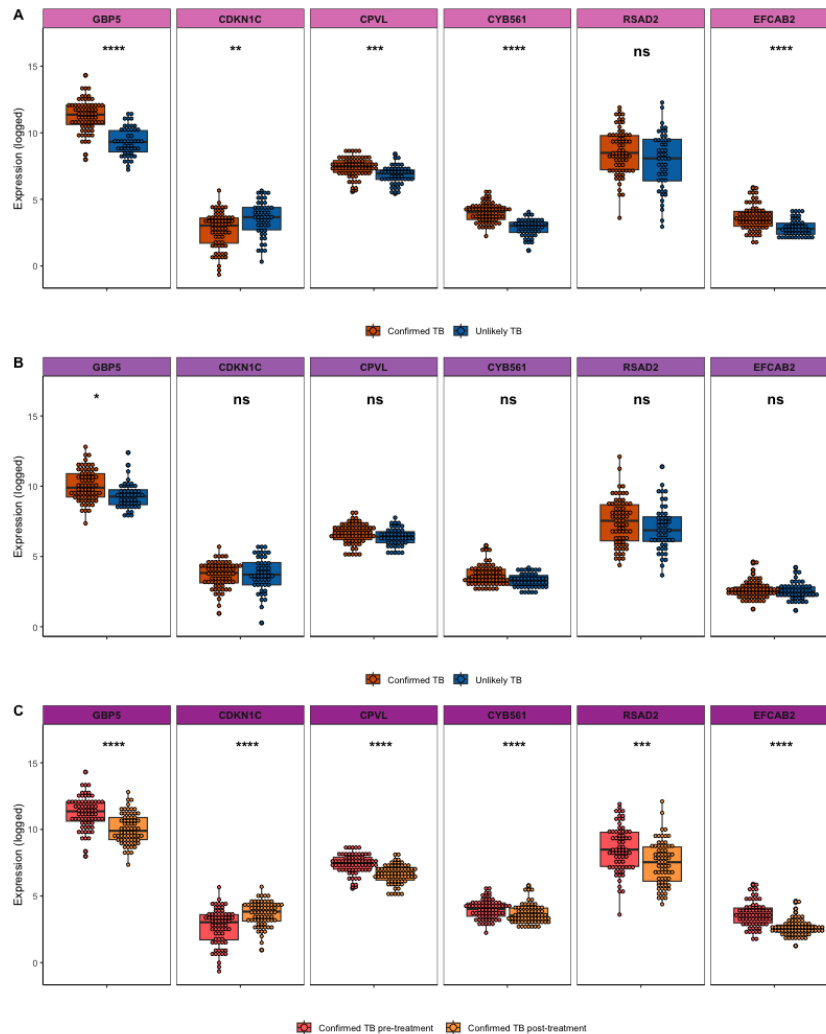


Figure 4.8: Logged expression values before treatment of the 6 candidate biomarker genes identified in Chapter 3 coloured according to disease group or time point.

Boxplots of the 6 candidate biomarker genes before treatment coloured according to disease group [Panel A]; Boxplots of the 6 candidate biomarker genes during treatment coloured according to disease group [Panel B]; Boxplots of the 6 candidate biomarker genes in the microbiologically confirmed TB group coloured according to time point [Panel C]. All p -values are adjusted for multiple testing using Benjamini-Hochberg. ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 .

The 6 genes identified as candidate biomarkers for paediatric TB in Chapter 3 appear to be SDE when comparing TB disease to unlikely TB at baseline, but their expression appears to return to healthy levels during treatment. The only gene that preserved a BH-adjusted p -value <0.05 during treatment was *GBP5*, which was one of the top SDE genes when comparing the two disease groups before treatment. The much higher BH-adjusted p -value, 9.5×10^{-16} and 1.7×10^{-02} before and during treatment respectively, suggests that the expression of this gene decreases when the patient improves, but it might require more than 3 months of treatment to return to its healthy level.

Figure 4.9 shows the performance of the 6 candidate biomarker genes when they were combined into a signature. The signature achieved an AUC of 94.5% [CI_{95%} 90.0%-98.1%], sensitivity of 85.7% [CI_{95%} 77.4%-94.3%] and specificity of 91.8% [CI_{95%} 83.7%-98.0%] in discriminating children with microbiologically confirmed TB from children with unlikely TB at baseline. There was a considerable drop when the signature performance was evaluated in the same patients during treatment. The

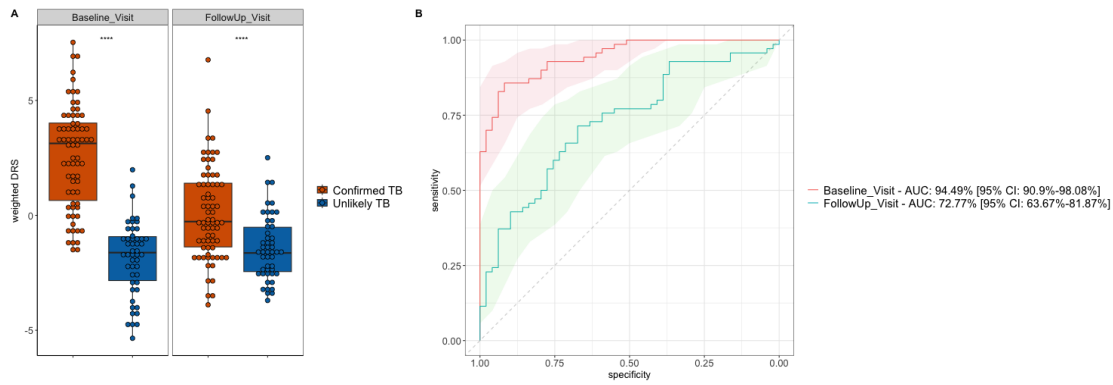


Figure 4.9: Boxplots of the wDRS of the 6-gene signature identified in Chapter 3 and the correspondent ROC curves when contrasting microbiological confirmed TB vs. unlikely TB before and during treatment. ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . wDRS = wDRS; Receiver Operating Characteristic = Receiver Operating Characteristic; AUC = Area Under the ROC curve

signature achieved an AUC of 72.8% [CI_{95%} 63.7%-81.9%], sensitivity of 71.4% [CI_{95%} 60.0%-81.4%] and specificity of 67.4% [CI_{95%} 55.1%-79.6%] in discriminating children with microbiologically confirmed TB from children with unlikely TB during treatment.

The weights/coefficients used to combine the genes into a signature were retrieved from the model trained on the discovery RNA-Seq dataset described in Section 3.3.1.

4.4 Summary

Treatment outcomes in children (younger than 15 years old) with TB disease have an average treatment success rate of 88% [2]. However, the treatment success rate is lower in children with certain forms of TB (i.e., TB meningitis) or other conditions, such as HIV co-infection or malnutrition [84]. Therefore, it is critical to identify a child not responding to treatment as early as possible to allow appropriate interventions, especially in groups where the treatment outcomes are less good.

Sputum smear or culture conversion from positive to negative after 2 months of treatment is a widely used tool to monitor treatment response. However, sputum smear or culture conversion can be used only in PTB cases, it is not a sensitive or specific indicator and a 2-month period is a long turnaround time to identify a child who is failing treatment. Moreover, most of the children treated for TB do not have microbiological confirmation of the disease at baseline, hence microbiological conversion cannot be used in this large proportion of patients [84].

Biomarkers have been extensively explored as diagnostic tools. Some recent studies have shown the potential of biomarkers in monitoring treatment response [147, 159], but studies in children are very limited [148]. Longitudinal transcriptomic studies could identify new candidate gene biomarkers for revealing the evolution and resolution of the immune response in children with TB.

In this chapter, I have analysed and described a longitudinal RNA-Seq dataset obtained from children with microbiologically confirmed TB and unlikely TB. The biological samples were collected at enrollment (i.e., before any treatments) and after 3 months of treatment. Children with microbiologically confirmed TB were treated with anti-TB treatment. Children with unlikely TB were treated based on their symptoms and diagnosis, but not with anti-TB treatment. DE analysis performed on microbiologically confirmed TB vs. unlikely TB at baseline, and on microbiologically confirmed TB vs. unlikely TB after

3 months of treatment showed that many of the genes that were SDE between the two disease groups before treatment, were returned to healthy levels during treatment. These results suggest that diagnostic candidate gene biomarkers at baseline could also be used to monitor TB treatment response since their diagnostic accuracy decreases over time during treatment. The 6-gene diagnostic signature identified and described in Chapter 3 achieved an AUC of 94.5% [CI_{95%} 90.0%-98.1%], sensitivity of 85.7% [CI_{95%} 77.4%-94.3%] and specificity of 91.8% [CI_{95%} 83.7%-98.0%] in discriminating children with microbiologically confirmed TB from children with unlikely TB at baseline. The AUC dropped to 72.8% [CI_{95%} 63.7%-81.9%], sensitivity to 71.4% [CI_{95%} 60.0%-81.4%] and specificity to 67.4% [CI_{95%} 55.1%-79.6%] in discriminating children with microbiologically confirmed TB from children with unlikely TB during treatment. *GBP5* was the only gene in the signature to be SDE between the two disease groups after 3 months of treatment, albeit with a much lower log₂FC and much higher BH-adjusted *p*-value compared to its levels before treatment. Although the majority of the genes returned to their healthy expression level after 3 months of treatment, some might require a longer therapy to resolve completely. The 6 candidate gene biomarkers were SDE when contrasting microbiologically confirmed TB during treatment vs. microbiologically confirmed TB before treatment, suggesting again that diagnostic signatures could be used to monitor treatment response.

This analysis has several limitations:

- The disease groups included were microbiologically confirmed TB and unlikely TB. This does not reflect a “real-world” situation where the majority of children who start on TB treatment do not have microbiologically confirmation. Studies are needed to assess the performance of the candidate biomarkers in the unconfirmed TB group
- Only children with PTB were included in the study. Since treatment outcomes are less good in children with other forms of TB, studies with both PTB and EPTB cases are needed
- Children were recruited in South Africa. To generalise the findings, a multi-country cohort is needed in the discovery phase
- A validation study is needed to assess the performance in independent cohorts

The aim of this chapter was to assess whether transcriptomic biomarkers could serve as a tool to monitor treatment response in children with TB disease. Although the cohort was relatively small, recruited only in one country, and included only PTB cases, the results showed that a diagnostic sparse signature could also serve as an indicator of treatment response. This provides motivation for further exploration in bigger cohorts for identifying biomarkers that, not only can be used as a tool to diagnose the disease and monitor the treatment response but also as a prognostic tool to identify individuals who progress from *Mtb* infection to TB disease and provide preventive therapy.

CHAPTER 5

A NOVEL AND PARSIMONIOUS PROTEOMIC SIGNATURE FOR DISTINGUISHING BETWEEN TUBERCULOSIS DISEASE AND OTHER RESPIRATORY INFECTIONS IN CHILDREN

The advent of high-throughput technologies in the past few decades has driven the discovery of biomarkers, allowing the identification of patterns in RNA expression or protein abundance in patients with a certain disease. A diagnostic signature is a set of host molecules (i.e., genes, proteins) that, when combined, can distinguish between the disease groups of interest or indicate a specific biological state or condition. A rapid, whole-blood based and inexpensive POCT able to generate a TB diagnostic score by measuring transcriptomic or proteomic signatures might achieve the specific milestones set by the WHO and help accomplish the *End TB Strategy* targets by 2035 [74].

The diagnosis of TB, especially in children, is a critical challenge that central or peripheral healthcare centres in LMIC face. Transcriptomic or proteomic diagnostic signatures might overcome the limitations in diagnosing TB disease and constitute the basis for a POCT that meets the WHO TPPs. Whilst the identification and validation of a novel, sparse and accurate host gene signature to diagnose TB in children was described in Chapter 3, this chapter describes the identification and validation of a novel, sparse and accurate host protein signature to distinguish TB disease from unlikely TB in children.

5.1 Overview and chapter aims

This work was part of a project funded by the NIH [<https://grantome.com/grant/NIH/R01-AI128765-02>; accessed on April 2nd, 2023], which started in 2016 and whose aim was to discover and validate host diagnostic signatures for paediatric TB, along with their translation into an accurate POCT for LMIC where the disease is endemic and resources are limited.

Since foreign pathogens elicit a characteristic immune response in the host, identifying genes and/or proteins that change their expression and/or abundance in response to a specific pathogen might provide valuable insights and help identify new signatures with the aim of developing novel diagnostic tests. Most of the studies on TB biomarkers focused on adult cohorts with paediatric datasets rarely included in the discovery phase. Moreover, signatures discovered in adult cohorts underperform in paediatric datasets. Therefore, there is a need to discover new paediatric TB biomarkers.

The identification of protein biomarkers for paediatric TB started in the 2010s with a focus on single protein biomarkers [178–182]. Since single proteins are not accurate enough in distinguishing TB disease from other disease groups of interest, single biomarkers were combined into a signature to merge their diagnostic power and achieve better performance. However, none of the proteomic signatures for paediatric TB published so far meets the WHO TPPs for a new non-sputum-based diagnostic test. Moreover, most of the very few studies published included only HIV-uninfected children [183, 185–187]. Although the HIV cases are decreasing worldwide, it is critical to include HIV-infected children in discovery and validation studies since HIV infection increases the incident of TB in children

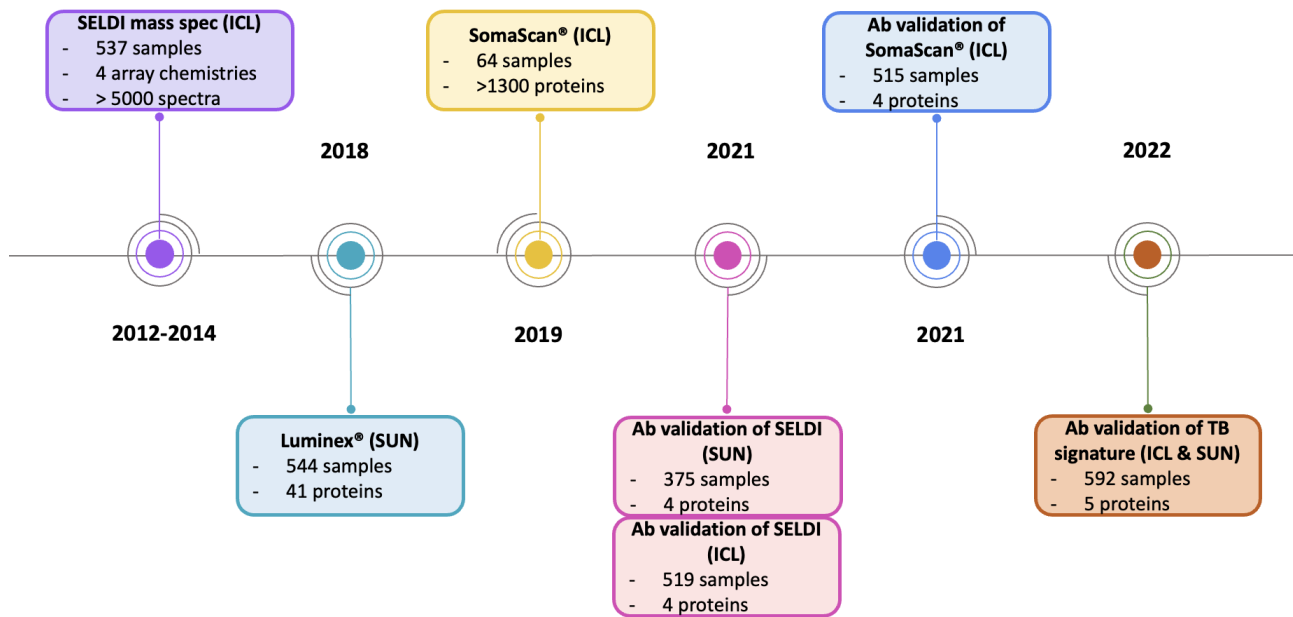


Figure 5.1: Timeline of discovery and validation phases of the diagnostic proteomic signature for childhood TB.

mass spec: Mass Spectrometry; SELDI: Surface-Enhanced Laser Desorption/Ionization; Ab: antibody; ICL: Imperial College London, London, United Kingdom; SUN: Stellenbosh University, Cape Town, South Africa.

with a more aggressive and deadly disease [92].

The main aims of the analysis described in this chapter are:

1. To explore the proteomic profiles of children with TB disease or unlikely TB
2. To identify the statistically different proteins between the two disease groups
3. To shortlist the key proteins able to distinguish between the disease groups of interest by using a feature selection algorithm
4. To evaluate the performance of the selected proteins in independent cohorts

The structure of this chapter is as follows. Section 5.2 describes the clinical cohorts whose serum samples were used in the discovery and validation phases, the study design, the data generation and analysis. Results are introduced in Section 5.3 and discussed and summarised in Section 5.4.

5.2 Cohorts and study design

This study is composed of two main phases listed below:

1. The discovery of a diagnostic signature for childhood TB in a dataset composed of three proteomic datasets generated by immunoassay
2. The independent validation by immunoassays of the discovered 4-protein diagnostic signature

Figure 5.1 describes the timeline and details of the proteomic discovery and validation experiments.

The discovery of the proteomic signature was performed on a proteomic dataset composed of:

1. The immunoassay validation of a paediatric TB diagnostic signature identified in a Surface-Enhanced Laser Desorption/Ionization (SELDI) experiment
2. The immunoassay validation of a paediatric TB diagnostic signature identified in a SomaScan® experiment
3. A list of 41 proteins measured by Luminex® assay

Measurements of protein abundance and data generation greatly differ among SELDI, SomaScan® and Luminex® assays. The SELDI assay is an untargeted proteomic technique based on mass spectrometry; the SomaScan® assay is a targeted aptamer-based technique measuring 1,301 human proteins [Section 1.11.3]; the Luminex® assay is an antibody-based technique targeting a limited set of proteins [Section 1.11.2]. Including datasets generated by three different techniques in the signature discovery phase provides a more heterogeneous comparison between TB disease and unlikely TB in children and captures the heterogeneity of the disease. The candidate protein biomarkers identified in the SELDI and SomaScan® datasets were cross-platform validated to ensure that they could be correctly targeted by immunoassays. In fact, a proteomic POCT needs to be reasonably priced, rapid, sensitive, specific, and not require extremely specialised and costly equipment. Immunoassays fulfil all these requirements. If the identified proteomic biomarkers are correctly measured by immunoassay, the translation to a POCT would be more easily achievable.

The three datasets (i.e., validation by immunoassay of the 4-protein signatures identified in the SELDI and SomaScan® assays and the Luminex® dataset) were pre-processed and analysed independently and then combined for the identification of the signature to distinguish TB disease from unlikely TB in children. 342 samples were in common across the three immunoassay experiments and therefore included in the discovery phase. The discovery dataset included microbiologically confirmed TB ($n=112$), unconfirmed TB ($n=37$) and unlikely TB ($n=193$). The independent validation cohort consisted of 592 samples with microbiologically confirmed TB ($n=123$), unconfirmed TB ($n=204$) and unlikely TB ($n=210$) patients. HIV-infected and -uninfected children were included in both the discovery and validation phases. Discovery and validation cohorts were recruited in South Africa, Malawi and Kenya in the studies described in Section 2.4. Their clinical features are shown in Table 5.1. Children in the microbiologically confirmed TB group were older (median age 59.5 [IQR 22.0-118.0], median age 52.0 [IQR 21.0-93.4] in the discovery and validation datasets respectively) than the patients in both unconfirmed and unlikely TB groups (median age 27.0 [IQR 13.0-40.4], 38.0 [IQR 14.0-86.0] in the discovery dataset; median age 26.9 [IQR 16.6-66.4], 32.3 [IQR 17.0-76.3] in the validation dataset). The percentage of females and males was around 50% in all disease groups in the discovery and validation datasets. The percentage of HIV-infected children was slightly higher in the confirmed TB and unlikely TB children in the discovery dataset (41.1% and 41.5% respectively) compared to the independent dataset (27.6% and 30.0% respectively). Most of the microbiologically confirmed and unconfirmed TB cases had PTB, whilst the most frequent diagnosis in the unlikely TB cases was respiratory tract infection in both datasets.

Figure 5.2 provides an overview of the discovery and validation experiments.

5.2.1 Case definitions

The serum used to generate the datasets described in this chapter was collected from children recruited in four African countries (South Africa, Malawi, Kenya and The Gambia) and across different studies

Discovery dataset			
	Confirmed TB (n=112)	Unconfirmed TB (n=37)	Unlikely TB (n=193)
Age (months), median (IQR)	59.5 (22.0-118.0)	27.0 (13.0-40.0)	38.0 (14.0-86.0)
missing	0	0	0
Female, no. (%)	52.0 (46.4%)	20.0 (54.1%)	81.0 (42.0%)
missing	0	0	0
HIV-infected, no. (%)	46.0 (41.1%)	11.0 (29.7%)	80.0 (41.5%)
missing	0	0	0
PTB, no. (%)	104.0 (92.9%)	33.0 (89.2%)	
missing	1 (0.9%)	0	
Respiratory tract infection, no. (%)			88.0 (45.6%)
missing			3.0 (1.5%)
Independent validation dataset			
	Confirmed TB (n=123)	Unconfirmed TB (n=204)	Unlikely TB (n=210)
Age (months), median (IQR)	52.0 (21.0-93.4)	26.9 (16.6-66.4)	32.3 (17.0-76.3)
missing	0	0	0
Female, no. (%)	52.0 (44.3%)	93 (45.6%)	92 (43.8%)
missing	0	0	0
HIV-infected, no. (%)	34.0 (27.6%)	43.0 (21.1%)	63.0 (30.0%)
missing	0	0	1 (0.5%)
PTB, no. (%)	116 (94.3%)	185 (90.7%)	
missing	1 (0.9%)	0	
Lower respiratory tract infection, no. (%)			66 (31.4%)
missing			99.0 (47.1%)

Table 5.1: Clinical features of patients included in the discovery and independent validation proteomic datasets for identifying novel paediatric TB protein biomarkers.

IQR: Interquartile range; HIV: Human Immunodeficiency Virus; PTB: Pulmonary Tuberculosis

[Section 2.4]. Children presenting to hospitals or healthcare facilities either for suspected TB disease or identified through contact tracing and with *Mtb* infection were recruited into the studies after formal consent provided by a parent or guardian. The children were then classified into 4 disease groups (microbiologically confirmed TB, unconfirmed TB, unlikely TB and *Mtb* infection) following the diagnostic algorithm described in Section 2.3. In brief, microbiologically confirmed TB diagnosis was based on the bacteriological confirmation either by culture or NAATs from at least 1 specimen obtained from the site of TB. A child was diagnosed with unconfirmed TB when the bacteriological confirmation was not obtained and the patient had at least 2 of the following: symptoms or signs suggestive of TB; close TB exposure or immunological evidence of *Mtb*; positive response to TB treatment; chest radiography or other evidence suggestive of TB. When the bacteriological confirmation was not obtained and the criteria for “unconfirmed TB” were not met, a symptomatic child was classified as unlikely TB. A healthy child with no clinical features of TB disease and a positive IGRA or TST result was classified as having *Mtb* infection.

Although the clinical algorithm describes the classification of patients into 4 diagnostic groups, only microbiologically confirmed TB, unconfirmed TB and unlikely TB patients were included in the analysis described in this chapter.

5.2.2 Data generation

SELDI dataset The SELDI technology combines two techniques: retentate chromatography and mass spectrometry. This technology involves depositing components of a protein mixture upon an array’s functionalised surface resulting in the adsorption of certain solubilised proteins [296]. After some washing steps, the retained proteins are desorbed, ionised through a laser and analysed by a mass

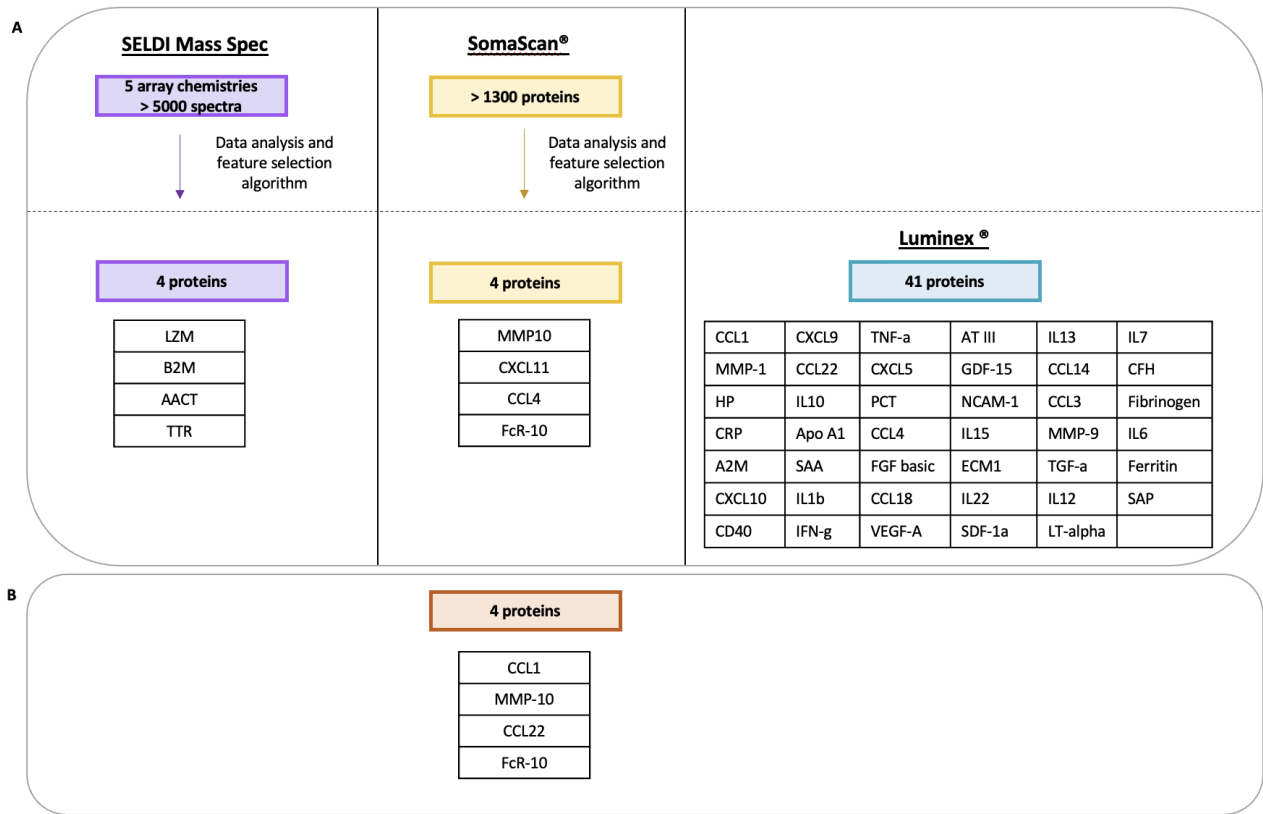


Figure 5.2: Overview of the biomarker discovery and validation phases carried out and described in Chapter 5.

The final pool of proteins in the signature discovery phase was composed of 49 proteins (4 identified in a SELDI experiment, 4 identified in a SomaScan® experiment and 41 measured by Luminex® assay) [Panel A]. The datasets in Panel A were combined and a feature selection algorithm identified 4 proteins as the best diagnostic biomarkers. The candidate biomarkers were finally validated in an independent cohort [Panel B].

SELDI: Surface-Enhanced Laser Desorption/Ionization

spectrometer, which measures the mass-to-charge ratio of the ionised molecules. Mass spectrometry, and therefore also SELDI, does not rely on the use of antibodies to target the molecules of interest. It is, in fact, an untargeted technique that detects and quantifies known and unknown proteins, or proteins for which antibodies do not exist.

The output data are, however, extremely complicated to analyse and require specific analytic pipelines. Furthermore, since SELDI is an untargeted technique, all the detectable proteins in the biological sample are measured and they need to be identified from their mass-to-charge ratio.

Data were generated by SELDI from serum samples collected from 537 children with TB disease, unlikely TB or *Mtb* infection. A 4-protein signature was identified and validated by immunoassay.

SomaScan® dataset Profiling of 1,301 human proteins was carried out in paediatric serum samples using the SomaScan® 1.3K multiplexed assay (SomaLogic, Inc.; 1.3K Assay) [297]. SomaScan® is based on protein detection using slow off-rate modified aptamers [Section 1.11.3]. It uses three levels of dilution (40%, 1% and 0.005%) to detect respectively low, medium and high abundance proteins. SomaLogic, the company that developed the SomaScan® assay, receives the biological samples and carries out the laboratory experiment. The data are then returned to the client in raw and normalised formats [Section 2.6.4].

A SomaScan® pilot experiment with 64 serum samples (40 microbiologically confirmed TB and 24

unlikely TB) was carried out to identify new diagnostic protein biomarkers for paediatric TB. The findings were then validated in the same set of samples (i.e., cross-platform validation) using either ELISA [Sections 1.11.1] or Luminex® [Section 1.11.2] immunoassays. Proteins were quantified using ELISA if they were not available via Luminex®. A final validation in 515 independent samples was finally run to assess the transferability of the findings.

Luminex® dataset Luminex® is a targeted immunoassay [Section 1.11.2]. The sample is added to a mixture of magnetic beads pre-coated with analyte-specific capture antibodies to bind to the analytes of interest. A fluorescent detection antibody is then added to the solution to bind to the specific bead-protein complexes. Once the bead-protein-antibody complexes are sorted by their spectral properties, the fluorescence, which is proportional to the abundance of the protein of interest, is measured. The Luminex® software extrapolates the concentration of the analytes of interest from the fluorescence values by using a 4- or 5-parameter logistic regression.

The Luminex® kits were supplied by Merck Millipore (Billerica, MA, USA) [HCCBP2MAG-58K, HCVD2MAG-67K, HCVD3MAG-67K, and HNDG1MAG-36K] and R&D systems (Bio-Techne, Minneapolis, USA) [custom-made kits]. The fluorescence, concentration values and detection limits of the 41 analytes were obtained using a Bio-Plex 200 system and the accompanying Bio-Plex Manager Software 6.0.

The Luminex® assay was used in both the discovery and validation phases.

Multi-platform integration and signature discovery

To overcome the important differences in terms of protein quantification and data generation between the SELDI, Somascan® and Luminex® assays, the SELDI and SomaScan® datasets were analysed individually and the candidate biomarkers identified in each of the two datasets were subsequently measured by immunoassays. Moreover, this reduced the search space to only the biomarkers with the strongest diagnostic ability in differentiating TB disease from unlikely TB in children. The three datasets used for the discovery of the signature were the following:

1. The validation by immunoassay of the 4 protein markers identified in the SELDI dataset
2. The validation by immunoassay of the 4 protein markers identified in the SomaScan® dataset
3. The 41 proteins measured by the Luminex® assay

342 were the samples in common among the three immunoassay experiments and included in the discovery of novel protein biomarkers for distinguishing between TB disease and unlikely TB. The individual datasets were pre-processed separately and then combined and analysed.

Validation of the proteomic signature

Validation of identified diagnostic biomarkers in a new set of samples that were not used for discovery is necessary to test the generalisability and reproducibility of the findings and assess whether the signature can reach clinical validation and be routinely used in clinical practice. Most of the already published potential biomarkers failed to be used in clinical practice since were either never validated or were found not to be significant in subsequent validation studies. This might be due to the small sample size of the discovery cohorts and/or to the incorrect use of the cross-validation, with consequent overfitting of the statistical model and overperforming of the signature in the discovery cohorts.

The 4 candidate protein biomarkers identified in the discovery phase were measured in the independent cohort using Luminex® or ELISA immunoassays according to the manufacturer's protocols. One of the 4 proteins was not available via Luminex® and its abundance was measured by ELISA. Two proteins were measured in the same Luminex® kit since they had similar dynamic ranges, and one protein was measured in a single-plex Luminex® kit. Samples in the independent cohort ($n=592$) were randomised across 96-well sample plates and processed according to the manufacturer's protocol. To accommodate 592 samples a total of 27 96-well plates were run, 9 plates for each Luminex® assay and 17 ELISA kits.

5.3 Results

The NIH-funded proteomic project aimed at discovering and validating a proteomic signature to diagnose TB disease in children. The project included the following datasets:

1. The discovery of a 4-protein diagnostic signature for childhood TB in a SELDI dataset and its validation by immunoassay
2. The discovery of a 4-protein diagnostic signature for childhood TB in a SomaScan® dataset and its validation by immunoassay
3. The discovery of a 4-protein diagnostic signature for childhood TB in a dataset composed of the immunoassay validations of the SELDI and SomaScan® signatures, and 41 proteins measured by Luminex® assay
4. The independent validation by immunoassays of the 4-protein signature

5.3.1 Discovery phase

SELDI signature

Discovery 537 children with TB disease, unlikely TB or *Mtb* infection, HIV-infected or -uninfected, were included in the discovery cohort. These children were recruited in South Africa, Malawi and Kenya during the studies described in Section 2.4. The signature was discovered in the South African and Malawian cohorts (346 children) and validated in the Kenyan cohort (191 children). After data pre-processing and analysis, 4 proteins were identified as candidate biomarkers to distinguish TB disease from unlikely TB in children.

After data pre-processing and analysis of the SELDI output data, 4 markers were identified:

1. Lysozyme (LZM) – m/z 14767
2. Transthyretin (TTR) – m/z 13769
3. Beta-2-Microglobulin (B2M) – m/z 10828
4. Alpha-1-Antichymotrypsin C-terminal fragment (A1ACT) – m/z 4627

The combination of the 4 markers achieved an AUC of 78.1% ($CI_{95\%}$ 72.1%-84.1%) in the training set of the discovery cohort, 82.4% ($CI_{95\%}$ 70.1%-97.1%) in the test set of the discovery cohort and 83.1% ($CI_{95\%}$ 74.2%-92.0%) in the Kenyan validation cohort. To assess if the same signature could achieve similar diagnostic power in different samples and test the reproducibility of the results, the 4 candidate biomarkers were measured by immunoassays in independent cohorts.

Validation The 4 proteins identified in the SELDI dataset as having diagnostic potential in discriminating between TB disease and unlikely TB in children were taken forward to validation in independent cohorts. Their levels were quantified in serum samples using the Luminex® immunoassay [Section 1.11.2]. Since the data were generated in two different laboratories (Imperial College London, London, United Kingdom; Stellenbosch University, Cape Town, South Africa), the Luminex® kits were ordered as part of the same production batch to minimise any source of technical variation. Samples were randomised across 96-well sample plates and processed according to the same manufacturer's protocol at each site. The data were pre-processed and analysed separately and then combined.

The concentration values and detection limits were determined from the standard curves generated from each kit's standards using the Bio-Plex Software Manager 6.1 (Bio-Rad Laboratories). The extrapolated concentration values, i.e., values outside but very close to the dynamic range that the Luminex® reader is not able to quantify but the software is able to extrapolate, were kept in the analysis. The concentrations for which the software could not extrapolate a numerical value because over the analyte dynamic range were excluded from the analysis to avoid artificially skewing the data. The concentrations not extrapolated because under the limit of detection were assigned either the lowest extrapolated value minus 10% for that specific analyte and Luminex® plate, or half the lowest limit of the dynamic range for that specific analyte and Luminex® plate, following the algorithm described in Section 2.6.3.

PCA was performed on the individual datasets and subsequently on the combined dataset to identify any source of technical variability. No plate effect was identified in the individual datasets. However, PCA on the merged dataset identified a batch effect due to the site where the samples were run (London or Cape Town). The function `normn_MA` [238] [Section 2.5], was used to minimise the batch effect between the individual datasets [Figure 5.3].

Since this dataset was meant to be merged with the other two discovery datasets (validation by immunoassay of the 4 proteins identified in the SomaScan® dataset and the 41 proteins measured by Luminex® assay), DA analysis and feature selection were not performed at this stage.

Somascan® signature

Discovery The pilot SomaScan® experiment included 64 serum samples (40 microbiologically confirmed TB and 24 unlikely TB) from Cape Town and Malawi, two cohorts described in Section 2.4. All the samples were HIV-uninfected based on SomaLogic's guidelines and included both PTB and EPTB. The data were returned pre-processed and normalised by SomaLogic. PCA was performed to identify any outliers or any source of technical variation. The protein abundances were $\log_2$ transformed and DA analysis was performed to identify proteins that were SDA between the microbiologically confirmed TB and unlikely TB groups. After testing the non-normality of the data with the Shapiro-Wilk test [Section 2.5] and since the cohort of samples was relatively small, the Mann-Whitney U non-parametric test [Section 2.5] was used to assess the significance of the proteins in the dataset. FS-PLS [Section 2.5] was used on the proteins with $\log_2\text{FC} > 0.5$ or $\log_2\text{FC} < -0.5$ and BH-adjusted p -value < 0.05 . A total of 26 proteins (13 more abundant and 13 less abundant in microbiologically confirmed TB) were SDA when contrasting microbiologically confirmed TB vs. unlikely TB [Table 5.2].

Whilst it is highly encouraging that many SDA proteins between microbiologically confirmed TB versus unlikely TB are biologically involved in inflammation and immunological processes, a minimum

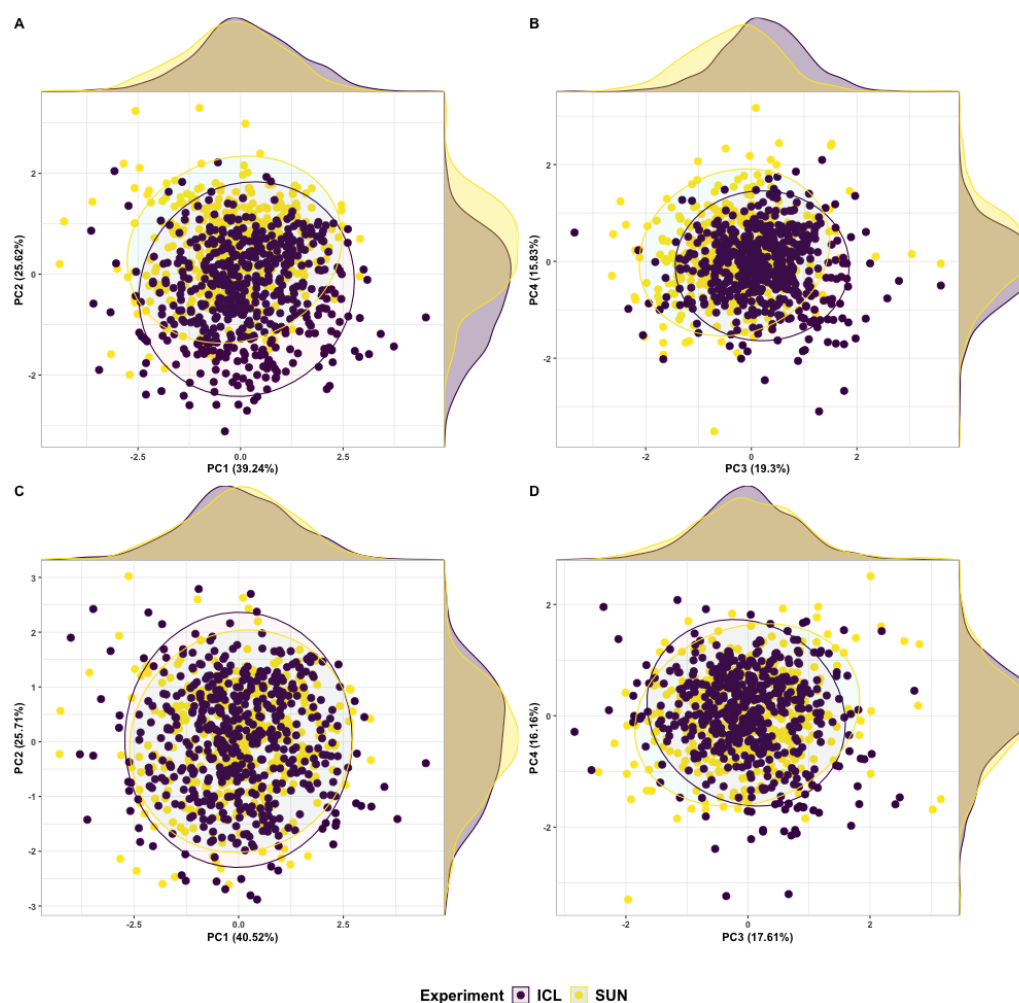


Figure 5.3: PCA biplots and density plots from PCA performed on the 4 proteins identified in the SELDI dataset and measured by immunoassay coloured according to batch.

PC 1-2 and PC 3-4 are shown before [Panel A and Panel B] and after batch correction [Panel C and D]. Points represent samples and are coloured according to the site where the assay was run. *PCA = Principal Component Analysis; PC: Principal Component; ICL: Imperial College London (London, United Kingdom); SUN: Stellenbosch University (Cape Town, South Africa)*

combination of proteins needed to be identified in order to take them forward to the validation phase. FS-PLS [Section 2.5] was used to identify the combination of SDA proteins with the highest diagnostic ability in distinguishing between TB disease and unlikely TB patients. FS-PLS was applied across 100 iterations, with a different training-test split (80:20) each time. Within each iteration, multiple signatures were discovered from the training set and the signature with the highest performance in terms of AUC in the test set was selected. At the end of the 100 iterations, a 4-protein signature, composed of Low affinity immunoglobulin gamma Fc region receptor III-B (FcR-10), Stromelysin-2 (MMP-10), C-X-C motif chemokine 11 (CXCL11) and C-C motif chemokine 4 (CCL4), was selected most frequently (37/100 iterations) and taken forward to cross-platform validation. Boxplots for each protein are shown in Figure 5.4.

Since the signature was selected 37 times in the cross-validation algorithm, the 4 independent variables of the statistical model (i.e., the 4 proteins) had different coefficients for each of the 37 iterations. A logistic regression statistical model, which is more appropriate for a binary outcome, was used to retrain the model and calculate the final coefficients for each protein. The wDRS values were used to plot the

Protein Name	UniProt ID	log ₂ FC	BH-adjusted p-value	AUC
S100A12	P80511	1.94	1.7x10 ⁻⁰³	83.3%
CXCL11	O14625	1.08	1.9x10 ⁻⁰²	78.0%
CXCL10	P02778	1.07	2.7x10 ⁻⁰²	74.9%
CXCL9	Q07325	0.99	1.8x10 ⁻⁰²	76.6%
S100-A9	Q07325	0.96	2.2x10 ⁻⁰²	75.5%
Kynureninase	Q16719	0.90	1.7x10 ⁻⁰²	77.0%
IgA	P01876/P01877	0.88	2.0x10 ⁻⁰³	81.8%
FcR-10	O75015	0.87	2.0x10 ⁻⁰³	80.5%
CAP-G	P40121	0.80	2.0x10 ⁻⁰³	81.5%
GPI	P06744	0.78	4.7x10 ⁻⁰³	73.4%
MHBD	Q99714	0.61	2.2x10 ⁻⁰²	76.2%
HtrA2	O43464	0.54	2.2x10 ⁻⁰²	75.6%
U-PAR	Q03405	0.52	3.1x10 ⁻⁰²	74.5%
CCL4	Q8NHW4	-0.51	2.8x10 ⁻⁰²	72.2%
PTN	P21246	-0.55	1.3x10 ⁻⁰²	77.5%
Contactin-2	Q02246	-0.57	2.0x10 ⁻⁰³	81.6%
Lumican	P51884	-0.58	1.8x10 ⁻⁰²	77.7%
PILR alpha	Q8IYJ0	-0.64	2.0x10 ⁻⁰³	82.2%
Dermatopontin	Q07507	-0.64	2.5x10 ⁻⁰³	73.8%
FAPalpha	Q12884	-0.65	2.2x10 ⁻⁰²	73.6%
MMP-10	P09238	-0.66	2.5x10 ⁻⁰²	77.5%
Cystatin-M	Q15828	-0.71	1.7x10 ⁻⁰³	82.7%
Thrombospondin-4	P35443	-0.79	1.9x10 ⁻⁰²	74.8%
Neurexin-1-beta	P58400	-0.86	1.3x10 ⁻⁰²	77.6%
IL-22RA2	Q969J5	-0.88	8.3x10 ⁻⁰³	79.0%
Osteomodulin	Q99983	-0.99	3.3x10 ⁻⁰³	79.1%

Table 5.2: List of the 26 SDA proteins, sorted by descending log₂FC, between microbiologically confirmed TB versus unlikely TB in the discovery SomaScan® dataset.

SDA = Significantly Differentially Abundant; FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve

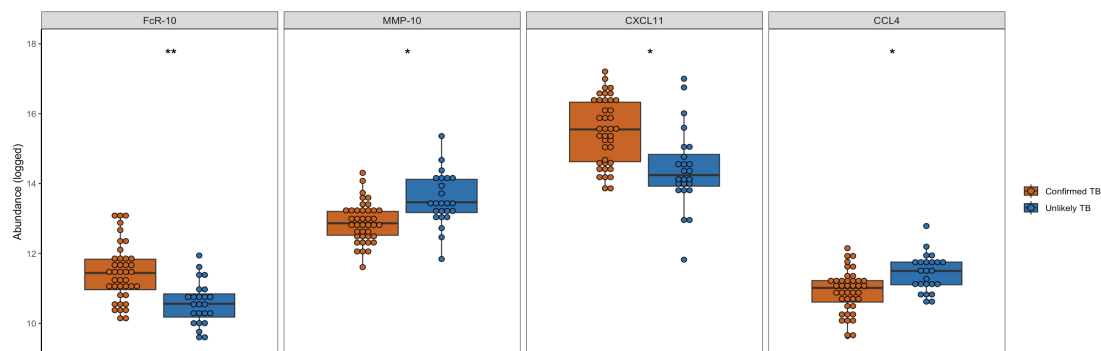


Figure 5.4: Logged abundance values of the 4 protein biomarkers identified in the SomaScan® dataset coloured according to disease group.

ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . All p -values are adjusted for multiple testing using Benjamini-Hochberg.

ROC curve and evaluate the performance of the signature in terms of AUC, sensitivity and specificity. The signature achieved an AUC of 96.7% [CI_{95%} 92.5%-100%], with a sensitivity and specificity of 90.0% [CI_{95%} 92.5%-97.2%] and 100% respectively [Figure 5.5].

Cross-platform and independent validation Cross-platform validation is critical to assess the transferability of the findings across different platforms, thus the 4 identified protein biomarkers were measured by immunoassay in the same samples ($n=64$) used for discovery. Independent validation was subsequently carried out on 515 samples not used for discovery to assess the transferability of

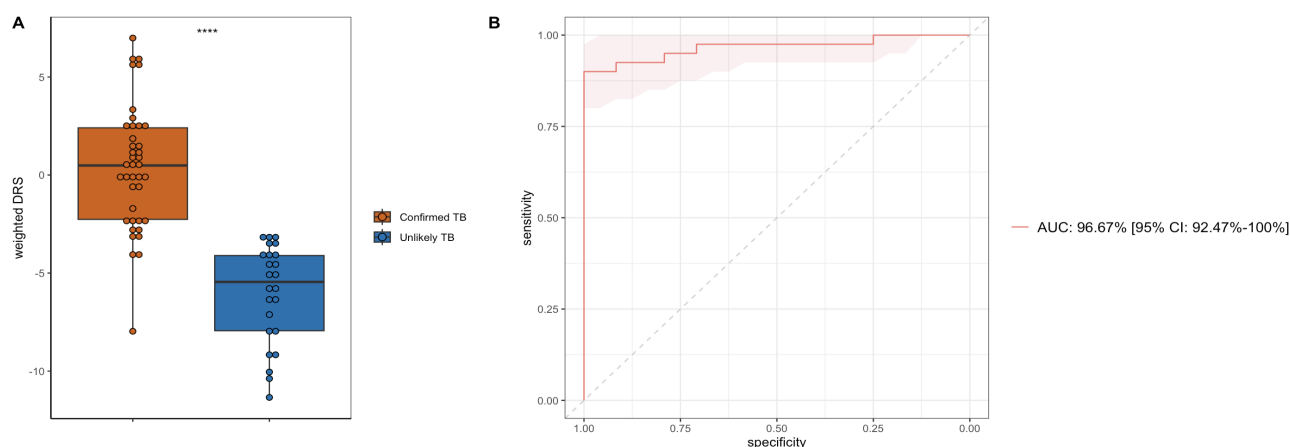


Figure 5.5: Boxplot of the wDRS of the 4-protein SomaScan® signature and its ROC curve.

All p -values are adjusted for multiple testing using BH. ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . wDRS = weighted Disease Risk Score; ROC = Receiver Operating Characteristic; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve; CI: Confidence Interval

the findings across different cohorts.

The 4 protein biomarkers taken forward to cross-platform validation were measured using either Luminex® or ELISA immunoassays according to the manufacturer's protocols. Luminex® was preferred over ELISA since it can be multiplexed and needs a smaller volume of biological samples, however, Low affinity immunoglobulin gamma Fc region receptor III-B (FcR-10) was measured by ELISA since it was not available via Luminex®. C-X-C motif chemokine 11 (CXCL11) and C-C motif chemokine 4 (CCL4) were measured in the same Luminex® kit since their dynamic ranges were compatible, Stromelysin-2 (MMP-10) was measured in a Luminex® single-plex kit. The Luminex® and ELISA data were combined subsequently. Since the aim of the cross-platform validation phase is to assess the transferability of the findings across different platforms, the 4 candidate protein biomarkers were measured in the 64 samples used in the discovery phase.

QC, including visual inspection of the standard curves and PCA, was performed to assess the quality of the data. The extrapolated concentration values were kept in the analysis, the values over the limit of detection were excluded, whilst the values under the limit of detection were assigned a value between zero and the lowest limit of detection as described in Section 2.6.3. The non-parametric Mann-Whitney U test was used to assess whether the median of the two comparative groups was statistically different, and $\log_2FC$, BH-adjusted p -value and AUC were calculated for each targeted protein [Table 5.3].

Single candidate biomarker proteins				
Protein Name	UniProt ID	$\log_2FC$	BH-adjusted p -value	AUC
FcR-10	O75015	1.25	2.2×10^{-06}	85.4%
CXCL11	O14625	0.92	1.4×10^{-02}	68.5%
CCL4	Q8NHW4	-0.90	4.0×10^{-03}	73.3%
MMP-10	P09238	-0.62	1.3×10^{-02}	69.4%
Combined 4-protein signature				
FcR-10, CXCL11, CCL4, MMP-10				91.7%

Table 5.3: $\log_2FC$, BH-adjusted p -value and AUC of the 4 protein biomarkers (individually and combined into a signature) identified in the SomaScan® discovery experiment when measured by immunoassays in the same biological samples.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve

The wDRS values were used to evaluate the performance of the signature. The 4 proteins combined achieved an AUC of 91.4% [CI_{95%} 84.6%-98.1%], a sensitivity of 75.0% [CI_{95%} 62.5%-87.5%] and a specificity of 91.7% [CI_{95%} 79.2%-100%].

The 4 identified proteins showed a good diagnostic performance not only when measured by the SomaScan® assay, but they kept their optimal diagnostic performance also when measured by immunoassays. The 4 proteins were finally measured in a larger cohort to assess their robustness and to determine whether the identified biomarkers had a similar diagnostic performance on a different set of samples. Samples in the independent validation cohort were randomised across 96-well plates and processed according to the manufacturer's protocol.

The extrapolated values were kept in the analysis, the values over the limit of detection were excluded, whilst the values under the limit of detection were assigned a value between zero and the lowest limit of detection. Standard curves were visually inspected to assess the quality of the data [Section 2.6.3]. PCA was performed to assess the presence of outlier samples or any source of technical variation. Outlier samples ($n=2$) were excluded and no source of variation was identified. Data were logged (base 2) and a linear model accounting for age and sex was used to test the significance of the proteins.

Table 5.4 lists log₂FC, BH-adjusted p -value and AUC of the 4 candidate biomarkers in the independent validation cohort.

Single candidate biomarker proteins				
Protein Name	UniProt ID	log ₂ FC	BH-adjusted p -value	AUC
FcR-10	O75015	0.80	1.7×10^{-13}	72.5%
CXCL11	O14625	0.79	4.0×10^{-07}	65.5%
CCL4	Q8NHW4	-0.36	4.4×10^{-03}	58.7%
MMP10	P09238	-0.57	1.1×10^{-08}	67.4%
Combined 4-protein signature				
FcR-10, CXCL11, CCL4, MMP-10				77.1%

Table 5.4: log₂FC, BH-adjusted p -value and AUC of the 4 protein biomarkers (individually and combined into a signature) identified in the SomaScan® dataset when measured by immunoassays in an independent set of samples.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve

All the proteins remained DA (i.e., log₂FC < -0.5 or log₂FC > 0.5 and BH-adjusted p -value < 0.05) apart from C-C motif chemokine 4 (CCL4), which had a significant BH-adjusted p -value but a log₂FC under the threshold. Stratification on HIV status showed that CCL4 was DA when comparing microbiologically confirmed TB versus unlikely TB in the HIV-uninfected samples. When the HIV-infected samples were included in the comparison, the difference in FC did not reach the threshold. This was not observed in the pilot and cross-validation SomaScan® experiments since the 64 samples were all collected from HIV-uninfected children. CCL4 is a protein that binds to C-C chemokine receptor type 5 (CCR5), which is one of the major HIV-suppressive factors produced by CD8⁺ T cells, and in HIV-infected patients it can be mainly found in its CCR5-ligand form to control HIV infection [298].

Luminex® dataset

Since the Luminex experiment included a considerable number of kits and plates, a graphical way to assess the number of indefinite or null concentration values was used. The OncoPrint function from the ComplexHeatmap R package [299] has been developed by the authors of the R package with the aim to

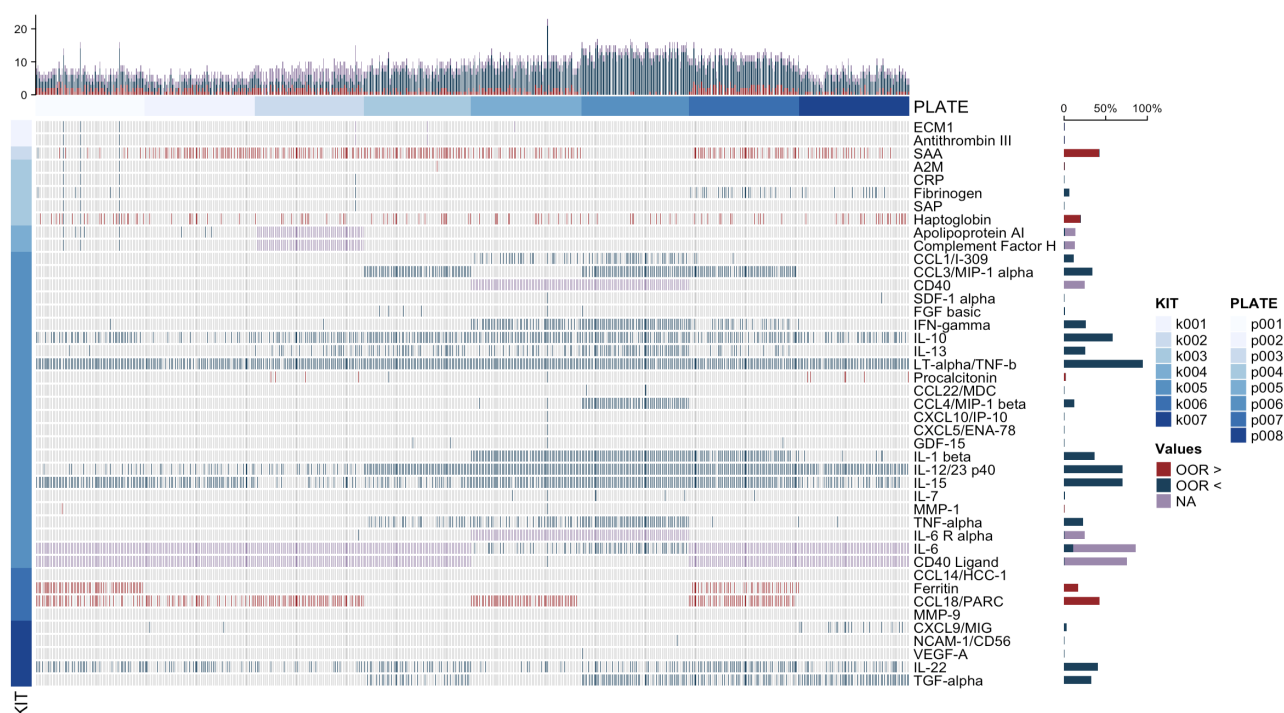


Figure 5.6: Overview of the indefinite values (under-the-limit-of-detection, over-the-limit-of-detection and NULL values) in the Luminex® experiment.

Samples are on the x-axis and analytes/proteins are on the y-axis. The different shades of blue in the top bar indicate the 8 plates in each Luminex® kit. The different shades of blue on the left-hand-side bar indicate the 7 Luminex® kits. Every single line (grey, red, blue or purple) is a concentration value for the correspondent sample-analyte pair. The red lines are the over-the-limit-of-detection values; the blue lines are the under-the-limit-of-detection values; the purple lines are the null values. The grey lines indicate a numerical value within the dynamic range. Broad summaries of the proportion of red, blue, purple and grey lines for each sample and analyte are plotted respectively at the top and right-hand side of the Figure.

OOR >: over-the-limit-of-detection value; OOR <: under-the-limit-of-detection value; NA: null

visualise multiple genomic alteration events by using a heatmap. In this thesis, the OncoPrint function was adapted in order to be used as an intuitive way to visualise, analyte by analyte and sample by sample, the total number of indefinite values generated in the experiment [Figure 5.6].

Some analytes, such as C-C motif chemokine 3 (CCL3), Interferon Gamma (IFN- γ), Interleukin-10 (IL-10), Interleukin-13 (IL-13), Lymphotoxin-alpha (LT-alpha), Interleukin-1 beta (IL-1 beta), Interleukin-12 (IL-12), Interleukin-15 (IL-15), Tumor necrosis factor (TNF-a), Interleukin-22 (IL-22) and Protransforming growth factor alpha (TGF-a), had a high number of under-the-limit-of-detection concentration values (i.e., OOR <) [Figure 5.6]. Since these values fall between zero and the lowest value of the dynamic range, in each plate and for each analyte, they were assigned either the lowest extrapolated value minus 10% or half of the lowest limit of detection if no values under the limit of detection were extrapolated. The analytes with the highest number of over-the-limit-of-detection values (i.e., OOR >) were Serum amyloid A-1 protein (SAA), Haptoglobin, Ferritin and C-C motif chemokine 18 (CCL18). Since the range the over-the-limit-of-detection values fall into is unknown, it is in fact between the highest limit of detection and infinity, a NULL value was assigned to them.

The list of analytes shown in Figure 5.6 includes 43 proteins rather than 41 as mentioned before in this section. Some of the plates were in fact ordered with two incorrect analytes (CD40 ligand rather than CD40 and IL6 Receptor alpha rather than IL6). These 4 proteins were therefore excluded from the analysis.

The analytes with *NULL* for over 10% of the observed concentration values were excluded from the analysis, along with 3 outlier samples identified by PCA. Once the outlier samples were excluded, a second PCA was performed to assess whether any source of technical variability was present in the data. The PCA plot (PC1 versus PC2) detected a plate effect, with plates 5 and 6, respectively p005 and p006 in Panel A of Figure 5.7, clustering towards the left of PC1 and the bottom of PC2. This variability due to technical differences between the sample plates was minimised [Figure 5.7 Panel C] by using the *normn.MA* function [238], described in section 2.5.

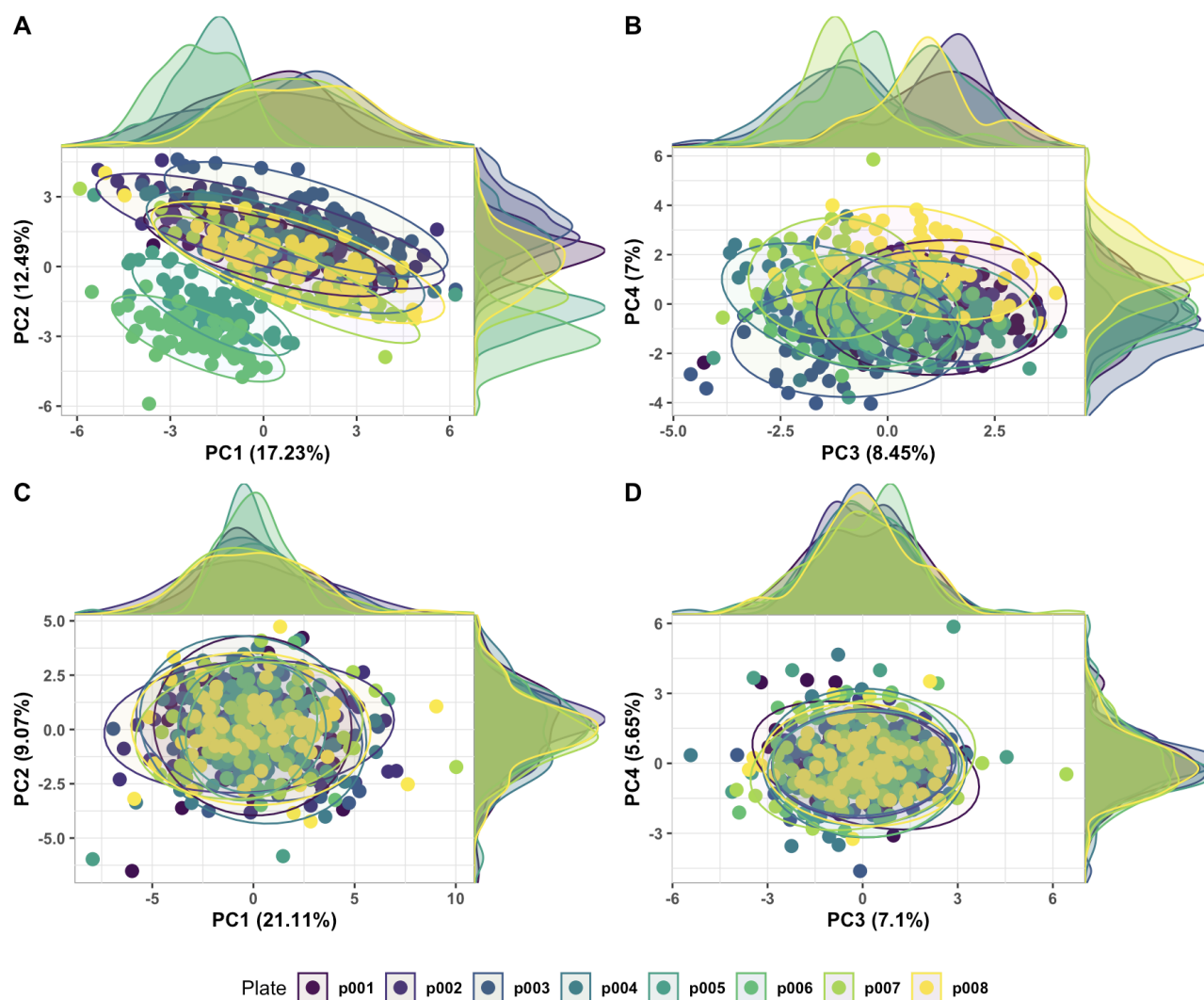


Figure 5.7: PCA biplots from PCA performed on the Luminex® dataset coloured according to sample plate. PC 1-2 and PC 3-4 are shown before [Panel A and Panel B] and after plate correction [Panel C and Panel D]. Points represent samples and are coloured according to sample plates. PCA = Principal Component Analysis; PC = Principal Component; p001 = sample plate 1; p002 = sample plate 2; p003 = sample plate 3; p004 = sample plate 4; p005 = sample plate 5; p006 = sample plate 6; p007 = sample plate 7; p008 = sample plate 8

After minimising the variability in the data due to technical differences across the sample plates, PCA was re-performed on the data. Since no other source of technical variability was detected, the dataset was combined with the SELDI and SomaScan® independent validation datasets to identify the best diagnostic protein markers for paediatric TB.

After data pre-processing and QC, including visual inspection of the standard curves, PCA and substitution of under-the-limit-of-detection values with a value between zero and the lowest limit of detection of each specific protein, the data were merged with the other two discovery datasets

(antibody validation of the SELDI and SomaScan® signatures).

Multi-platform integration

To overcome the important differences in terms of protein quantification and data generation between the SELDI, Somascan® and Luminex® assays, the SELDI and SomaScan® datasets were analysed individually and the candidate biomarkers identified in each of the two datasets were subsequently measured by immunoassays. Moreover, this reduced the search space to only the biomarkers with the strongest diagnostic ability in differentiating TB disease from unlikely TB in children. The three datasets used for the discovery of the signature were the following:

1. The validation by immunoassay of the 4 protein markers identified in the SELDI dataset
2. The validation by immunoassay of the 4 protein markers identified in the SomaScan® dataset
3. The 41 proteins measured by the Luminex® assay

342 samples were in common among the three immunoassay experiments and included in the discovery of novel protein biomarkers for distinguishing between TB disease and unlikely TB. The individual datasets were pre-processed separately and data from the 342 samples were combined. PCA was performed to assess if any source of technical variability was present [Figure 5.8]. The disease group was captured by PC1 with the microbiologically confirmed TB group towards the bottom right of PC1 and the unlikely TB group towards the top left of the plot [Figure 5.8; Panel A]. The unconfirmed TB patients lay in between the two other confirmed disease categories. HIV status was captured by PC1, with the HIV-infected group towards the top right of the plot [Figure 5.8 Panel B]. Age was captured by PC3, with the younger children towards the higher values of PC3 and the older children towards the lower values of PC3 [Figure 5.8; Panel J]. No differences were observed in terms of recruitment site [Figure 5.8; Panels E, F] or sex [Figure 5.8; Panels G, H].

DA analysis was performed to identify proteins that were SDA. Protein abundances were logged (base 2) and a linear model accounting for age and sex was used to identify the DA proteins between the microbiologically confirmed TB and unlikely TB groups. The unconfirmed TB group was not included in the biomarker identification phase since the final diagnosis is uncertain. Unconfirmed TB is, in fact, diagnosed through clinical examination but does not have microbiological confirmation of the disease and might introduce biases in discovery studies. Proteins with $\log_2FC > 0.5$ or $\log_2FC < -0.5$ and BH-adjusted p -value < 0.05 were considered SDA between the two comparative groups. 11 proteins (9 more abundant in microbiologically confirmed TB and 2 more abundant in the unlikely TB group) were SDA between the two disease groups of interest [Table 5.5].

Iterative FS-PLS with cross-validation [Section 2.5] was performed on the microbiologically confirmed TB and unlikely TB samples and on the 11 SDA proteins to identify the combination of protein biomarkers with the highest diagnostic performance. Iterative FS-PLS was run across 100 iterations, with a different training-test split (at 80:20) used each time. Within each iteration, the signature discovered in the training set and with the highest AUC in the test set was selected. The signature taken forward to independent validation was the signature selected more frequently across the 100 iterations (36/100 iterations). The signature was composed of C-C motif chemokine 1 (CCL1), Stromelysin-2 (MMP-10), C-C motif chemokine 22 (CCL22) and Low affinity immunoglobulin gamma Fc region receptor III-B (FcR-10). A logistic regression model, which is more appropriate for a binary outcome,

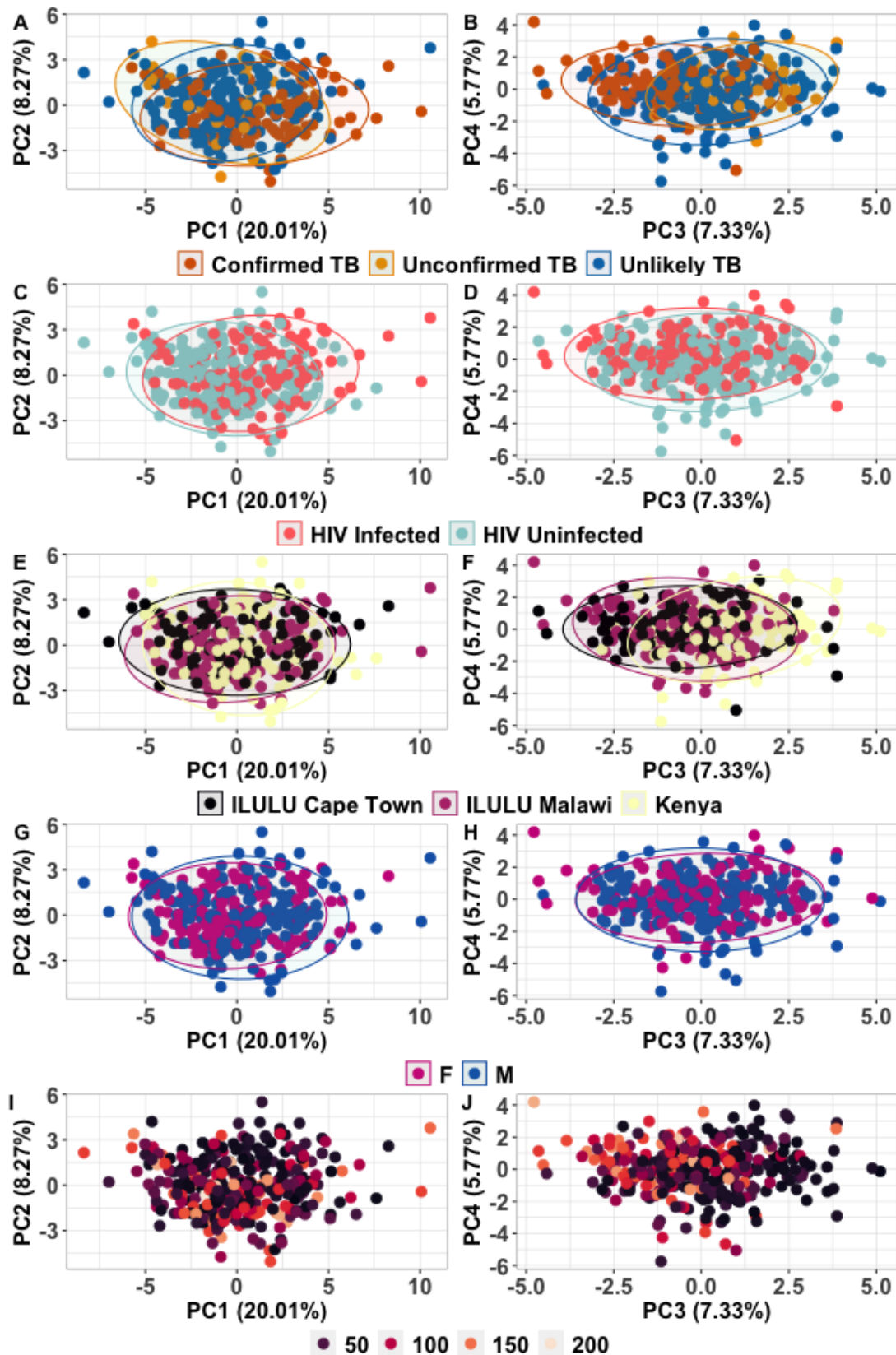


Figure 5.8: PCA biplots from PCA performed on the proteomic discovery dataset coloured according to clinical variables.

PC 1-2 and PC 3-4 are shown. Points represent samples and are coloured according to disease group [Panel A and Panel B], HIV status [Panel C and Panel D], recruitment site [Panel E and Panel F], sex [Panel G and Panel H] and age [Panel I and Panel J]. PCA = Principal Component Analysis; PC: Principal Component; HIV: Human Immunodeficiency Virus; F = female; M = male

Protein Name	UniProt ID	log ₂ FC	BH-adjusted <i>p</i> -value	AUC
CCL1	P22362	2.81	4.1×10^{-20}	82.9%
FcR-10	O75015	0.83	3.0×10^{-10}	73.2%
CXCL9	Q07325	0.78	5.7×10^{-10}	72.6%
CXCL10	P02778	1.15	4.8×10^{-07}	68.7%
MMP-10	P09238	-0.53	1.9×10^{-05}	66.1%
IFN- γ	P01579	0.68	6.5×10^{-05}	65.0%
CRP	P02741	1.31	1.2×10^{-04}	64.4%
CXCL11	O14625	0.72	1.4×10^{-04}	64.3%
MMP-1	P03956	0.57	7.1×10^{-04}	62.7%
CCL22	O00626	-0.68	2.1×10^{-03}	61.5%
IL-7	P13232	0.53	4.7×10^{-03}	60.6%

Table 5.5: log₂FC, BH-adjusted *p*-value and AUC of the 11 differentially abundant protein in the merged discovery dataset sorted by increasing BH-adjusted *p*-value.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve

was then used to retrain the model and calculate the final coefficients. Boxplots for each protein are shown in Figure 5.9.

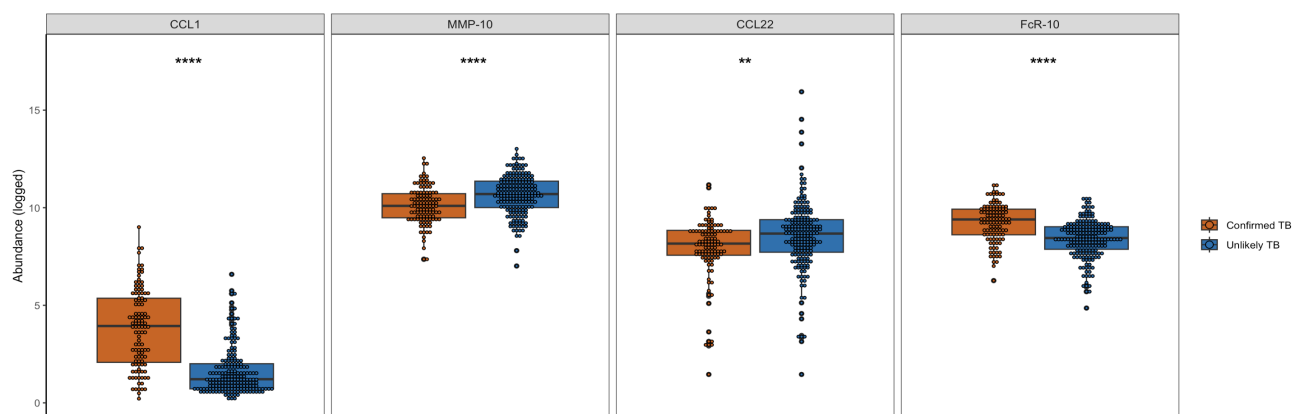


Figure 5.9: Logged abundance values of the 4 protein biomarkers identified in the discovery dataset coloured according to disease group.

All *p*-values are adjusted for multiple testing using Benjamini-Hochberg. ns: *p*-value >0.05; *: *p*-value <0.05; **: *p*-value ≤0.01; ***: *p*-value ≤0.001; ****: *p*-value ≤0.0001.

The wDRS values were used for calculating the AUC, sensitivity and specificity of the 4-protein signature. The signature achieved an AUC of 87.0% [CI_{95%} 82.9%-91.2%] and sensitivity and specificity respectively of 72.3% [CI_{95%} 63.4%-80.4%] and 88.1% [CI_{95%} 83.4%-92.8%] in discriminating microbiologically confirmed TB versus unlikely TB [Figure 5.10].

The 4-protein diagnostic signature identified in the discovery phase was then taken forward to validation in an independent set of paediatric samples that were not used for discovery.

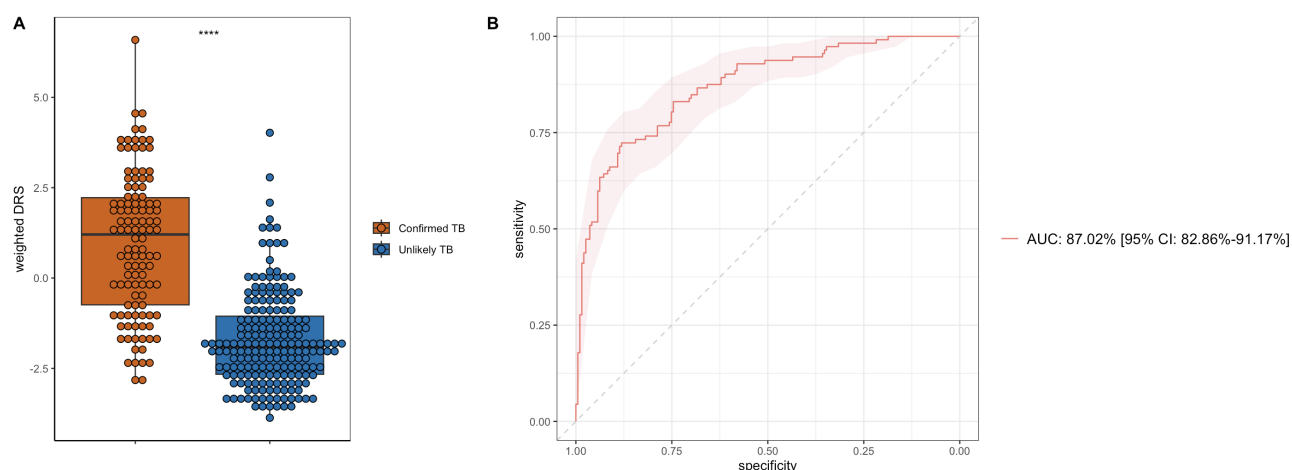


Figure 5.10: Boxplot of the wDRS of the 4-protein signature and its ROC curve in the discovery cohort. ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . wDRS = weighted Disease Risk Score; ROC = Receiver Operating Characteristic; AUC = Area Under the ROC Curve; CI: Confidence Interval

5.3.2 Validation phase

The 4-protein signature identified in the discovery phase was composed of: C-C motif chemokine 1 (CCL1), Stromelysin-2 (MMP-10), C-C motif chemokine 22 (CCL22) and Low affinity immunoglobulin gamma Fc region receptor III-B (FcR-10). Whilst Low affinity immunoglobulin gamma Fc region receptor III-B (FcR-10) was quantified using the ELISA assay since it was not available via Luminex®, C-C motif chemokine 1 (CCL1), Stromelysin-2 (MMP-10) and C-C motif chemokine 22 (CCL22) were quantified using the Luminex® assay. C-C motif chemokine 1 (CCL1) and C-C motif chemokine 22 (CCL22) were measured in the same Luminex® kit since they had similar dynamic ranges. Stromelysin-2 (MMP-10) was quantified in a single-plex Luminex® kit.

Pre-processing and analysis of the Luminex® and ELISA data were carried out as described in Section 5.2.2. In brief, the standard curves for each analyte in each plate were visually examined. The values extrapolated by the Luminex® or ELISA software were kept in the analysis. The concentration values under the limit of detection (i.e., OOR $<$) were assigned either the lowest extrapolated value minus 10% or half of the lowest limit of detection if no values under the limit of detection were extrapolated. None of the quantifications was over the limit of detection. PCA was performed to assess the presence of any technical variability in the data [Figure 5.11].

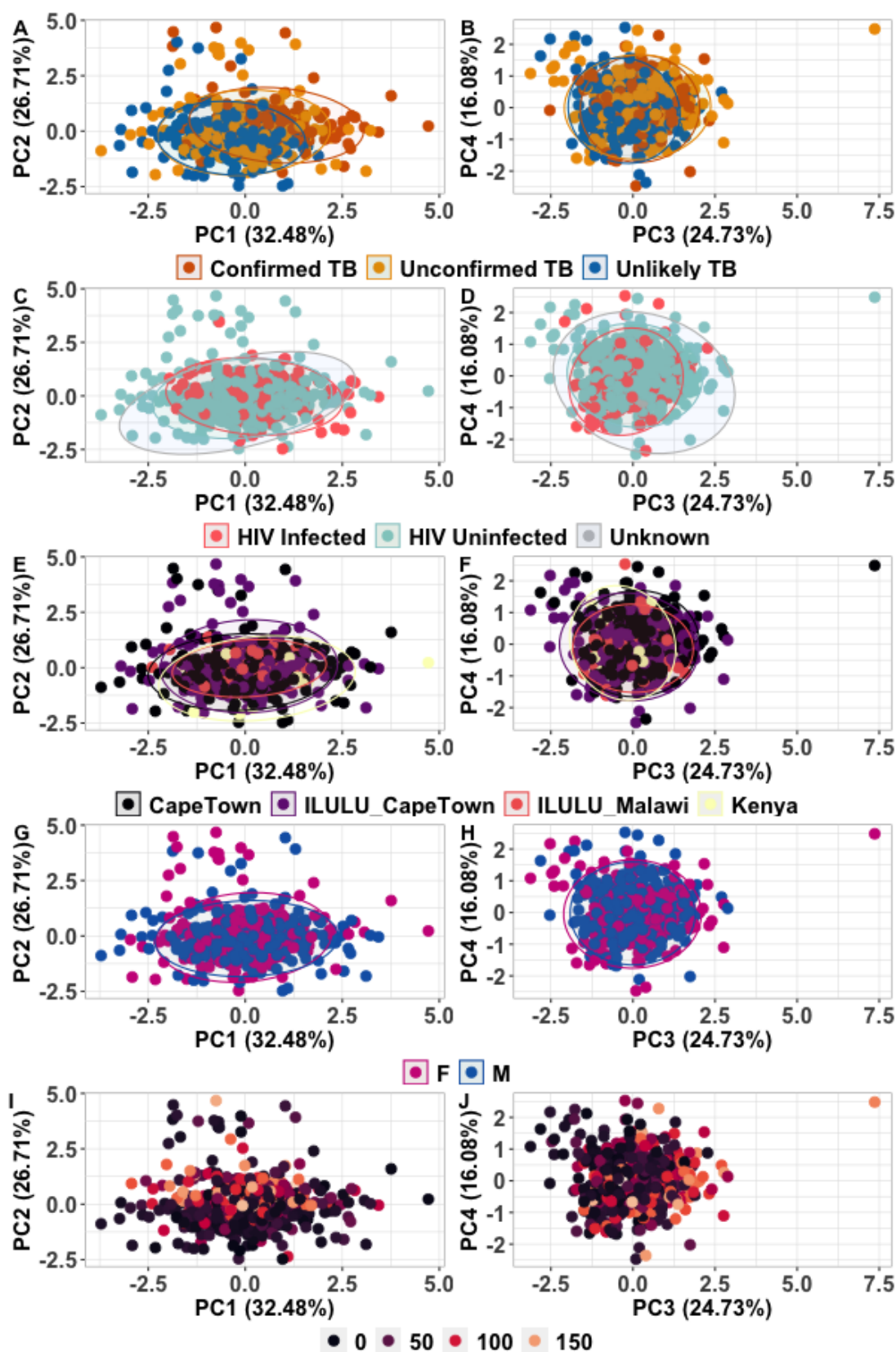


Figure 5.11: PCA biplots from PCA performed on the proteomic validation dataset coloured according to clinical variables.

PC 1-2 and PC 3-4 are shown. Points represent samples and are coloured according to disease group [Panel A and Panel B], HIV status [Panel C and Panel D], recruitment site [Panel E and Panel F], sex [Panel G and Panel H] and age [Panel I and Panel J]. PCA = Principal Component Analysis; PC: Principal Component; HIV: Human Immunodeficiency Virus; F = female; M = male

The protein concentrations were then logged ($\log_2$) and a linear model accounting for age and sex

was used to compare the abundance of the 4 proteins between the two groups of interest (i.e., microbiologically confirmed TB and unlikely TB). The BH-adjusted p -values and $\log_2FC$ were calculated for each protein and listed in Table 5.6. C-C motif chemokine 1 (CCL1), Stromelysin-2 (MMP-10) and Low affinity immunoglobulin gamma Fc region receptor III-B (FcR-10) were DA in this comparison, however, C-C motif chemokine 22 (CCL22) was not significant (i.e., BH-adjusted p -value >0.05). Boxplots for each protein are shown in Figure 5.12.

Single candidate biomarker proteins				
Protein Name	UniProt ID	$\log_2FC$	BH-adjusted p -value	AUC
CCL1	P22362	2.12	1.6×10^{-14}	74.6%
MMP-10	P09238	-0.49	5.5×10^{-05}	65.0%
CCL22	O00626	-0.05	0.24	54.3%
FcR-10	O75015	1.0	9.4×10^{-23}	82.6%
Combined 4-protein signature				
CCL1, MMP-10, CCL22, FcR-10				88.7%

Table 5.6: $\log_2FC$, BH-adjusted p -value and AUC of the 4 candidate protein biomarkers (individually and combined into a signature) in the validation dataset.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve; CI: Confidence Interval.

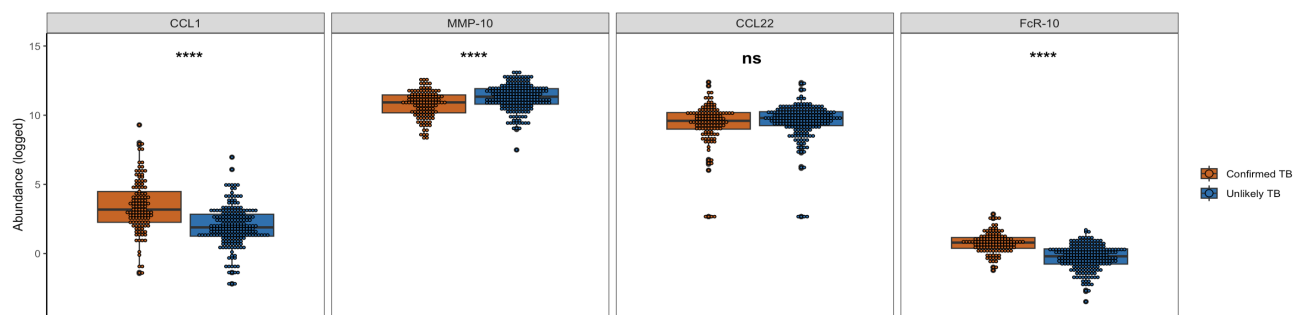


Figure 5.12: Logged abundance values of the 4 protein biomarkers in the validation dataset coloured according to disease group.

All p -values are adjusted for multiple testing using Benjamini-Hochberg. ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 .

When the statistical additive model identified in the discovery phase was used, the signature achieved an AUC of 82.7% [CI_{95%} 77.8%-87.5%] and sensitivity and specificity respectively of 77.7% [CI_{95%} 69.4%-85.1%] and 77.4% [CI_{95%} 71.6%-83.2%] [Figure 5.13] in the independent validation cohort. When the logistic regression model was retrained and new coefficients were calculated, the signature achieved an AUC of 88.7% [CI_{95%} 85.0%-92.4%] and sensitivity and specificity respectively of 78.5% [CI_{95%} 71.1%-85.6%] and 86.1% [CI_{95%} 81.3%-90.4%] [Figure 5.13]. A simple DRS was also used, subtracting the total abundance of features that decreased (MMP10 and CCL22) from the total abundance of features that increased (CCL1 and FcR-10) in the microbiologically confirmed TB for each individual. This model achieved an AUC of 85.2% [CI_{95%} 80.8%-89.5%] and sensitivity and specificity respectively of 77.7% [CI_{95%} 70.3%-85.1%] and 84.1% [CI_{95%} 78.9%-88.9%] [Figure 5.13]. When wDRSs and simple DRS were calculated in the unconfirmed TB group, the median was in between the microbiologically confirmed and unlikely TB groups [Figure 5.13 Panel A].

Table 5.7 shows the performance of the signature when stratifying by HIV status, site of TB infection and cohort. The signature achieved better performance in children diagnosed with any forms of EPTB. In this comparison, the sensitivity was 100%, meaning that all the positive cases (i.e., TB disease) were

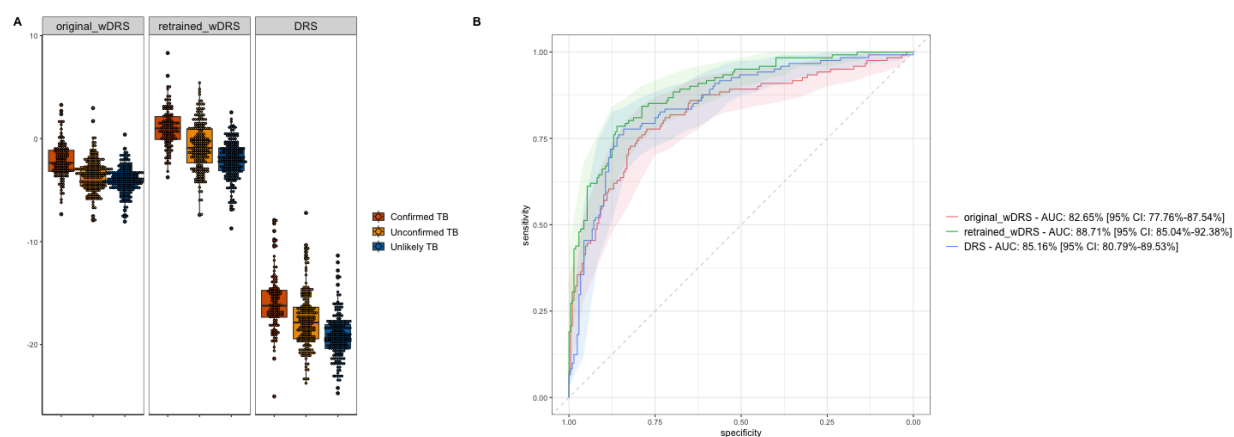


Figure 5.13: Boxplot of the original wDRS, retrained wDRS and the simple DRS of the 4-protein signature and their ROC curves in the validation cohort.

ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . wDRS = weighted Disease Risk Score; DRS = Disease Risk Score; ROC = Receiver Operating Characteristic; AUC = Area Under the ROC Curve; CI: Confidence Interval

diagnosed correctly. When the candidate protein biomarkers were combined into a simple DRS, the performance of the signature in the HIV-infected and -uninfected patients was very similar. When a wDRS was calculated, the 4-protein signature performed better in differentiating TB from unlikely TB in HIV-infected children. The signature performance was lower in the Malawian cohort, whilst it was similar among the other cohorts.

Original wDRS			
	AUC [CI _{95%}]	Sensitivity [CI _{95%}]	Specificity [CI _{95%}]
HIV-uninfected	81.7% [CI _{95%} 75.8%-87.7%]	76.2% [CI _{95%} 66.7%-85.7%]	75.7% [CI _{95%} 68.8%-82.6%]
HIV-infected	84.6% [CI _{95%} 75.3%-93.9%]	82.4% [CI _{95%} 67.7%-94.1%]	85.7% [CI _{95%} 77.8%-93.7%]
PTB	79.5% [CI _{95%} 72.8%-86.1%]	73.1% [CI _{95%} 62.7%-83.6%]	77.4% [CI _{95%} 71.6%-82.7%]
PTB+EPTB	87.1% [CI _{95%} 80.4%-93.7%]	82.2% [CI _{95%} 71.1%-93.3%]	82.7% [CI _{95%} 77.4%-87.5%]
EPTB	88.0% [CI _{95%} 75.6%-100.0%]	100%	70.7% [CI _{95%} 64.4%-76.9%]
South Africa ILULU	85.7% [CI _{95%} 78.1%-93.3%]	72.7% [CI _{95%} 60.4%-85.4%]	87.9% [CI _{95%} 79.3%-96.6%]
Malawi ILULU	68.3% [CI _{95%} 49.6%-87.1%]	71.4% [CI _{95%} 50.0%-92.9%]	73.0% [CI _{95%} 56.8%-86.5%]
South Africa NIH	85.7% [CI _{95%} 79.4%-92.0%]	83.3% [CI _{95%} 72.2%-92.6%]	77.3% [CI _{95%} 69.1%-85.6%]
Kenya	77.5% [CI _{95%} 44.23%-100.0%]	80.0% [CI _{95%} 40.0%-100.0%]	87.5 [CI _{95%} 68.7%-100.0%]
Retrained wDRS			
	AUC [CI _{95%}]	Sensitivity [CI _{95%}]	Specificity [CI _{95%}]
HIV-uninfected	88.3% [CI _{95%} 83.9%-92.7%]	81.0% [CI _{95%} 71.4%-89.3%]	82.6% [CI _{95%} 76.4%-88.9%]
HIV-infected	91.6% [CI _{95%} 85.6%-97.5%]	82.4% [CI _{95%} 67.7%-94.1%]	88.9% [CI _{95%} 81.0%-95.2%]
PTB	86.1% [CI _{95%} 81.0%-91.3%]	80.6% [CI _{95%} 70.2%-89.6%]	78.9% [CI _{95%} 73.6%-84.6%]
PTB+EPTB	91.4% [CI _{95%} 86.8%-96.0%]	82.2% [CI _{95%} 71.1%-93.3%]	86.1% [CI _{95%} 81.3%-90.4%]
EPTB	92.7% [CI _{95%} 86.5%-98.9%]	100.0%	87.0% [CI _{95%} 82.2%-91.4%]
South Africa ILULU	90.6% [CI _{95%} 84.7%-96.5%]	83.3% [CI _{95%} 72.9%-93.8%]	87.9% [CI _{95%} 79.3%-94.8%]
Malawi ILULU	80.5% [CI _{95%} 66.4%-94.6%]	85.7% [CI _{95%} 64.3%-100.0%]	70.3% [CI _{95%} 56.8%-83.8%]
South Africa NIH	90.2% [CI _{95%} 85.5%-95.0%]	79.6% [CI _{95%} 68.5%-90.7%]	85.6% [CI _{95%} 78.4%-91.8%]
Kenya	96.3% [CI _{95%} 89.0%-100.0%]	100.0%	87.5 [CI _{95%} 68.8%-100.0%]
DRS			
	AUC [CI _{95%}]	Sensitivity [CI _{95%}]	Specificity [CI _{95%}]
HIV-uninfected	85.8% [CI _{95%} 80.7%-90.9%]	76.2% [CI _{95%} 66.7%-84.5%]	86.1% [CI _{95%} 80.5%-91.7%]
HIV-infected	85.1% [CI _{95%} 76.4%-93.7%]	82.4% [CI _{95%} 67.7%-94.1%]	82.5% [CI _{95%} 73.0%-90.5%]
PTB	83.8% [CI _{95%} 77.9%-89.6%]	74.6% [CI _{95%} 64.2%-85.1%]	86.1% [CI _{95%} 81.3%-86.1%]
PTB+EPTB	86.7% [CI _{95%} 80.9%-92.5%]	80.0% [CI _{95%} 66.7%-91.1%]	84.1% [CI _{95%} 78.9%-88.9%]
EPTB	88.3% [CI _{95%} 80.2%-96.5%]	100%	78.9% [CI _{95%} 72.6%-84.1%]
South Africa ILULU	88.1% [CI _{95%} 81.3%-94.9%]	72.9% [CI _{95%} 60.4%-85.4%]	89.7% [CI _{95%} 81.0%-96.6%]
Malawi ILULU	75.1% [CI _{95%} 58.0%-92.2%]	78.6% [CI _{95%} 57.1%-100.0%]	78.4% [CI _{95%} 64.9%-91.9%]
South Africa NIH	85.7% [CI _{95%} 79.5%-91.9%]	79.6% [CI _{95%} 68.5%-88.9%]	86.6% [CI _{95%} 79.4%-92.8%]
Kenya	87.5% [CI _{95%} 70.7%-100.0%]	80.0% [CI _{95%} 40.0%-100.0%]	87.5 [CI _{95%} 68.8%-100.0%]

Table 5.7: Proteomic signature performance in clinical subgroups.

CI: Confidence Interval; HIV: Human Immunodeficiency Virus; PTB: Pulmonary Tuberculosis; EPTB: Extra-Pulmonary Tuberculosis

5.3.3 Biological functions of the 4 proteins

Proteomic diagnostic biomarkers are usually identified in biological samples such as serum, plasma, urine or saliva. The whole set of proteins expressed by cells or tissues of an organism is defined as proteome; the subset of proteins synthesised in a cell and secreted outside the cell is called secretome. Similar to the proteome, the secretome profile is highly dynamic. It changes, in fact, in organ and tissue development and maturation and in response to environmental stimuli or pathologies [300]. The secreted proteins reflect the various stages of pathological conditions and cell-cell communication. They are therefore a source of biomarkers for diagnosis, prognosis or therapy monitoring [300]. Furthermore, secreted proteins are detectable in blood with no extra steps needed for lysing the cells, easing the pipeline to measure the proteins of interest in a potential protein biomarker-based POCT.

Of the 4 proteins in the signature, 2 were more abundant in TB disease, C-C motif chemokine 1 (CCL1) and Low affinity immunoglobulin gamma Fc region receptor III-B (FcR-10), and 2 were more abundant in unlikely TB, Stromelysin-2 (MMP-10) and C-C motif chemokine 22 (CCL22). All of them are known to be involved in host immunity.

C-C motif chemokine 1 (CCL1), which is more abundant in TB disease compared to unlikely TB, is

a glycoprotein secreted by activated monocytes/macrophages, T lymphocytes and endothelial cells [301]. It is therefore involved in the immune system processes, inflammatory response and signalling. CCL1 binds to C-C chemokine receptor type 8 (CCR8), which is expressed in monocytes/macrophages, T helper 2 cells (Th2) and regulatory T lymphocytes [302]. CCL1 has been identified as a potential protein biomarker for TB disease in previous paediatric or adult studies [188, 303–305].

Stromelysin-2 (MMP-10), which is less abundant in TB disease compared to unlikely TB, is a secreted protein of the matrix metalloproteinases (MMPs) family. MMP proteins are involved in the breakdown of extracellular matrix in normal physiological or disease processes. They are considered to play a crucial role in TB since the *Mtb* bacilli are released into the airways after the breakdown of the lung extracellular matrix [306]. MMP-10 has not been well studied in TB. A recent study showed that MMP-10 was more abundant in induced sputum and BronchoAlveolar Lavage (BAL) samples in adults with TB disease compared to unlikely TB [307]. MMP-10 was consistently less abundant in children with TB disease compared to children with unlikely TB in our SomaScan® dataset ($\log_2\text{FC} = -0.66$, BH-adjusted $p\text{-value} = 2.6e^{-02}$), its cross-platform validation by immunoassays ($\log_2\text{FC} = -0.62$, BH-adjusted $p\text{-value} = 1.3e^{-02}$), its validation in a larger set of samples ($\log_2\text{FC} = -0.57$, BH-adjusted $p\text{-value} = 1.1e^{-08}$) [Section 5.2.2] and, finally, in the independent validation of the proteomic diagnostic signature ($\log_2\text{FC} = -0.49$, BH-adjusted $p\text{-value} = 3.8e^{-05}$) [Section 5.2.2]. The difference in abundance of MMP-10 when contrasting TB disease versus unlikely TB in adults and children might be due to the different biological samples used in the studies (BAL and induced sputum versus serum), or the different pathogenesis of TB disease in children and adults.

C-C motif chemokine 22 (CCL22), which is less abundant in TB disease compared to unlikely TB, is a protein secreted by dendritic cells and macrophages. It plays a role in the immune system processes, inflammatory response and signalling. CCL22 was shown to be significantly less abundant in a HIV-infected adult cohort with TB disease compared to unlikely TB [308]. This protein has not been investigated in paediatric TB so far. It was selected as part of the proteomic diagnostic signature from the Luminex® experiment, and it was also differentially abundant ($\log_2\text{FC} = -0.44$, BH-adjusted $p\text{-value} = 3.5e^{-02}$) in our SomaScan® dataset described in Section 5.2.2.

Low affinity immunoglobulin gamma Fc region receptor III-B (FcR-10), which is more abundant in TB disease compared to unlikely TB, is a secreted protein involved in immune system processes and signalling. It is expressed by granulocytes (neutrophils, mast cells and eosinophils). FcR-10 plays an important role in both promoting and regulating the immune and inflammatory response to immune complexes [309]. Its role in TB has not been investigated so far. FcR-10 was consistently more abundant in children with TB disease compared to children with unlikely TB in our SomaScan® dataset ($\log_2\text{FC} = 0.87$, BH-adjusted $p\text{-value} = 8.9e^{-06}$), its cross-platform validation by immunoassays ($\log_2\text{FC} = 1.25$, BH-adjusted $p\text{-value} = 2.2e^{-06}$), its validation in a larger set of samples ($\log_2\text{FC} = 0.8$, BH-adjusted $p\text{-value} = 1.7e^{-13}$) and, finally, in the independent validation of the proteomic diagnostic signature ($\log_2\text{FC} = 0.83$, BH-adjusted $p\text{-value} = 3.0e^{-10}$) [Section 5.2.2].

5.4 Summary

Novel and rapid diagnostic tests for TB are urgently needed, especially in children where the diagnosis is more complex than in adults. Children have in fact low bacteria loads, which causes many cases of childhood TB to go underdiagnosed or missed. Moreover, collecting biological samples from paediatric

patients can be challenging and often requires invasive procedures, such as induced sputum, gastric aspiration or biopsy. A biomarker-based diagnostic test would provide many benefits. Firstly, it would be much faster at providing results compared to other approaches currently used. In addition, because a host protein diagnostic signature would capture the changes in protein abundances in the blood, it would be able to detect the presence of *Mtb* regardless of the infection site.

In this chapter, I have used proteomic datasets to identify protein candidate biomarkers for paediatric TB. Firstly, SELDI and SomaScan® platforms were used to measure a larger pool of proteins in the discovery phase. The two datasets were processed and analysed separately, and a diagnostic signature for each dataset was identified. The two signatures were then validated by immunoassay and the data were merged with a third dataset generated by a Luminex® assay including quantifications of 41 analytes. A 4-protein diagnostic signature was identified by FS-PLS and validated in an independent cohort. The final signature was composed of C-C motif chemokine 1 (CCL1), Stromelysin-2 (MMP-10), C-C motif chemokine 22 (CCL22) and Low affinity immunoglobulin gamma Fc region receptor III-B (FcR-10). In the independent validation cohort, the signature achieved an AUC of 82.2% [CI_{95%} 77.2%-87.3%], which was improved to 88.7% [CI_{95%} 84.9%-92.4%] when the model weights were retrained. When a simple DRS was calculated, the AUC achieved 85.1% [CI_{95%} 80.6%-89.6%]. Whilst the wDRS with the retrained model weights had the best diagnostic performance, the simple DRS would be probably the best choice to take forward to the next steps (i.e., POCT development and clinical trial) since there is less probability of overfitting.

The work presented in this chapter has some limitations. Whilst this is a multi-country study, the cohorts included in the discovery and validation phases were from different African countries. Ideally, a novel biomarker-based diagnostic test for TB should have optimal diagnostic performance globally. Further validation in new cohorts from other countries is therefore needed. Discovery studies are usually based on untargeted techniques to be able to identify the candidate biomarkers from all the proteins present in the biological samples analysed. Although a SELDI experiment was carried out, just 4 identified candidate biomarkers were taken forward to the discovery and validation phases. The use of a limited pool of proteins in a discovery study might leave out good diagnostic biomarkers.

This multi-platform, multi-cohort study has identified, for the first time, a host protein signature to distinguish microbiologically confirmed TB versus unlikely TB. Further steps are now required to combine the 4 proteins into a prototype test. After its development, further clinical studies are needed to compare the result of the newly developed POCT to the current gold standard. The identification of the host protein signature for paediatric TB is therefore a very important first step towards a new, rapid and cheap diagnostic test for this disease.

To conclude, in this chapter, I have presented my work on the identification of a novel host protein signature for TB disease in children with optimal diagnostic performance. This might pave the way towards the development of a POCT for paediatric TB and reduce the TB cases worldwide, especially in LMIC where this disease is endemic.

CHAPTER 6

INTEGRATING TRANSCRIPTOMIC AND PROTEOMIC DATASETS TO IDENTIFY MULTI-OMICS DIAGNOSTIC BIOMARKERS FOR PAEDIATRIC TUBERCULOSIS

Modern high-throughput technologies measure a large number of biological molecules simultaneously. However, considering just one single layer of -omics may be missing the complex relationships between molecular layers and pathways. Statistical methods have been mainly applied to single types of -omics to identify sparse signatures able to explain or predict biological conditions. Although various single-omics signatures to diagnose TB disease in children have been identified, none of them meets the TPPs published by the WHO in 2014 [44] for a biomarker-based POCT. Since multi-omics datasets capture the heterogeneity of the biological system, we ought to explore whether diagnostic multi-omics signatures have the potential to overperform single-omics biomarkers.

In Chapters 3 and 5, I presented the discovery and validation of a sparse transcriptomic and proteomic signature respectively, both with very high accuracy in diagnosing TB disease in children. In this chapter, I present a novel approach to identifying and evaluating the performance of a multi-omics signature for diagnosing TB in children.

6.1 Overview and chapter aims

Multi-omics analysis considers molecules from more than one functional level (i.e., transcriptomic, proteomics, metabolomics). Whilst the scientific community has frequently adopted a single-omics approach for biomarker discovery, the complex nature of a biological system requires an integrative and multidisciplinary approach to better characterise the relationships between the different layers of -omics. This is hugely important since the host undergoes various changes during infection at the genomic, transcriptomic, proteomic and metabolomic levels [310]. To obtain a more complete picture of the biological system, holistic approaches are usually used. A holistic approach considers several variables simultaneously to capture and present the complete systems rather than focusing on single types of variables. This approach is referred to as multi-omics integration, which refers to combining single-omics datasets into a dataset with quantification of more than one type of biological molecule. Recently, platforms able to measure gene expression and protein abundance simultaneously have been developed. NanoString technology (Seattle, WA, USA) offers a multiplex platform targeting both genes and proteins. Multi-omics POCTs are therefore achievable, although more expensive and less suitable for LMIC.

The main aims of the analysis described in this chapter are:

1. To identify a sparse multi-omics signature able to differentiate between TB disease and unlikely TB in children
2. To compare the diagnostic performance of the newly identified multi-omics signature to the

diagnostic performance of a transcriptomic and a proteomic diagnostic signature identified in the same cohort

The structure of this chapter is as follows. Section 6.2 describes the case definitions, clinical cohorts, data integration and analysis. Results are introduced in Section 6.3 and summarised and discussed in Section 6.4. Biological insights are outlined in Section 6.5.

6.2 Cohorts and study design

To identify multi-omics biomarkers for paediatric TB, the transcriptomic and proteomic discovery datasets described in Chapters 3 and 5 were analysed. Host whole blood transcriptomic profiles from RNA-Seq and measurements of protein abundances generated by ELISA and Luminex® assays were included in the multi-omics integration analysis. Only the samples in common between the transcriptomic and proteomic datasets were kept in the analysis. Since a validation dataset was not available, the samples were randomly split into training and test sets (80:20 split) to limit overfitting. DE and DA analyses were performed on the transcriptomic and proteomic datasets respectively and, subsequently, FS-PLS was performed to identify a transcriptomic and a proteomic signature in the two training sets. The two single-omics training sets were then integrated. MixOmics [203] and FS-PLS were used to identify the set of genes and proteins that combined into a diagnostic signature achieved the optimal performance in discriminating between TB disease and unlikely TB in children. The performances of the transcriptomic, proteomic and multi-omic signatures were then evaluated in the respective test sets and compared.

The subset of the cohort analysed here included 195 children (<15 years of age) recruited in South Africa, Malawi or Kenya during the studies described in Section 2.4. The children were diagnosed with microbiologically confirmed TB ($n=69$) or unlikely TB ($n=126$), with or without HIV-infection. The microbiologically confirmed TB disease group included children with PTB or EPTB. The clinical features of the children in the whole cohort, training set and test set are shown in Table 6.1. Overall, children with confirmed TB were older (median age 53.0 [IQR 21.0-108.0]) than children with unlikely TB (median age 31.0 [IQR 13.0-79.5]). The percentage of females was higher in the confirmed TB group (53.6%) compared to the unlikely TB group (42.9%). The percentage of HIV-infected children was 46% in both disease groups. Most of the microbiologically confirmed TB cases had PTB, whilst 45.2% of the unlikely TB cases had a respiratory tract infection. Confirmed TB children were older (training set: median age 53.0 [IQR 22.0-95.0]; test set: median age 92.5 [IQR 16.8-131.8]) than children with unlikely TB (training set: median age 38.0 [IQR 13.8-82.3]; test set: median age 22.5 [IQR 13.0-53.8]) in both training and test sets. The percentage of females was slightly higher in the confirmed TB groups (training set: 54.5%; test set: 47.0%) compared to the unlikely TB group (training set: 50.0%; test set: 26.9%). The percentage of HIV-infected children in the confirmed TB group was 45.5% and 50.0% in training and test sets respectively. In the unlikely TB group, 48.0% and 38.5% of the children were HIV-infected in training and test sets respectively. Most of the confirmed TB cases were PTB in both training and test sets, and almost 50% of unlikely TB cases had a respiratory tract infection.

6.2.1 Case definitions

A diagnostic algorithm was used to classify the children into microbiologically confirmed TB, unconfirmed TB, unlikely TB and *Mtb* infection [Section 2.3]. In brief, microbiologically confirmed TB diagnosis was

Multi-omics dataset			
		Confirmed TB (n=69)	Unlikely TB (n=126)
Age (months), median (IQR)		53.0 (21.0-108.0)	31.0 (13.0-79.5)
	missing	0	0
Female, no. (%)		37.0 (53.6%)	54.0 (42.9%)
	missing	0	0
HIV-infected, no. (%)		32.0 (46.4%)	58.0 (46.0%)
	missing	0	0
PTB, no. (%)		64.0 (92.8%)	
	missing	0	
Respiratory tract infection, no. (%)			57.0 (45.2%)
	missing		0
Training set			
		Confirmed TB (n=55)	Unlikely TB (n=100)
Age (months), median (IQR)		53.0 (22.0-95.0)	38.0 (13.8-82.3)
	missing	0	0
Female, no. (%)		30.0 (54.5%)	47.0 (47.0%)
	missing	0	0
HIV-infected, no. (%)		25.0 (45.5%)	48.0 (48.0%)
	missing	0	0
PTB, no. (%)		50.0 (90.9%)	
	missing	0	
Respiratory tract infection, no. (%)			44.0 (44.0%)
	missing		0
Test set			
		Confirmed TB (n=14)	Unlikely TB (n=26)
Age (months), median (IQR)		92.5 (16.8-131.8)	22.5 (13.0-53.8)
	missing	0	0
Female, no. (%)		7.0 (50.0%)	7.0 (26.9%)
	missing	0	0
HIV-infected, no. (%)		7.0 (50.0%)	10.0 (38.5%)
	missing	0	0
PTB, no. (%)		14.0 (100.0%)	
	missing	0	
Respiratory tract infection, no. (%)			13.0 (50.0%)
	missing		0

Table 6.1: Clinical features of patients included in the transcriptomic and proteomic datasets, training set, and test set used to identify and test novel multi-omics biomarkers for distinguishing between TB disease and unlikely TB in children.

IQR: Interquartile range; HIV: Human Immunodeficiency Virus; PTB: Pulmonary Tuberculosis

based on the bacteriological confirmation either by culture or NAATs from at least 1 specimen obtained from the site of TB. A child was diagnosed with unconfirmed TB when the bacteriological confirmation was not obtained and the patient had at least 2 of the following: symptoms or signs suggestive of TB; close TB exposure or immunological evidence of *Mtb*; positive response to TB treatment; chest radiography or other evidence suggestive of TB. When the bacteriological confirmation was not obtained and the criteria for “unconfirmed TB” were not met, the symptomatic child was classified as unlikely TB. A healthy child with no clinical features of TB disease and a positive IGRA or TST result was classified as having *Mtb* infection.

Although the clinical algorithm describes the classification of patients into 4 diagnostic groups, only microbiologically confirmed TB and unlikely TB patients were included in the analysis described in this chapter.

6.2.2 Data generation

Transcriptomic dataset The transcriptomic dataset was generated using bulk RNA-Seq. Whole blood was collected into PAXgene Blood RNA Tubes (PreAnalytiX) from patients at the earliest time point after recruitment and prior to any treatment. PAXgene tubes were frozen within 6 hours after collection and stored at -80°C until extraction. Total RNA was isolated using the PAXgene Blood RNA Kits following the recommended methodology and sent for sequencing at The Wellcome Centre for Human Genetics in Oxford, United Kingdom.

The protocol followed to generate the RNA-Seq data was as follows. Genomic material was quantified using RiboGreen (Invitrogen) on the FLUOstar OPTIMA plate reader (BMG Labtech) and the size profile and integrity were analysed on the 2200 TapeStation (Agilent, RNA ScreenTape). Input material was normalised, and strand-specific library preparation was completed using NEBNext® Ultra™ II mRNA kit (NEB) and NEB rRNA/globin depletion probes following the manufacturer's instructions. Libraries were prepared on a Tetrad (Bio-Rad) using *in-house* unique dual indexing primers [269]. Individual libraries were normalised using Qubit and pooled together. The pooled library was diluted to ~10 nM for storage, denatured and further diluted prior to loading on the sequencer. Samples were randomised across RNA-Seq plates. Sequencing was performed at the Oxford Genomics Centre using a Novaseq6000 platform at 150 base-pair paired-end configuration, generating a raw read count of 30 million reads per sample. Following sequencing, the raw RNA-Seq reads were processed and the count matrix was generated as outlined in Section 2.5.2.

Proteomic datasets The proteomic dataset was generated by immunoassays. Sera were collected from patients at recruitment in serum-separating tubes and stored at -80°C. Luminex® and ELISA assays were used for quantification of the target proteins. 48 proteins were measured by Luminex® and 1 protein was measured by ELISA since it was not available via Luminex®. The 49 target proteins were selected from the literature or previously identified as biomarkers for TB disease or inflammation [Section 5.2.2]. Patients were randomised across the Luminex®/ELISA plates. Sera were diluted and target proteins were measured as per manufacturers' protocols.

6.3 Results

The aim of this study was to identify a multi-omics diagnostic signature for distinguishing between TB disease vs. unlikely TB in children. RNA-Seq and targeted proteomics, either ELISA or Luminex® assays, were used to generate the transcriptomic and proteomic data, which were subsequently combined to identify multi-omics biomarkers. The transcriptomic and proteomic datasets were first analysed in a single-omics manner and a gene and a protein signature for distinguishing between microbiologically confirmed TB vs. unlikely TB were identified. Finally, the two single-omics datasets were combined for the identification of a multi-omics signature. Single-omics and multi-omics analyses were carried out on the same cohort of samples for benchmarking the performance of the single-omics signature against the performance of the multi-omics one.

6.3.1 Transcriptomic data

Data pre-processing Samples not in common with the proteomic dataset were excluded. Pre-processing and normalisation were performed as described in Section 2.6. In brief, genes with low counts were

excluded with the `filterByExpr` function [243] to remove genes that did not have sufficiently large counts to be retained in the statistical analysis [Section 2.6.2]. No outlier samples were identified by PCA in this dataset.

The pre-processed counts were randomly split into training and test sets (80:20 split) and normalised separately using `edgeR` [243] [Section 2.6.2]. PCA performed on the training set did not capture any source of technical variation in the data [Figure 6.1]. The disease group was captured by PC1 [Figure 6.1; Panel A] with the microbiologically confirmed TB patients clustering towards the higher values of PC1 and the unlikely TB samples towards lower values. The HIV status was captured by PC2 [Figure 6.1; Panel C] with the HIV-infected samples clustering at the top of the plot. The recruitment site was captured by PC4 [Figure 6.1; Panel F] with the Kenyan samples clustering towards the higher values of PC4 and the ILULU Cape Town and ILULU Malawi cohorts clustering towards the bottom of the plot. No differences were observed in terms of sex or age [Figure 6.1; Panels G, H, I, J].

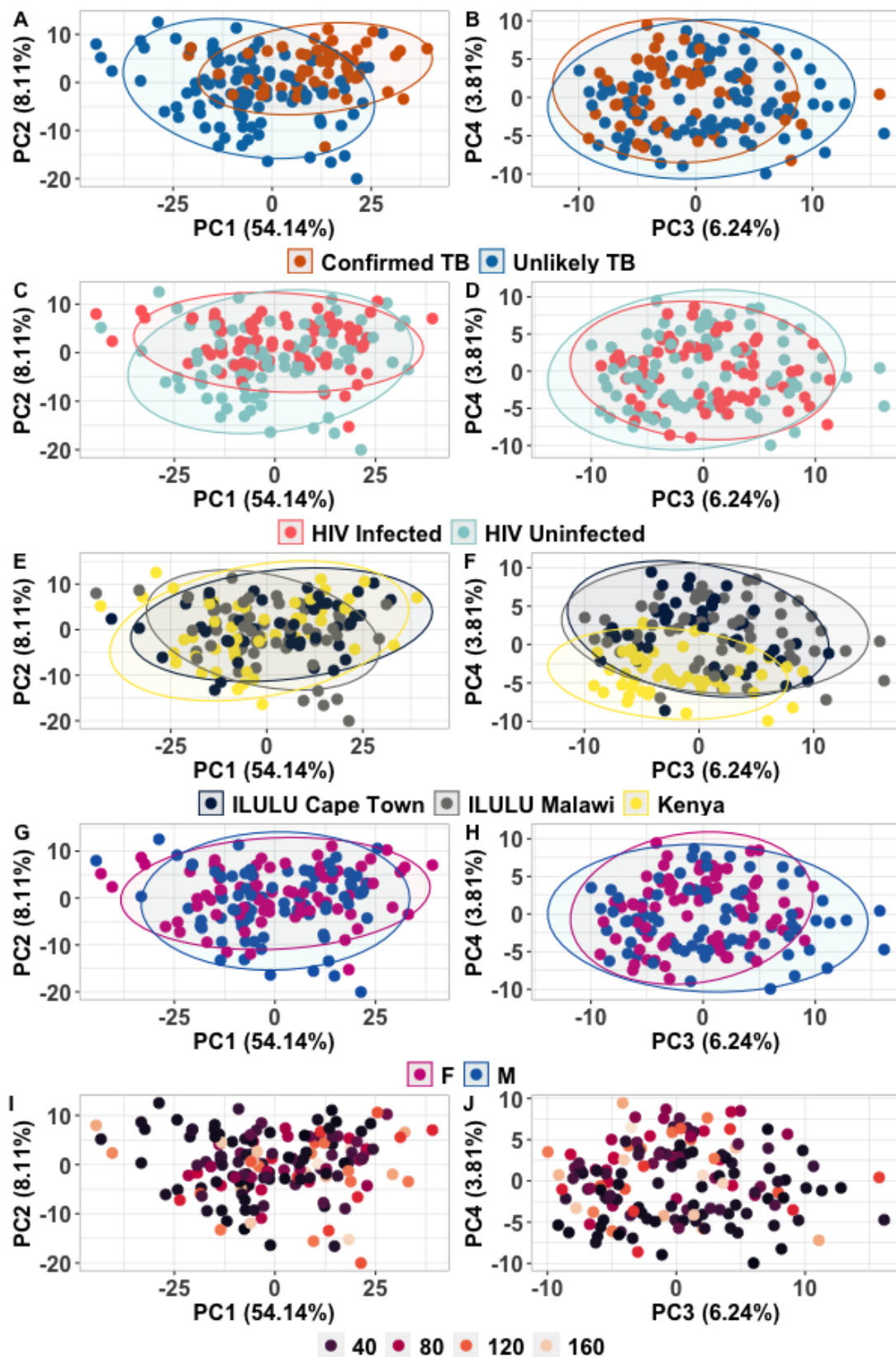


Figure 6.1: PCA biplots from PCA performed on the RNA-Seq training dataset used in the multi-omics integration analysis coloured according to clinical variables.

PC 1-2 and PC 3-4 are shown. Points represent samples and are coloured according to disease group [Panel A and Panel B], HIV status [Panel C and Panel D], recruitment site [Panel E and Panel F], sex [Panel G and Panel H] and age [Panel I and Panel J]. PCA = Principal Component Analysis; RNA-Seq = RNA Sequencing; PC: Principal Component; HIV: Human Immunodeficiency Virus; F = female; M = male

Differential Expression analysis Subsequently, DE analysis was performed on the training set using edgeR [243]. A linear model accounting for age and sex was fitted and 498 genes (32 under- and 466 over-expressed) [Appendix F] were identified as SDE ($\log_2FC > 0.5$ or $\log_2FC < -0.5$ and BH-adjusted p -value < 0.05) between microbiologically confirmed TB and unlikely TB disease groups [Figure 6.2].

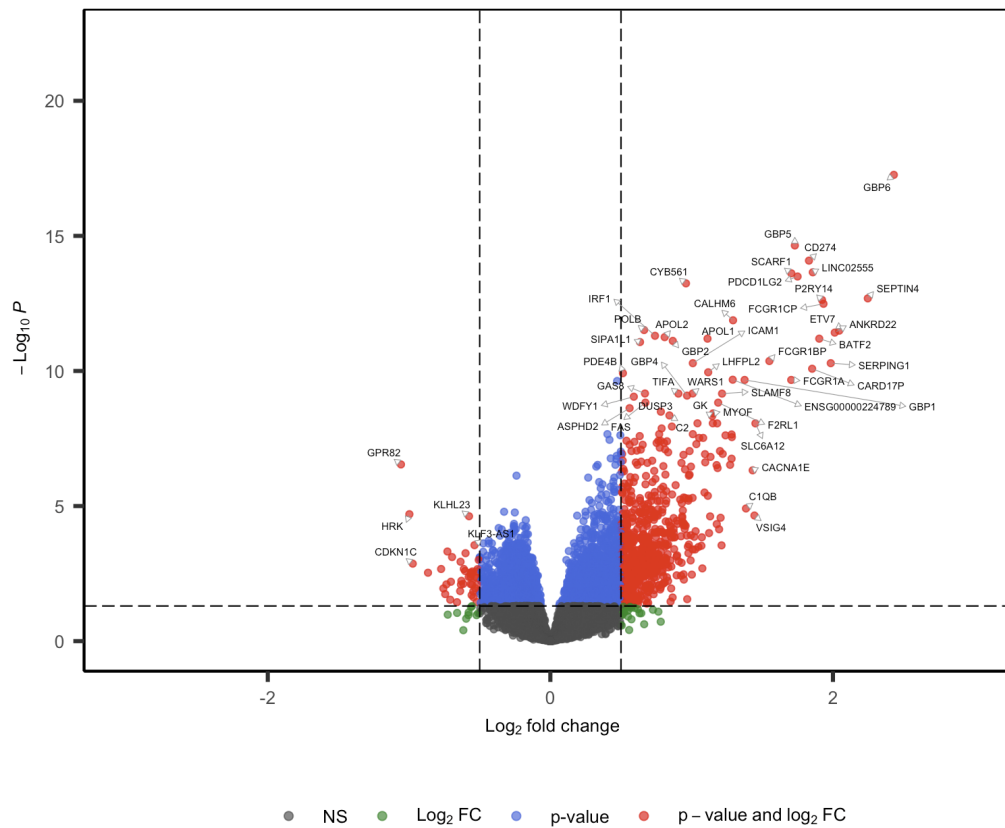


Figure 6.2: Volcano plot showing $\log_2FC$ and BH-adjusted p -values from Differential Expression (DE) analysis comparing microbiologically confirmed TB vs. unlikely TB in the training set of the RNA-Seq data. Points represent genes and are coloured as follows: red - genes with BH-adjusted p -values < 0.05 and absolute $\log_2FC > 0.5$; blue - genes with BH-adjusted p -values < 0.05 and absolute $\log_2FC < 0.5$; green - genes with BH-adjusted p -values > 0.05 and absolute $\log_2FC > 0.5$; grey - genes with BH-adjusted p -values > 0.05 and absolute $\log_2FC < 0.5$. Genes to the right of zero are over-expressed in TB. FC = Fold Change; BH = Benjamini-Hochberg; RNA -Seq = RNA Sequencing; DE = Differential Expression

Feature selection FS-PLS with cross-validation was performed on the training set and on the SDE genes to identify a sparse diagnostic signature for distinguishing between the two disease groups of interest. The feature selection method identified a 6-gene signature composed of *GBP6* (Guanylate Binding Protein 6), *GM2A* (Ganglioside GM2 Activator), *CYB561* (Cytochrome B561), *TPPP3* (Tubulin Polymerization Promoting Protein Family Member 3), *IFIT2* (Interferon Induced Protein With Tetratricopeptide Repeats 2) and *CALHM6* (Calcium Homeostasis Modulator Family Member 6).

$\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity in training and test sets of the 6 candidate biomarker genes are listed in Table 6.2, and their boxplots are shown in Figure 6.3.

Training set						
Gene Name	Ensembl ID	log ₂ FC	BH-adjusted p-value	AUC	Sensitivity	Specificity
GBP6	ENSG00000183347	2.40	5.5×10^{-18}	90.0%	87.3%	81.0%
GM2A	ENSG00000196743	0.60	8.9×10^{-07}	76.6%	78.2%	70.0%
CYB561	ENSG00000008283	0.96	5.7×10^{-14}	86.1%	92.7%	66.0%
TPPP3	ENSG00000159713	-0.77	2.1×10^{-03}	66.2%	45.5%	82.0%
IFIT2	ENSG00000119922	0.86	1.7×10^{-03}	70.3%	56.4%	78.0%
CALHM6	ENSG00000188820	1.33	1.3×10^{-12}	84.1%	83.6%	75.0%
Test set						
Gene Name	Long Gene Name	log ₂ FC	BH-adjusted p-value	AUC	Sensitivity	Specificity
GBP6	ENSG00000183347	2.32	7.1×10^{-03}	92.9%	92.6%	80.8%
GM2A	ENSG00000196743	0.34	0.70	66.5%	64.3%	73.1%
CYB561	ENSG00000008283	1.19	3.6×10^{-02}	90.9%	78.6%	100.0%
TPPP3	ENSG00000159713	-0.01	0.99	50.3%	64.3%	50.0%
IFIT2	ENSG00000119922	1.19	0.66	70.9%	92.9%	50.0%
CALHM6	ENSG00000188820	1.59	8.4×10^{-03}	90.0%	92.9%	80.8%

Table 6.2: log₂FC, BH-adjusted *p*-value, AUC, sensitivity and specificity of the 6 candidate biomarkers in the training and test sets of the RNA-Seq dataset when contrasting microbiologically confirmed TB vs. unlikely TB.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve.

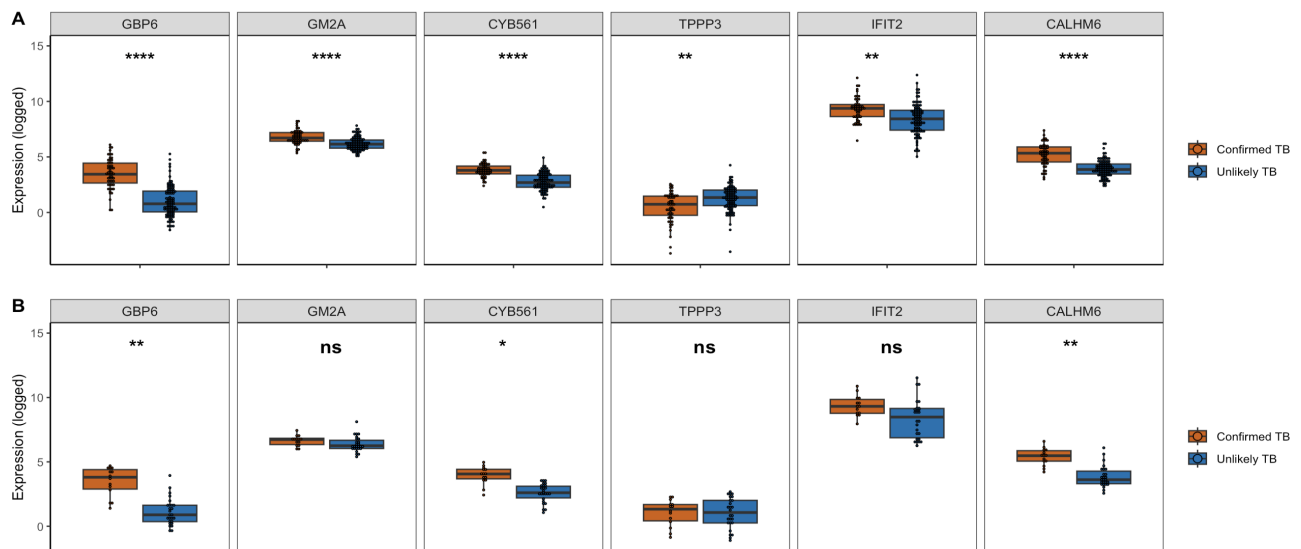


Figure 6.3: Logged expression values of the 6 candidate biomarker genes in the training set [Panel A] and test set [Panel B] of the RNA-Seq dataset coloured according to disease group. ns: *p*-value > 0.05 ; *: *p*-value < 0.05 ; **: *p*-value ≤ 0.01 ; ***: *p*-value ≤ 0.001 ; ****: *p*-value ≤ 0.0001 . All *p*-values are adjusted for multiple testing using Benjamini-Hochberg.

Signature performance evaluation The output model of FS-PLS was retrained using logistic regression since a logistic regression model is more suitable for a binary outcome. The performance of the final signature was evaluated in terms of AUC, sensitivity and specificity using a wDRS for each patient [Section 2.7]. The wDRS was calculated by multiplying the logged (base 2) gene expression values of each of the 6 candidate gene biomarkers by their weights/coefficients obtained from the logistic regression model for each patient. The 6-gene signature achieved an AUC of 95.5% [CI_{95%} 91.7%-98.2%], sensitivity of 80.0% [CI_{95%} 69.1%-89.1%] and specificity of 97.0% [CI_{95%} 93.0%-100.0%] in the training set, and an AUC of 95.6% [CI_{95%} 90.2%-100.0%], sensitivity of 100.0% and specificity of 84.6% [CI_{95%} 69.2%-96.1%] in the test set [Figure 6.4]. When the specificity was

fixed at 80.0%, the sensitivity achieved 100.0% [CI_{95%} 64.3%-100.0%] in the test set, meeting the optimal WHO Target Product Profiles (TPPs) criteria for a triage test but not for a diagnostic test for TB.

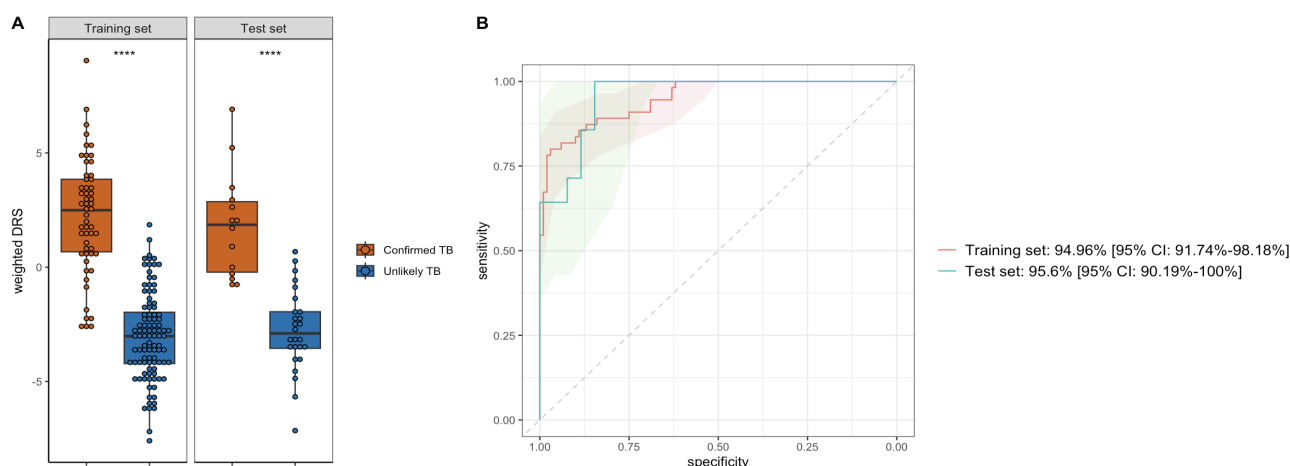


Figure 6.4: Boxplot of the wDRS of the 6-gene signature and its ROC curve in the training and test sets of the RNA-Seq dataset.

ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . wDRS = weighted Disease Risk Score; ROC = Receiver Operating Characteristic; RNA-Seq = RNA Sequencing; AUC = Area Under the ROC Curve

6.3.2 Proteomic data

The targeted proteomic datasets included in this analysis are the following:

1. The validation by immunoassay of a 4-protein diagnostic signature for childhood TB identified in a SELDI dataset
2. The validation by immunoassay of a 4-protein diagnostic signature for childhood TB identified in a SomaScan® dataset
3. A pool of 41 proteins measured by Luminex® assay

Data pre-processing The samples not in common with the transcriptomic datasets were excluded. Pre-processing was performed as described in Section 2.6.3. In brief, the standard curves for each analyte in each plate were visually examined. The values extrapolated by the Luminex® or ELISA software were kept in the analysis. The under-the-limit-of-detection concentration values (i.e., OOR $<$) were assigned either the lowest extrapolated value minus its 10% or half of the lowest limit of detection when no values under the limit of detection were extrapolated. The concentrations over the detection limit were set to *NULL*. The pre-processed concentration values were split into training and test sets, following the sample split used in the transcriptomic data. PCA performed on the training set did not capture any source of technical variability in the data [Figure 6.5]. The disease group was captured by PC1 [Figure 6.5; Panel A] with the microbiologically confirmed TB patients clustering towards the higher values of PC1 and the unlikely TB samples towards the left side of the PC1 vs. PC2 biplot. The HIV status was captured by PC3 [Figure 6.5; Panel D] with the HIV-uninfected samples clustering at the bottom left of the plot. The recruitment site was captured by PC3 and PC4 [Figure 6.5; Panel F] with the Kenyan samples clustering towards the top right of PC3 vs. PC4 and the ILULU Cape Town and ILULU Malawi cohorts clustering towards the bottom left of the plot. No differences were observed in terms of sex or age [Figure 6.5; Panels G, H, I, J].

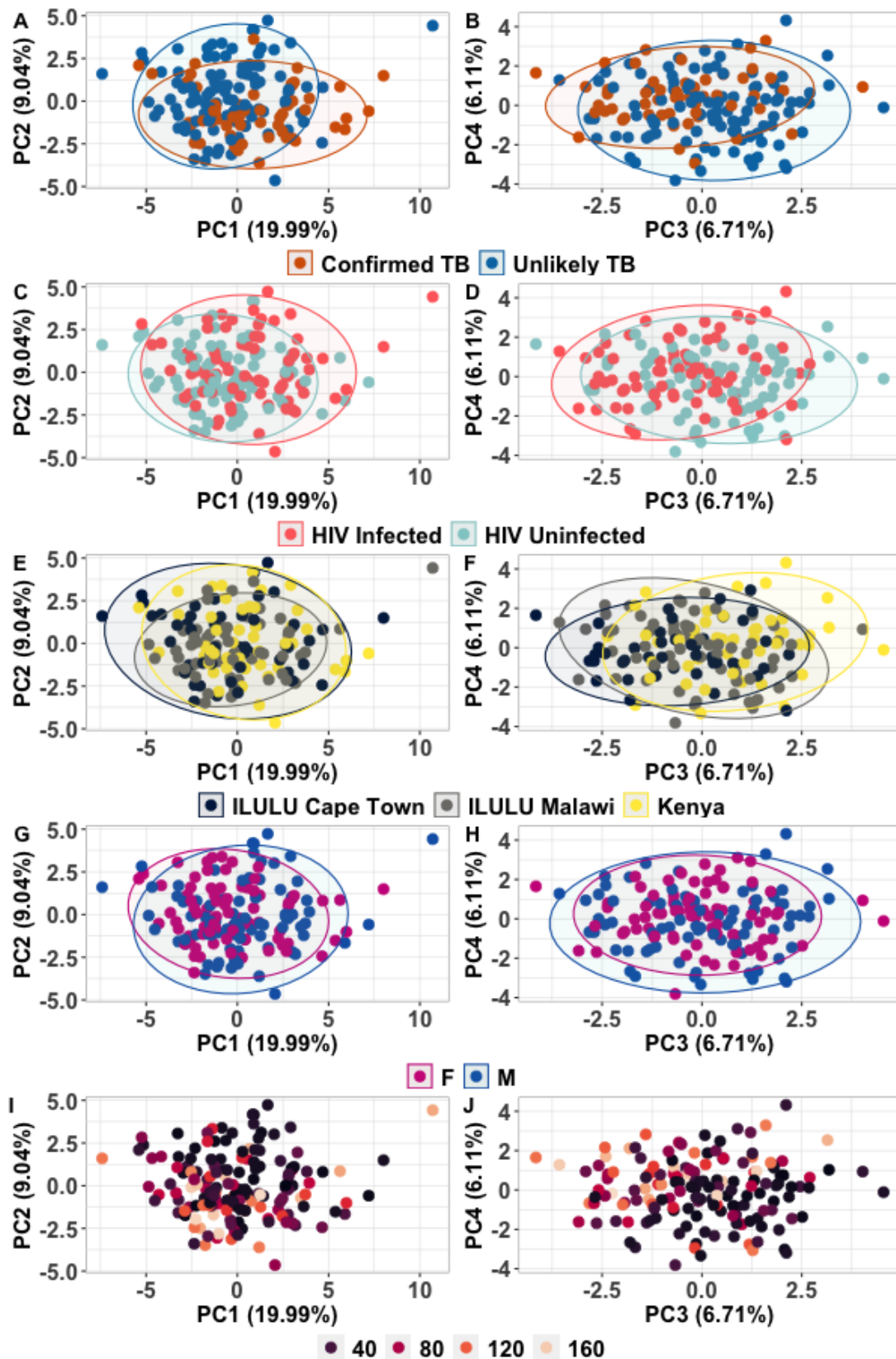


Figure 6.5: Principal Component Analysis (PCA) biplots from PCA performed on the training proteomic dataset used in the multi-omics integration analysis coloured according to clinical variables.

PC 1-2 and PC 3-4 are shown. Points represent samples and are coloured according to disease group [Panel A and Panel B], HIV status [Panel C and Panel D], recruitment site [Panel E and Panel F], sex [Panel G and Panel H] and age [Panel I and Panel J]. *PC*: Principal Component; *HIV*: Human Immunodeficiency Virus; *F* = female; *M* = male

Differential Abundance analysis A linear model accounting for age and sex was used to calculate $\log_2FC$ and BH-adjusted p -value of the 49 proteins in the training set. Proteins with BH-adjusted p -value <0.05 and $\log_2FC < -0.5$ or $\log_2FC > 0.5$ were considered SDA. A total of 7 proteins (CCL1, CXCL9, FcR-10, CRP, CXCL10, IFN- γ , IL-7), all more abundant in microbiologically confirmed TB, were SDA when contrasting microbiologically confirmed TB versus unlikely TB.

Feature selection FS-PLS with cross-validation was performed on the training set and on the list of SDA proteins to identify the best performing proteomic diagnostic signature in discriminating between the two disease groups. The feature selection algorithm identified a 5-protein signature composed of CCL1 (C-C motif chemokine 1), CRP (C-reactive protein), IL-7 (Interleukin-7), CXCL10 (C-X-C motif chemokine 10) and FcR-10 (Low affinity immunoglobulin gamma Fc region receptor III-B).

$\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity in training and test sets of the 5 candidate biomarker proteins are listed in Table 6.3, and their boxplots are shown in Figure 6.6.

Training set						
Protein Name	UniProt ID	$\log_2FC$	BH-adjusted p -value	AUC	Sensitivity	Specificity
CCL1	P22362	2.92	2.3×10^{-13}	82.9%	81.8%	75.0%
CRP	P02741	1.59	3.1×10^{-03}	65.2%	89.1%	49.0%
IL-7	P13232	0.60	4.1×10^{-02}	61.1%	43.6%	77.0%
CXCL10	P02778	1.03	7.6×10^{-03}	67.1%	49.1%	84.0%
FcR-10	O75015	0.71	1.1×10^{-03}	69.2%	56.4%	81.0%
Test set						
Protein Name	UniProt ID	$\log_2FC$	BH-adjusted p -value	AUC	Sensitivity	Specificity
CCL1	P22362	2.94	3.9×10^{-04}	88.3%	92.9%	80.8%
CRP	P02741	1.03	0.82	63.7%	64.3%	69.2%
IL-7	P13232	0.81	0.19	69.0%	78.6%	61.5%
CXCL10	P02778	0.87	0.82	70.3%	78.6%	69.2%
FcR-10	O75015	1.41	1.0×10^{-03}	89.8%	85.7%	92.3%

Table 6.3: $\log_2FC$, BH-adjusted p -value, AUC, sensitivity and specificity of the 5 candidate biomarkers in the training and test sets of the proteomic dataset when contrasting microbiologically confirmed TB vs. unlikely TB.

FC = Fold Change; BH = Benjamini-Hochberg; AUC = Area Under the ROC Curve.

Signature performance evaluation The output model of FS-PLS was retrained using logistic regression since a logistic regression model is more suitable for a binary outcome. The performance of the final signature was evaluated in terms of AUC, sensitivity and specificity using a wDRS for each patient [Section 2.7]. The wDRS was calculated by multiplying the logged (base 2) abundance of each of the 5 candidate protein biomarkers by their weights/coefficients for each patient. The 5-protein signature achieved an AUC of 84.6% [$CI_{95\%}$ 77.7%-91.5%], sensitivity of 69.1.0% [$CI_{95\%}$ 56.4%-80.0%] and specificity of 92.0% [$CI_{95\%}$ 86.0%-97.0%] in the training set, and an AUC of 92.9% [$CI_{95\%}$ 83.9%-100.0%], sensitivity of 78.6% [$CI_{95\%}$ 57.1%-100.0%] and specificity of 96.2% [$CI_{95\%}$ 88.5%-100.0%] in the test set [Figure 6.7]. When the specificity was fixed at 98.0%, sensitivity achieved 78.6% [$CI_{95\%}$ 50.0%-92.8%] in the test set, meeting the WHO TPPs criteria for a diagnostic test. However, the signature did not meet the optimal WHO TPPs criteria for a triage test (sensitivity achieved 85.7% [$CI_{95\%}$ 64.3%-100.0%] when the specificity was fixed at 80.0%).

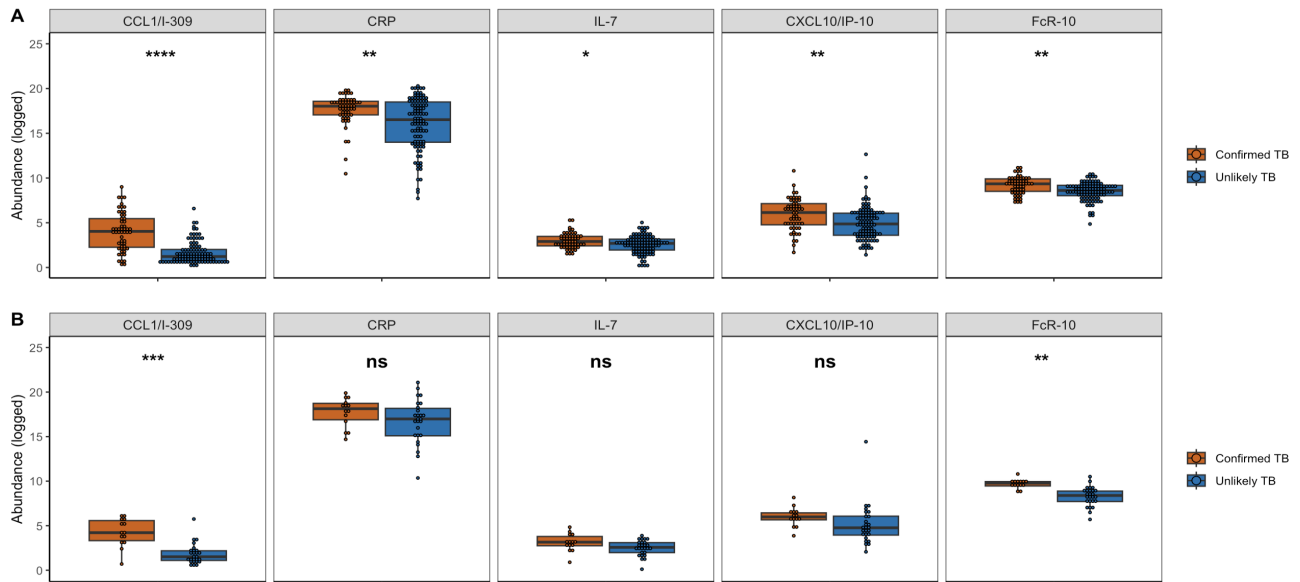


Figure 6.6: Logged expression values of the 5 candidate biomarkers in the training set [Panel A] and test set [Panel B] of the proteomic dataset coloured according to disease group.

ns: p -value > 0.05; *: p -value < 0.05; **: p -value ≤ 0.01; ***: p -value ≤ 0.001; ****: p -value ≤ 0.0001. All p -values are adjusted for multiple testing using Benjamini-Hochberg.

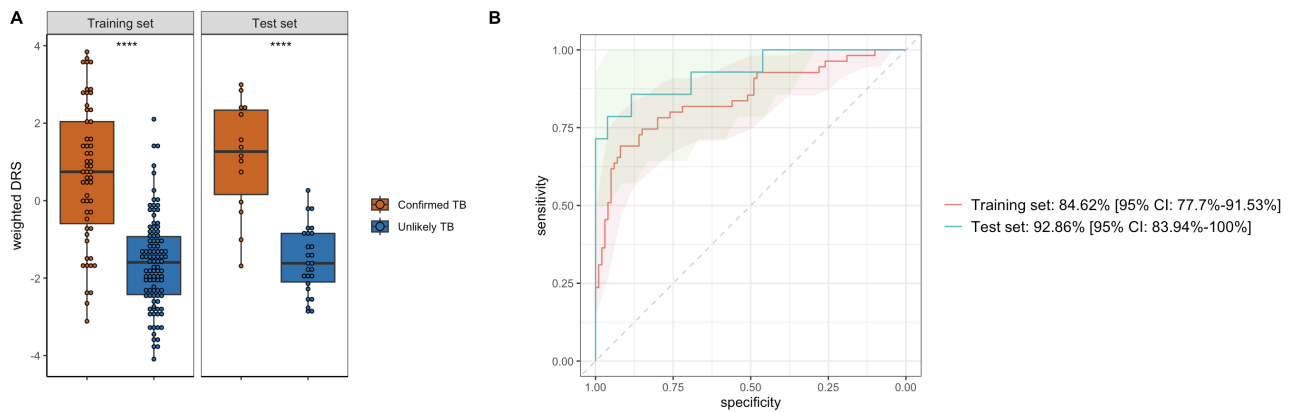


Figure 6.7: Boxplot of the wDRS of the 5-protein signature and its ROC curve in the training and test sets of the proteomic dataset.

ns: p -value > 0.05; *: p -value < 0.05; **: p -value ≤ 0.01; ***: p -value ≤ 0.001; ****: p -value ≤ 0.0001. wDRS = weighted Disease Risk Score; ROC = Receiver Operating Characteristic; AUC = Area Under the ROC Curve

6.3.3 Multi-omics integration

Data pre-processing The pre-processed transcriptomic and proteomic training sets were combined. PCA was performed to assess whether any source of technical variation was captured in the data [Figure 6.8]. The disease group was captured by PC1 [Figure 6.8; Panel A] with the microbiologically confirmed TB patients clustering towards the higher values of PC1 and the unlikely TB samples towards lower values. The HIV status was captured by PC2 [Figure 6.8; Panel C] with the HIV-infected samples clustering at the top of the PC1 vs. PC2 biplot. The recruitment site was captured by PC4 [Figure 6.8; Panel F] with the Kenyan samples clustering towards the lower values of PC4 and the ILULU Cape Town and ILULU Malawi cohorts clustering towards the top of the PC3 vs. PC4 biplot. No differences were observed in terms of sex or age [Figure 6.8; Panels G, H, I, J].

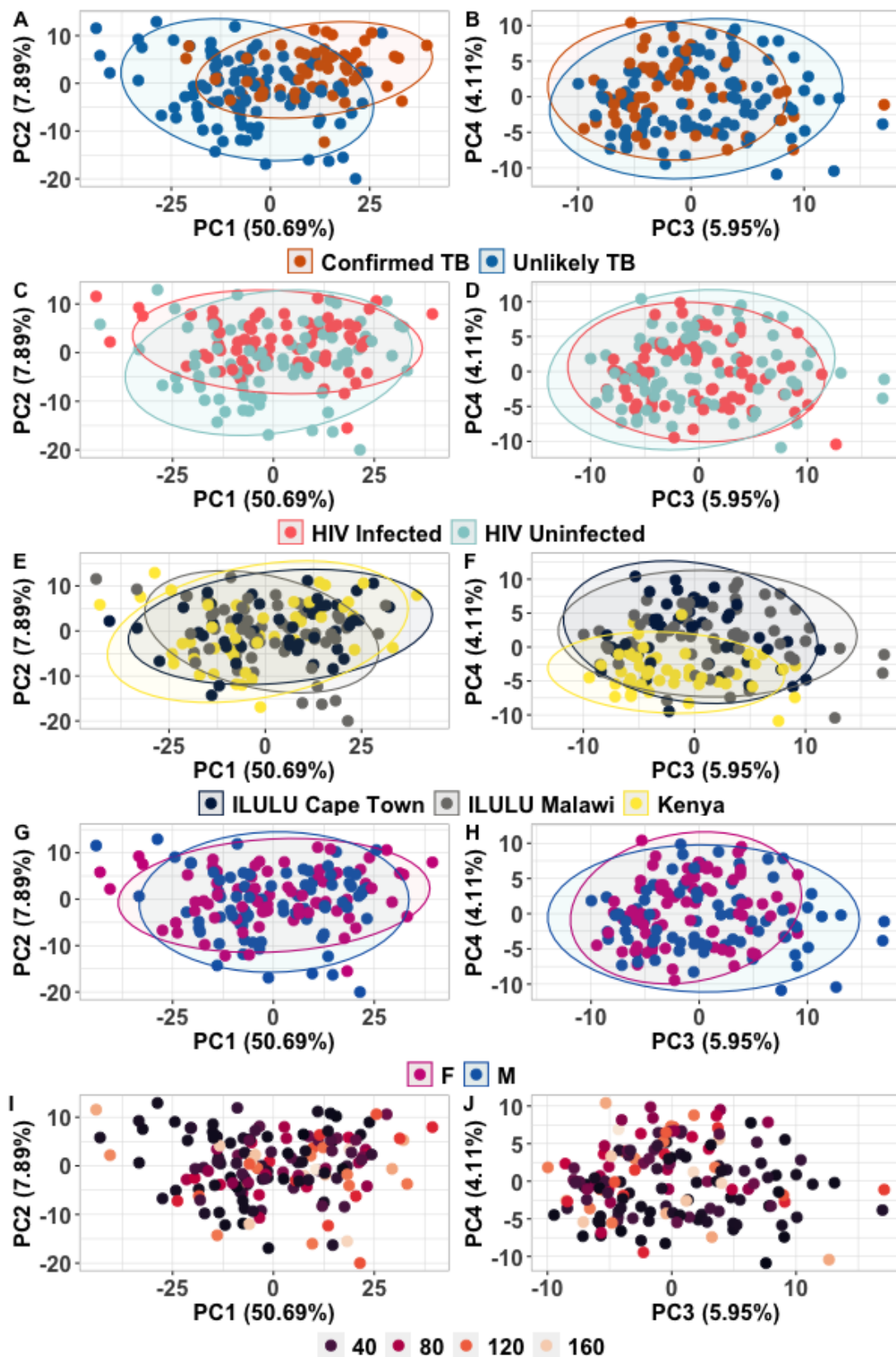


Figure 6.8: Principal Component Analysis (PCA) biplots from PCA performed on the integrated RNA-Seq and proteomic training datasets coloured according to clinical variables.

PC 1-2 and PC 3-4 are shown. Points represent samples and are coloured according to disease group [Panel A and Panel B], HIV status [Panel C and Panel D], recruitment site [Panel E and Panel F], sex [Panel G and Panel H] and age [Panel I and Panel J]. *PC*: Principal Component; *HIV*: Human Immunodeficiency Virus; *F* = female; *M* = male

Feature selection DIABLO [265] was used for multi-omics biomarker identification. DIABLO can be used on the full set of features (i.e., with no restrictions in terms of $\log_2FC$ and BH-adjusted p -value). However, for benchmarking reasons and ensuring a fair comparison between the performance of single- and multi-omics signatures, it was used on the SDE genes and SDA proteins identified in the single-omics analyses. The most correlated among the single-omics datasets and most discriminating between the disease groups of interest biomarkers were identified. Correlation between the single-omics dataset is one of the input parameters needed by DIABLO to identify the list of biomarkers. The estimated global correlation between the two single-omics was calculated as correlation between the variates of a partial least square regression model and used in the design matrix of the `block.splsda` function [265].

7 genes and 7 proteins were identified as the most discriminant features between the two disease groups of interest [Figure 6.9].

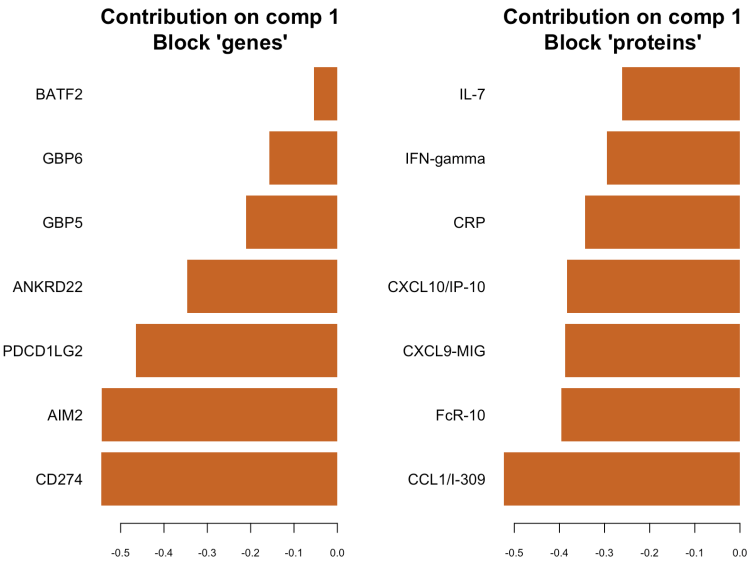


Figure 6.9: Most discriminant genes and proteins in the multi-omics training set. The single bars describe the weight of each corresponding feature in the first latent component produced by DIABLO. The higher the value, the more discriminant the feature between the two disease groups of interest in the first latent component.

FS-PLS was successively used to reduce the number of biomarkers in the signature. DIABLO, in fact, is a tool developed to identify biomarkers accounting for characteristics of the multi-omics dataset, such as which single-omics dataset each feature is part of or the correlation between the single-omics datasets. Most of the feature selection algorithms, such as FS-PLS, identify the best-performing subset of features without considering such parameters, which are critical in multi-omics analyses. Therefore, FS-PLS was used on the 14 features (7 genes and 7 proteins) identified by DIABLO on the multi-omics training set to select a smaller diagnostic signature with high accuracy in distinguishing microbiologically confirmed TB from unlikely TB. FS-PLS was used with cross-validation [Section 2.7] in the training set and a 5-feature signature, composed of *GBP6* (Guanylate Binding Protein 6), *CCL1* (C-C motif chemokine 1), *CXCL10* (C-X-C motif chemokine 10), *GBP5* (Guanylate Binding Protein 5) and *IL-7* (Interleukin-7), was identified.

Signature performance evaluation The output model of FS-PLS was retrained using logistic regression, which is more suitable for a binary outcome. The performance of the final signature was evaluated in terms of AUC, sensitivity and specificity using a wDRS for each patient [Section 2.7]. The

wDRS was calculated by multiplying the logged (base 2) expression/abundance of each of the 5 features in the final signature by their weights/coefficients for each patient. The 5-protein signature achieved an AUC of 93.1% [CI_{95%} 88.5%-97.6%], sensitivity of 89.1% [CI_{95%} 80.0%-96.4%] and specificity of 88.0% [CI_{95%} 82.0%-94.0%] in the training set, and an AUC of 97.5% [CI_{95%} 93.6%-100.0%], sensitivity of 92.9% [CI_{95%} 78.6%-100.0%] and specificity of 92.3% [CI_{95%} 80.8%-100.0%] in the test set [Figure 6.10]. When the specificity was set to 98.0% in the test set, sensitivity achieved 85.7% [CI_{95%} 57.1%-100.0%], meeting the optimal values for sensitivity and specificity for a POCT for TB disease [44]. The signature met also the optimal WHO TPPs criteria for a triage test with sensitivity achieving 100.0% [CI_{95%} 78.6%-100.0%] when the specificity was set at 80.0%.

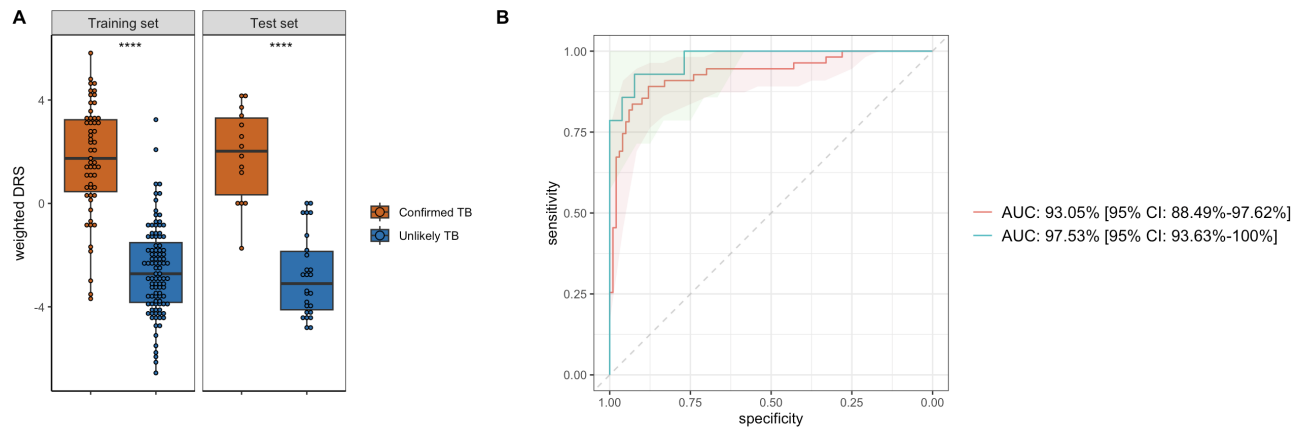


Figure 6.10: Boxplot of the wDRS of the 5-feature signature and its ROC curve in the training and test sets of the multi-omics dataset.

ns: p -value >0.05 ; *: p -value <0.05 ; **: p -value ≤ 0.01 ; ***: p -value ≤ 0.001 ; ****: p -value ≤ 0.0001 . wDRS = weighted Disease Risk Score; ROC = Receiver Operating Characteristic; AUC = Area Under the ROC Curve

Table 6.4 describes the performance of the transcriptomic, proteomic and multi-omics signatures based on HIV status, site of TB or cohort in the test set. Whilst the transcriptomic signature had a similar performance in HIV-infected and HIV-uninfected patients, the proteomic and multi-omics signature had a better performance in HIV-uninfected samples. In terms of site of TB, the transcriptomic and multi-omics signatures had a better performance in patients without any form of EPTB, whilst the proteomic signature had a better performance in patients diagnosed with both PTB and EPTB. Finally, the signatures had a better performance in the ILULU Cape Town cohort, compared to the ILULU Malawi and Kenyan cohorts. The transcriptomic signature achieved 100% in terms of sensitivity in each subgroup, thus it was able to detect all true TB cases.

6.4 Biological function of the 5 features

The 5-feature signature identified in the multi-omics training set and tested in the multi-omics test set is composed of *GBP6* (Guanylate Binding Protein 6), *CCL1* (C-C motif chemokine 1), *CXCL10* (C-X-C motif chemokine 10), *GBP5* (Guanylate Binding Protein 5) and *IL-7* (Interleukin-7). The 2 genes and 3 proteins have biological functions that might provide insights into the pathogenesis of childhood TB.

GBP6 and *GBP5*, both over-expressed in TB disease compared to unlikely TB, are protein-coding genes belonging to the GTPase superfamily. GBP genes provide defence against pathogens and are induced by the interferon (IFN) cytokine family [311]. GBP genes, such as *GBP2*, *GBP5* and *GBP6*, appear in several TB signatures [312]. *GBP5* was one of the candidate biomarker genes in the transcriptomic signature identified and validated in Chapter 3 of this thesis.

Transcriptomic signature - wDRS			
	AUC [CI _{95%}]	Sensitivity [CI _{95%}]	Specificity [CI _{95%}]
HIV-uninfected	95.5% [CI _{95%} 88.0%-100.0%]	100.0%	87.5% [CI _{95%} 68.8%-100.0%]
HIV-infected	95.7% [CI _{95%} 87.4%-100.0%]	100.0%	80.0% [CI _{95%} 50.0%-100.0%]
PTB	95.8% [CI _{95%} 90.3%-100.0%]	100.0%	84.6% [CI _{95%} 69.2%-96.2%]
PTB+EPTB	94.9% [CI _{95%} 83.8%-100.0%]	100.0%	84.6% [CI _{95%} 69.2%-96.2%]
ILULU Cape Town	100.0%	100.0%	100.0%
ILULU Malawi	95.2% [CI _{95%} 82.0%-100.0%]	100.0%	85.7% [CI _{95%} 57.1%-100.0%]
Kenya	94.0 % [CI _{95%} 81.3%-100.0%]	100.0%	90.0% [CI _{95%} 70.0%-100.0%]
Proteomic signature - wDRS			
	AUC [CI _{95%}]	Sensitivity [CI _{95%}]	Specificity [CI _{95%}]
HIV-uninfected	98.21% [CI _{95%} 94.0%-100.0%]	100.0%	87.5% [CI _{95%} 68.8%-100.0%]
HIV-infected	85.7% [CI _{95%} 64.9%-100.0%]	85.7% [CI _{95%} 57.1%-100.0%]	80.0% [CI _{95%} 50.0%-100.0%]
PTB	90.9% [CI _{95%} 79.8%-100.0%]	81.8% [CI _{95%} 54.6%-100.0%]	88.5% [CI _{95%} 76.9%-100.0%]
PTB+EPTB	100.0%	100.0%	100.0%
ILULU Cape Town	98.2% [CI _{95%} 93.0%-100.0%]	100.0%	88.9% [CI _{95%} 66.7%-100.0%]
ILULU Malawi	95.2% [CI _{95%} 82.0%-100.0%]	100.0%	85.7% [CI _{95%} 57.1%-100.0%]
Kenya	84.0 % [CI _{95%} 60.8%-100.0%]	60.0% [CI _{95%} 20.0%-100.0%]	100.0%
Multi-omics signature - wDRS			
	AUC [CI _{95%}]	Sensitivity [CI _{95%}]	Specificity [CI _{95%}]
HIV-uninfected	98.2% [CI _{95%} 94.1%-100.0%]	100.0%	93.8% [CI _{95%} 81.3%-100.0%]
HIV-infected	94.3% [CI _{95%} 82.2%-100.0%]	85.7% [CI _{95%} 57.1%-100.0%]	100.0%
PTB	97.5% [CI _{95%} 93.1%-100.0%]	90.0% [CI _{95%} 72.7%-100.0%]	96.2% [CI _{95%} 88.5%-100.0%]
PTB+EPTB	97.4% [CI _{95%} 91.3%-100.0%]	100.0%	92.3% [CI _{95%} 80.7%-100.0%]
ILULU Cape Town	100.0%	100.0%	100.0%
ILULU Malawi	95.2% [CI _{95%} 82.0%-100.0%]	100.0%	85.7% [CI _{95%} 57.1%-100.0%]
Kenya	96.0% [CI _{95%} 86.6%-100.0%]	100.0%	80.0% [CI _{95%} 50.0%-100.0%]

Table 6.4: Transcriptomic, proteomic and multi-omics signatures performance in the test set broken down by clinical subgroups.

CI: Confidence Interval; HIV: Human Immunodeficiency Virus; PTB: Pulmonary Tuberculosis; EPTB: Extra-Pulmonary Tuberculosis

C-C motif chemokine 1 (CCL1), which is more abundant in TB disease compared to unlikely TB, is a glycoprotein secreted by activated monocytes/macrophages, T lymphocytes and endothelial cells [301]. It is involved in the immune system processes, inflammatory response and signalling. CCL1 binds to C-C chemokine receptor type 8 (CCR8), which is expressed in monocytes/macrophages, T helper 2 cells (Th2) and regulatory T lymphocytes [302]. CCL1 has been identified as a potential protein biomarker for TB disease in previous paediatric or adult studies [188, 303–305]. This protein was part of the 4-protein diagnostic signature identified and validated in Chapter 5 of this thesis.

C-X-C motif chemokine 10 (CXCL10) is a small cytokine belonging to the CXC chemokine family. It was more abundant in children with TB disease compared to children with unlikely TB in our dataset. CXCL10, also known as Interferon gamma-induced protein 10 (IP-10), is produced by activated macrophages and antigen-presenting cells. It might be involved in cell growth and in stimulating angiogenesis in the presence of TB lesions [313]. CXCL10 might also play a role in limiting *Mtb* replication [305]. CXCL10 was identified as a potential biomarker for TB disease in previous studies [305, 313, 314].

Interleukin-7 (IL-7) is a cytokine belonging to the hemopoietin family. It was more abundant in children with TB disease compared to children with unlikely TB in our dataset. IL-7 is involved in host defence by regulating the development and homeostasis of immune cells, such as T lymphocytes, B lymphocytes, and NK cells [315]. It has been shown in a previous study that levels of IL-7 were significantly increased in adults with TB disease compared to healthy controls. Hence, it was previously identified as a

candidate biomarker for TB disease [316].

6.5 Summary

In this study, I used two different -omics modalities to identify a sparse diagnostic multi-omics signature to distinguish TB disease from unlikely TB in children. None of the single-omics diagnostic signatures for childhood TB disease published so far meets the WHO TPPs criteria for a POCT [44]. Single-omics datasets do not overlap extensively and each of them captures an incomplete picture of the underlying biology. Single-omics signatures could, therefore, miss important biological information captured by other biological molecules. A holistic approach based on the integration of multi-omics and complementary datasets might bridge this information gap.

Transcriptomic and proteomic data generated from biological samples of 195 children with microbiologically confirmed TB ($n=69$) and unlikely TB ($n=126$) were used to identify a gene signature, a proteomic signature and a multi-omics signature. The three signatures were identified in the same set of samples to be able to accurately compare their diagnostic performance. The newly identify 6-gene signature, composed of *GBP6* (Guanylate Binding Protein 6), *GM2A* (Ganglioside GM2 Activator), *CYB561* (Cytochrome B561), *TPPP3* (Tubulin Polymerization Promoting Protein Family Member 3), *IFIT2* (Interferon Induced Protein With Tetratricopeptide Repeats 2) and *CALHM6* (Calcium Homeostasis Modulator Family Member 6), achieved an AUC of 95.6% [$CI_{95\%}$ 90.2%-100.0%], sensitivity of 100.0% and specificity of 84.6% [$CI_{95\%}$ 69.2%-96.1%] in the test set. The 5-protein signature, composed of CCL1 (C-C motif chemokine 1), CRP (C-reactive protein), IL-7 (Interleukin-7), CXCL10 (C-X-C motif chemokine 10) and FcR-10 (Low affinity immunoglobulin gamma Fc region receptor III-B), achieved an AUC of 92.9% [$CI_{95\%}$ 83.9%-100.0%], sensitivity of 78.6% [$CI_{95\%}$ 57.1%-100.0%] and specificity of 96.2% [$CI_{95\%}$ 88.5%-100.0%] in the test set. The 6-gene signature meets the optimal WHO TPPs criteria for a triage test (sensitivity achieved 100.0% [$CI_{95\%}$ 64.3%-100.0%] when the specificity was fixed at 80.0%), but not for a diagnostic test (sensitivity achieved 64.3% [$CI_{95\%}$ 35.7%-92.9%] when the specificity was fixed at 98.0%) in the test set. Since the 5-protein signature is more specific than the transcriptomic signature, it met the WHO TPPs criteria for a diagnostic test (sensitivity achieved 78.6% [$CI_{95\%}$ 50.0%-92.8%] when the specificity was fixed at 98.0%), but not the optimal WHO TPPs criteria for a triage test (sensitivity achieved 85.7% [$CI_{95\%}$ 64.3%-100.0%] when the specificity was fixed at 80.0%) for TB in children. None of the two single-omics signatures is accurate enough to achieve the optimal sensitivity and specificity for both a triage and diagnostic test for paediatric TB. A 5-feature signature, composed of *GBP6* (Guanylate Binding Protein 6), CCL1/I-309 (C-C motif chemokine 1), CXCL10/IP-10 (C-X-C motif chemokine 10), *GBP5* (Guanylate Binding Protein 5) and IL-7 (Interleukin-7), was identified after integration of the transcriptomic and proteomic datasets. The multi-omics signature achieved an AUC of 97.5% [$CI_{95\%}$ 93.6%-100.0%], sensitivity of 92.9% [$CI_{95\%}$ 78.6%-100.0%] and specificity of 92.3% [$CI_{95\%}$ 80.8%-100.0%] in the test set. When the specificity was set to 98.0%, the sensitivity achieved 85.7% [$CI_{95\%}$ 57.1%-100.0%] in the test set, meeting the optimal values for sensitivity and specificity for a diagnostic test [44]. The signature meets also the optimal WHO TPPs criteria for a triage test (sensitivity achieved 100.0% [$CI_{95\%}$ 78.6%-100.0%] when the specificity was set at 80.0%). This is the first time that a biosignature achieved the optimal performance in terms of sensitivity and specificity for both a diagnostic and a triage test for paediatric TB.

This study is not without limitations. Although the multi-omics signature has been identified in

the training set using a feature selection algorithm with cross-validation to avoid overfitting, and finally tested on the test set, further validation in large independent cohorts is needed to ensure the transferability of the findings. The samples used in this study to generate the data were collected from children recruited in South Africa, Malawi and Kenya. Although this is an African multi-country cohort, it is necessary to evaluate the performance of the signature in cohorts recruited in other countries. Moreover, simultaneous measurement of different types of molecules/biomarkers makes the development of a POCT more challenging and expensive.

Multi-omics integration for biomarker discovery is a novel field that researchers have started exploring just recently. This study has enlightened the potential of sparse multi-omics signatures and the better diagnostic performance they can achieve. Multi-omics signatures are a very promising tool for the development of biomarker-based POCT not just for paediatric TB disease but for many other diseases.

CHAPTER 7

DISCUSSION AND FUTURE DIRECTIONS

This thesis has explored the potential of host -omics profiles as diagnostic tools for paediatric TB through the identification and assessment of candidate diagnostic biomarkers. Chapter 1 has introduced the current global burden of paediatric TB disease, the diagnostic tests, their advantages and limitations as well as how -omics technologies can be used in the field of diagnostics. In Chapter 2, the clinical cohorts that provided the samples for generating the -omics data were introduced, along with a description of the statistical methods and machine learning approaches used to analyse the datasets.

Chapters 3 to 6 have described 4 projects whose common aim was to use -omics data to diagnose TB in children or monitor their treatment response.

Chapter 3 has explored gene expression profiles in children with TB disease and unlikely TB (i.e., other infections with symptomatology resembling TB) at baseline to identify and validate a sparse diagnostic transcriptomic signature. I have demonstrated that a combination of 6 genes meets the minimal WHO requirements of sensitivity and specificity for a biomarker-based triage test for childhood TB [44]. In Chapter 4, I describe the host blood RNA profiles of a longitudinal dataset with samples collected from children with microbiologically confirmed TB and unlikely TB at baseline (i.e., before treatment) and at 3 months (i.e., during treatment). I further tested the performance of the 6-gene signature in this dataset, which achieved optimal performance at baseline with a considerable drop during treatment. These results suggest that genes that are SDE at the time of diagnosis return to normal expression levels during treatment. Gene biomarkers could, therefore, serve not only as diagnostic tools but also to monitor treatment response.

Chapter 5 has explored protein abundance profiles to identify a signature for discriminating between TB disease vs. unlikely TB in children. A 4-protein signature, identified in a multi-country African cohort and validated in independent samples, achieved excellent performance in distinguishing between the two disease groups.

Finally, I have used a pipeline to integrate single-omics datasets and identify a multi-omics signature. Multi-omics signatures, since they include molecules from more than one biological layer in a complementary way, could better capture the heterogeneity of the diseases and achieve improved diagnostic performance. The 5-feature multi-omics signature identified in Chapter 6 was composed of 2 genes and 3 proteins and exceeded the performance of the single-omics transcriptomic and proteomic signatures identified in the same cohort. The combination of 5 biological molecules achieved the optimal WHO requirements for sensitivity and specificity for both a biomarker-based diagnostic and triage test for childhood TB [44].

To my knowledge, this was the first time that sparse transcriptomic and proteomic signatures with such high performances were identified in a multi-country cohort, with HIV-infected and -uninfected children. Although the majority of the TB patients were diagnosed with PTB, children with EPTB were also included in the study. The heterogeneity of the cohort allowed us to assess the performance

of the identified signatures in the clinical subgroups. The whole blood signatures identified in this thesis achieved better performance in distinguishing children with EPTB vs. unlikely TB compared to the PTB cases. This is an extremely useful finding since the current diagnostic tests for paediatric TB underperform in EPTB cases [84].

7.1 Limitations and future directions

The work described in this thesis has some limitations.

Although the cohorts used were certainly heterogeneous, with children recruited in 4 African countries and including HIV-infected and -uninfected patients, further validation in cohorts from other countries and continents is needed to assess the generability of the signatures. Ideally, a diagnostic signature for TB disease should have optimal performance globally enabling the POCT to be utilised without any geographical limitations. One potential future direction for the research presented in this thesis would be to validate the identified biomarkers in heterogeneous and independent cohorts recruited in different countries.

The transcriptomic signature identified and validated in Chapter 3 outperforms the proteomic signature identified and validated in Chapter 5 in terms of AUC, sensitivity and specificity. This might be due to several reasons, such as the smaller discovery cohort used in the proteomic study or the techniques used to generate the data, which have different limits of detection and sensitivity, and affect also the size of the hypothesis space (i.e., untargeted vs. targeted techniques). In the transcriptomic discovery study, in fact, an untargeted technique was used to measure the expression of all the genes present in the biological samples at the time of sampling. Contrarily, the proteomic discovery study was carried out using mostly targeted techniques and, therefore, targeting a limited set of proteins. The larger the set of biomarkers in the discovery phase, the more likely is to select better features in terms of discriminatory power between the groups of interest.

The children recruited in the discovery cohorts were diagnosed either with microbiologically confirmed TB or unlikely TB. Children with unconfirmed TB were not included in the discovery since their diagnosis is uncertain and might introduce biases. Achieving microbiological confirmation in children is difficult because of the paucibacillary form of the paediatric TB disease. A rapid diagnostic test would be extremely useful for these individuals as it would act as another component of the diagnostic toolkit. However, without including this group of patients in the discovery cohorts used to identify the diagnostic biomarkers, we are unlikely to obtain signatures that are optimised to classify these individuals. A future direction of the work presented in Chapters 3 to 6 would be to determine whether extra features could be added to the signatures that can distinguish between TB disease and unlikely TB when the unconfirmed phenotypic group is included. This would overcome a huge limitation of the current diagnostic tests for TB disease and lead to more certain diagnoses of a group of patients which is very hard to achieve microbiologically confirmation for.

The new diagnostic POCTs are moving towards the detection of multiple pathogens simultaneously [317, 318]. The signatures presented in Chapters 3, 5 and 6 are modelled as binary classifiers (i.e., TB disease vs. unlikely TB). Diagnostic decisions are rarely binary. Identifying several binary signatures and performing multiple dichotomous diagnostic tests is impractical. An extension of the work presented in this thesis would be to identify multi-class diagnostic signatures to diagnose more than two distinct diseases simultaneously.

Although Chapter 6 provides proof of principle that multi-omics signatures could outperform single-omics signatures, further analysis and validation studies are needed. Moreover, a multi-omics signature is more challenging and expensive to be translated into a POCT. Multi-omics analyses are surely more comprehensive since they include features from more than one -omics layer. This could help better understand the biology underpinning paediatric TB or other diseases through multi-omics pathway analyses.

7.2 Conclusion

This thesis has demonstrated that, despite the clinical similarities between TB disease and unlikely TB, they are distinguishable at the host molecular level. I have used transcriptomic and proteomic datasets to identify a transcriptomic, a proteomic and a multi-omics signature to discriminate TB disease from unlikely TB in children. Biomarkers and signatures can be translated into POCTs, which are medical tests that can be performed at the bedside, are rapid, cheap and do not require expensive systems or reagents.

These are extremely important findings that pave the way for the development of rapid POCTs for paediatric TB disease. Such tests are highly needed in LMIC where TB is endemic. Moreover, they would ensure that more children receive the diagnosis and treatment that they require more rapidly, and would help reduce the global TB burden.

BIBLIOGRAPHY

- [1] World Health Organization. *Global tuberculosis report 2021*. Report. 2021.
- [2] World Health Organization. *Global tuberculosis report 2022*. Report. 2022.
- [3] W. Cruz-Knight and L. Blake-Gumbs. "Tuberculosis: an overview". In: *Prim Care* 40.3 (2013), pp. 743–56. ISSN: 1558-299X. DOI: 10.1016/j.pop.2013.06.003. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23958367>.
- [4] C. Dye et al. "Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project". In: *JAMA* 282.7 (1999), pp. 677–86. ISSN: 0098-7484. DOI: 10.1001/jama.282.7.677. URL: <https://www.ncbi.nlm.nih.gov/pubmed/10517722>.
- [5] A. Cohen et al. "The global prevalence of latent tuberculosis: a systematic review and meta-analysis". In: *Eur Respir J* 54.3 (2019). ISSN: 1399-3003 (Electronic) 0903-1936 (Linking). DOI: 10.1183/13993003.00655-2019. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31221810>.
- [6] J. L. Flynn and J. Chan. "Immune cell interactions in tuberculosis". In: *Cell* 185.25 (2022), pp. 4682–4702. ISSN: 1097-4172. DOI: 10.1016/j.cell.2022.10.025. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36493751>.
- [7] J. I. Campbell et al. "Childhood Tuberculosis: Historical Perspectives, Recent Advances, and a Call to Action". In: *J Pediatric Infect Dis Soc* 11.Suppl 3 (2022), S63–S66. ISSN: 2048-7207. DOI: 10.1093/jpids/piac051. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36314551>.
- [8] Salmaan Keshavjee and Paul E. Farmer. "Tuberculosis, Drug Resistance, and the History of Modern Medicine". In: *New England Journal of Medicine* 367.10 (2012), pp. 931–936. ISSN: 0028-4793. DOI: 10.1056/nejmra1205429. URL: <https://dx.doi.org/10.1056/NEJMra1205429>.
- [9] S. Fatima et al. "Tuberculosis vaccine: A journey from BCG to present". In: *Life Sci* 252 (2020), p. 117594. ISSN: 1879-0631. DOI: 10.1016/j.lfs.2020.117594. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32305522>.
- [10] A. Roy et al. "Effect of BCG vaccination against *Mycobacterium tuberculosis* infection in children: systematic review and meta-analysis". In: *BMJ* 349 (2014), g4643. ISSN: 1756-1833. DOI: 10.1136/bmj.g4643. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25097193>.
- [11] I. Motta et al. "Recent advances in the treatment of tuberculosis". In: *Clin Microbiol Infect* (2023). ISSN: 1469-0691. DOI: 10.1016/j.cmi.2023.07.013. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37482332>.
- [12] M. Z. Imperial et al. "A patient-level pooled analysis of treatment-shortening regimens for drug-susceptible pulmonary tuberculosis". In: *Nat Med* 24.11 (2018), pp. 1708–1715. ISSN: 1546-170X. DOI: 10.1038/s41591-018-0224-2. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30397355>.
- [13] M. Z. Imperial et al. "Precision-Enhancing Risk Stratification Tools for Selecting Optimal Treatment Durations in Tuberculosis Clinical Trials". In: *Am J Respir Crit Care Med* 204.9 (2021), pp. 1086–1096. ISSN: 1535-4970. DOI: 10.1164/rccm.202101-01170C. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34346856>.
- [14] In: *WHO consolidated guidelines on tuberculosis: Module 4: Treatment - Drug-susceptible tuberculosis treatment*. WHO Guidelines Approved by the Guidelines Review Committee. Review Book. Geneva, 2022. ISBN: 978-92-4-004812-6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35727905>.
- [15] World Health Organization. *Operational Handbook on Tuberculosis. Module 5: Management of Tuberculosis in Children and Adolescents*. Report. 2022.
- [16] A. Turkova et al. "Shorter Treatment for Nonsevere Tuberculosis in African and Indian Children". In: *N Engl J Med* 386.10 (2022), pp. 911–922. ISSN: 1533-4406. DOI: 10.1056/NEJMoa2104535. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35263517>.
- [17] R. L. Goodall et al. "Evaluation of two short standardised regimens for the treatment of rifampicin-resistant tuberculosis (STREAM stage 2): an open-label, multicentre, randomised, non-inferiority trial". In: *Lancet* 400.10366 (2022), pp. 1858–1868. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(22)02078-5. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36368336>.
- [18] B. T. Nyang'wa et al. "A 24-Week, All-Oral Regimen for Rifampin-Resistant Tuberculosis". In: *N Engl J Med* 387.25 (2022), pp. 2331–2343. ISSN: 1533-4406. DOI: 10.1056/NEJMoa2117166. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36546625>.
- [19] A. Esmail et al. "An All-Oral 6-Month Regimen for Multidrug-Resistant Tuberculosis: A Multicenter, Randomized Controlled Clinical Trial (the NExT Study)". In: *Am J Respir Crit Care Med* 205.10 (2022), pp. 1214–1227. ISSN: 1535-4970. DOI: 10.1164/rccm.202107-17790C. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35175905>.

- [20] In: *WHO consolidated guidelines on tuberculosis: Module 4: treatment - drug-resistant tuberculosis treatment, 2022 update*. WHO Guidelines Approved by the Guidelines Review Committee. Review Book. Geneva, 2022. ISBN: 978-92-4-006312-9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36630546>.
- [21] D. D. Chaplin. "Overview of the immune response". In: *J Allergy Clin Immunol* 125.2 Suppl 2 (2010), S3–23. ISSN: 1097-6825. DOI: 10.1016/j.jaci.2009.12.980. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20176265>.
- [22] R. Medzhitov and Jr. Janeway C. A. "Innate immunity: impact on the adaptive immune response". In: *Curr Opin Immunol* 9.1 (1997), pp. 4–9. ISSN: 0952-7915. DOI: 10.1016/s0952-7915(97)80152-5. URL: <https://www.ncbi.nlm.nih.gov/pubmed/9039775>.
- [23] J. Sprent and C. D. Surh. "Generation and maintenance of memory T cells". In: *Curr Opin Immunol* 13.2 (2001), pp. 248–54. ISSN: 0952-7915. DOI: 10.1016/s0952-7915(00)00211-9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/11228420>.
- [24] C. Schiweck et al. "Regulation of CD4(+) and CD8(+) T Cell Biology by Short-Chain Fatty Acids and Its Relevance for Autoimmune Pathology". In: *Int J Mol Sci* 23.15 (2022). ISSN: 1422-0067. DOI: 10.3390/ijms23158272. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35955407>.
- [25] B. B. Finlay and G. McFadden. "Anti-immunology: evasion of the host immune system by bacterial and viral pathogens". In: *Cell* 124.4 (2006), pp. 767–82. ISSN: 0092-8674. DOI: 10.1016/j.cell.2006.01.034. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16497587>.
- [26] A. M. Cadena et al. "Concurrent infection with *Mycobacterium tuberculosis* confers robust protection against secondary infection in macaques". In: *PLoS Pathog* 14.10 (2018), e1007305. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1007305. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30312351>.
- [27] P. Chandra, S. J. Grigsby, and J. A. Philips. "Immune evasion and provocation by *Mycobacterium tuberculosis*". In: *Nat Rev Microbiol* 20.12 (2022), pp. 750–766. ISSN: 1740-1534. DOI: 10.1038/s41579-022-00763-4. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35879556>.
- [28] K. B. Urdahl, S. Shafiani, and J. D. Ernst. "Initiation and regulation of T-cell responses in tuberculosis". In: *Mucosal Immunol* 4.3 (2011), pp. 288–93. ISSN: 1935-3456. DOI: 10.1038/mi.2011.10. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21451503>.
- [29] M. de Martino et al. "Immune Response to *Mycobacterium tuberculosis*: A Narrative Review". In: *Front Pediatr* 7 (2019), p. 350. ISSN: 2296-2360. DOI: 10.3389/fped.2019.00350. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31508399>.
- [30] Y. Li, Y. Wang, and X. Liu. "The role of airway epithelial cells in response to mycobacteria infection". In: *Clin Dev Immunol* 2012 (2012), p. 791392. ISSN: 1740-2530. DOI: 10.1155/2012/791392. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22570668>.
- [31] S. Ehlers and U. E. Schaible. "The granuloma in tuberculosis: dynamics of a host-pathogen collusion". In: *Front Immunol* 3 (2012), p. 411. ISSN: 1664-3224. DOI: 10.3389/fimmu.2012.00411. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23308075>.
- [32] A. O'Garra et al. "The immune response in tuberculosis". In: *Annu Rev Immunol* 31 (2013), pp. 475–527. ISSN: 1545-3278. DOI: 10.1146/annurev-immunol-032712-095939. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23516984>.
- [33] W. W. Reiley et al. "ESAT-6-specific CD4 T cell responses to aerosol *Mycobacterium tuberculosis* infection are initiated in the mediastinal lymph nodes". In: *Proc Natl Acad Sci U S A* 105.31 (2008), pp. 10961–6. ISSN: 1091-6490. DOI: 10.1073/pnas.0801496105. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18667699>.
- [34] N. W. Schluger and W. N. Rom. "The host immune response to tuberculosis". In: *Am J Respir Crit Care Med* 157.3 Pt 1 (1998), pp. 679–91. ISSN: 1073-449X. DOI: 10.1164/ajrccm.157.3.9708002. URL: <https://www.ncbi.nlm.nih.gov/pubmed/9517576>.
- [35] E. Vynnycky and P. E. Fine. "Lifetime risks, incubation period, and serial interval of tuberculosis". In: *Am J Epidemiol* 152.3 (2000), pp. 247–63. ISSN: 0002-9262. DOI: 10.1093/aje/152.3.247. URL: <https://www.ncbi.nlm.nih.gov/pubmed/10933272>.
- [36] E. Whittaker et al. "Examining the Complex Relationship Between Tuberculosis and Other Infectious Diseases in Children". In: *Front Pediatr* 7 (2019), p. 233. ISSN: 2296-2360. DOI: 10.3389/fped.2019.00233. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31294001>.
- [37] A. Pawlowski et al. "Tuberculosis and HIV co-infection". In: *PLoS Pathog* 8.2 (2012), e1002464. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1002464. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22363214>.
- [38] A. M. Cooper. "T cells in mycobacterial infection and disease". In: *Curr Opin Immunol* 21.4 (2009), pp. 378–84. ISSN: 1879-0372. DOI: 10.1016/j.coi.2009.06.004. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19646851>.
- [39] I. M. Orme et al. "Cytokine secretion by CD4 T lymphocytes acquired in response to *Mycobacterium tuberculosis* infection". In: *J Immunol* 151.1 (1993), pp. 518–25. ISSN: 0022-1767. URL: <https://www.ncbi.nlm.nih.gov/pubmed/8100846>.
- [40] J. Chan and J. Flynn. "The immunological aspects of latency in tuberculosis". In: *Clin Immunol* 110.1 (2004), pp. 2–12. ISSN: 1521-6616. DOI: 10.1016/s1521-6616(03)00210-9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/14986673>.

- [41] N. R. Patel et al. "HIV impairs TNF-alpha mediated macrophage apoptotic response to *Mycobacterium tuberculosis*". In: *J Immunol* 179.10 (2007), pp. 6973–80. ISSN: 0022-1767. DOI: 10.4049/jimmunol.179.10.6973. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17982088>.
- [42] P. Chandrasekaran et al. "Malnutrition: Modulator of Immune Responses in Tuberculosis". In: *Front Immunol* 8 (2017), p. 1316. ISSN: 1664-3224. DOI: 10.3389/fimmu.2017.01316. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29093710>.
- [43] N. Mafirakureva et al. "Cost-effectiveness analysis of interventions to improve diagnosis and preventive therapy for paediatric tuberculosis in 9 sub-Saharan African countries: A modelling study". In: *PLoS Med* 20.9 (2023), e1004285. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1004285. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37672524>.
- [44] World Health Organization. *High-priority Target Product Profiles for New Tuberculosis Diagnostic: Report of a Consensus Meeting*. Report. 2014.
- [45] World Health Organization. *Same-day Diagnosis of Tuberculosis by Microscopy: Policy Statement*. Report. 2016.
- [46] J. I. Campbell and D. Menzies. "Testing and Scaling Interventions to Improve the Tuberculosis Infection Care Cascade". In: *J Pediatric Infect Dis Soc* 11.Suppl 3 (2022), S94–S100. ISSN: 2048-7207. DOI: 10.1093/jpids/piac070. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36314552>.
- [47] J. R. Campbell et al. "A systematic review on TST and IGRA tests used for diagnosis of LTBI in immigrants". In: *Mol Diagn Ther* 19.1 (2015), pp. 9–24. ISSN: 1179-2000. DOI: 10.1007/s40291-014-0125-0. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25579159>.
- [48] R. N. van Zyl-Smit et al. "Within-subject variability of interferon-g assay results for tuberculosis and boosting effect of tuberculin skin testing: a systematic review". In: *PLoS One* 4.12 (2009), e8517. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0008517. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20041113>.
- [49] M. Pai, A. Zwerling, and D. Menzies. "Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: an update". In: *Ann Intern Med* 149.3 (2008), pp. 177–84. ISSN: 1539-3704. DOI: 10.7326/0003-4819-149-3-200808050-00241. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18593687>.
- [50] M. Pai et al. "Gamma interferon release assays for detection of *Mycobacterium tuberculosis* infection". In: *Clin Microbiol Rev* 27.1 (2014), pp. 3–20. ISSN: 1098-6618. DOI: 10.1128/CMR.00034-13. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24396134>.
- [51] S. E. Dorman. "New diagnostic tests for tuberculosis: bench, bedside, and beyond". In: *Clin Infect Dis* 50 Suppl 3 (2010), S173–7. ISSN: 1537-6591. DOI: 10.1086/651488. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20397945>.
- [52] P. Desikan. "Sputum smear microscopy in tuberculosis: is it still relevant?" In: *Indian J Med Res* 137.3 (2013), pp. 442–4. ISSN: 0971-5916. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23640550>.
- [53] World Health Organization. *Fluorescent Light-Emitting Diode (LED) microscopy for Diagnosis of Tuberculosis: Policy Statement*. Report. 2011.
- [54] World Health Organization. *Chest Radiography in Tuberculosis Detection*. Report. 2016.
- [55] K. R. Steingart et al. "Sputum processing methods to improve the sensitivity of smear microscopy for tuberculosis: a systematic review". In: *Lancet Infect Dis* 6.10 (2006), pp. 664–74. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(06)70602-8. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17008175>.
- [56] J. M. Hong et al. "Point-of-care diagnostic tests for tuberculosis disease". In: *Sci Transl Med* 14.639 (2022), eabj4124. ISSN: 1946-6242. DOI: 10.1126/scitranslmed.abj4124. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35385338>.
- [57] V. N. Chihota et al. "Liquid vs. solid culture for tuberculosis: performance and cost in a resource-constrained setting". In: *Int J Tuberc Lung Dis* 14.8 (2010), pp. 1024–31. ISSN: 1815-7920. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20626948>.
- [58] S. Rishi et al. "A comparative study for the detection of Mycobacteria by BACTEC MGIT 960, Lowenstein Jensen media and direct AFB smear examination". In: *Indian J Med Microbiol* 25.4 (2007), pp. 383–6. ISSN: 0255-0857. DOI: 10.4103/0255-0857.37344. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18087090>.
- [59] E. MacLean et al. "Advances in Molecular Diagnosis of Tuberculosis". In: *J Clin Microbiol* 58.10 (2020). ISSN: 1098-660X. DOI: 10.1128/JCM.01582-19. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32759357>.
- [60] World Health Organization. *Automated Real-time Nucleic Acid Amplification Technology for Rapid and Simultaneous Detection of Tuberculosis and Rifampicin Resistance: Xpert MTB/RIF System*. Report. 2011.
- [61] K. R. Steingart et al. "Xpert(R) MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults". In: *Cochrane Database Syst Rev* 1 (2014), p. CD009593. ISSN: 1469-493X. DOI: 10.1002/14651858.CD009593.pub3. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24448973>.

- [62] S. D. Lawn and A. I. Zumla. "Diagnosis of extrapulmonary tuberculosis using the Xpert® MTB/RIF assay". In: *Expert Rev Anti Infect Ther* 10.6 (2012), pp. 631–5. ISSN: 1744-8336. DOI: 10.1586/eri.12.43. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22734954>.
- [63] L. Maynard-Smith et al. "Diagnostic accuracy of the Xpert MTB/RIF assay for extrapulmonary and pulmonary tuberculosis when testing non-respiratory samples: a systematic review". In: *BMC Infect Dis* 14 (2014), p. 709. ISSN: 1471-2334. DOI: 10.1186/s12879-014-0709-7. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25599808>.
- [64] World Health Organization. *WHO Consolidated Guidelines on Tuberculosis. Module 3: Diagnosis - Rapid Diagnostics for Tuberculosis Detection, 2021 Update*. Report. 2021.
- [65] M. M. Ninan et al. "Xpert Ultra in diagnosing extrapulmonary TB: accuracy and trace calls". In: *Int J Tuberc Lung Dis* 26.5 (2022), pp. 441–445. ISSN: 1815-7920. DOI: 10.5588/ijtld.21.0480. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35505481>.
- [66] World Health Organization. *Xpert MTB/RIF Implementation Manual: Technical and Operational 'how-to': Practical Considerations*. Report. 2014.
- [67] J. Creswell et al. "Results from early programmatic implementation of Xpert MTB/RIF testing in nine countries". In: *BMC Infect Dis* 14 (2014), p. 2. ISSN: 1471-2334. DOI: 10.1186/1471-2334-14-2. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24383553>.
- [68] N. Raizada et al. "Feasibility of decentralised deployment of Xpert MTB/RIF test at lower level of health system in India". In: *PLoS One* 9.2 (2014), e89301. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0089301. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24586675>.
- [69] T. Agizew et al. "Peripheral clinic versus centralized laboratory-based Xpert MTB/RIF performance: Experience gained from a pragmatic, stepped-wedge trial in Botswana". In: *PLoS One* 12.8 (2017), e0183237. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0183237. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28817643>.
- [70] World Health Organization. *The use of loop-mediated isothermal amplification (TB-LAMP) for the diagnosis of pulmonary tuberculosis*. Report. 2016.
- [71] P. B. Shete et al. "Diagnostic accuracy of TB-LAMP for pulmonary tuberculosis: a systematic review and meta-analysis". In: *BMC Infect Dis* 19.1 (2019), p. 268. ISSN: 1471-2334. DOI: 10.1186/s12879-019-3881-y. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30890135>.
- [72] F. W. Basile et al. "Pediatric Tuberculosis Diagnostics: Present and Future". In: *J Pediatric Infect Dis Soc* 11.Suppl 3 (2022), S85–S93. ISSN: 2048-7207. DOI: 10.1093/jpids/piac082. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36314546>.
- [73] L. M. Parsons et al. "Laboratory diagnosis of tuberculosis in resource-poor countries: challenges and opportunities". In: *Clin Microbiol Rev* 24.2 (2011), pp. 314–50. ISSN: 1098-6618. DOI: 10.1128/CMR.00059-10. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21482728>.
- [74] World Health Organization. *The end TB strategy*. Report. 2015.
- [75] J. R. Campbell and M. L. Bastos. "No time to waste: preventing tuberculosis in children". In: *Lancet* 395.10228 (2020), pp. 924–926. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(20)30532-8. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32199475>.
- [76] L. M. Cranmer et al. "What's Old and New in Tuberculosis Vaccines for Children". In: *J Pediatric Infect Dis Soc* 11.Suppl 3 (2022), S110–S116. ISSN: 2048-7207. DOI: 10.1093/jpids/piac078. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36314550>.
- [77] C. Lienhardt et al. "Risk factors for tuberculosis infection in children in contact with infectious tuberculosis cases in the Gambia, West Africa". In: *Pediatrics* 111.5 Pt 1 (2003), e608–14. ISSN: 1098-4275. DOI: 10.1542/peds.111.5.e608. URL: <https://www.ncbi.nlm.nih.gov/pubmed/12728119>.
- [78] B. J. Marais et al. "The clinical epidemiology of childhood pulmonary tuberculosis: a critical review of literature from the pre-chemotherapy era". In: *Int J Tuberc Lung Dis* 8.3 (2004), pp. 278–85. ISSN: 1027-3719. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15139465>.
- [79] M. C. Raviglione, Jr. Snider D. E., and A. Kochi. "Global epidemiology of tuberculosis. Morbidity and mortality of a worldwide epidemic". In: *JAMA* 273.3 (1995), pp. 220–6. ISSN: 0098-7484. URL: <https://www.ncbi.nlm.nih.gov/pubmed/7807661>.
- [80] B. J. Marais et al. "The burden of childhood tuberculosis: a public health perspective". In: *Int J Tuberc Lung Dis* 9.12 (2005), pp. 1305–13. ISSN: 1027-3719. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16466051>.
- [81] D. Shingadia and V. Novelli. "Diagnosis and treatment of tuberculosis in children". In: *Lancet Infect Dis* 3.10 (2003), pp. 624–32. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(03)00771-0. URL: <https://www.ncbi.nlm.nih.gov/pubmed/14522261>.
- [82] A. T. Cruz and J. R. Starke. "Clinical manifestations of tuberculosis in children". In: *Paediatr Respir Rev* 8.2 (2007), pp. 107–17. ISSN: 1526-0542. DOI: 10.1016/j.prrv.2007.04.008. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17574154>.
- [83] J. A. Seddon et al. "The evolving research agenda for paediatric tuberculosis infection". In: *Lancet Infect Dis* 19.9 (2019), e322–e329. ISSN: 1474-4457. DOI: 10.1016/S1473-3099(18)30787-4. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31221543>.
- [84] M. Kaforou et al. "Transcriptomics for child and adolescent tuberculosis". In: *Immunol Rev* 309.1 (2022), pp. 97–122. ISSN: 1600-065X. DOI: 10.1111/imr.13116. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35818983>.

- [85] E. Whittaker et al. "Regulatory T Cells and Pro-inflammatory Responses Predominate in Children with Tuberculosis". In: *Front Immunol* 8 (2017), p. 448. ISSN: 1664-3224. DOI: 10.3389/fimmu.2017.00448. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28487695>.
- [86] R. Basu Roy et al. "Tuberculosis susceptibility and protection in children". In: *Lancet Infect Dis* 19.3 (2019), e96–e108. ISSN: 1474-4457. DOI: 10.1016/S1473-3099(18)30157-9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30322790>.
- [87] J. A. Seddon et al. "The Wonder Years: What Can Primary School Children Teach Us About Immunity to *Mycobacterium tuberculosis*?" In: *Front Immunol* 9 (2018), p. 2946. ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.02946. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30619306>.
- [88] S. M. Newton et al. "Paediatric tuberculosis". In: *Lancet Infect Dis* 8.8 (2008), pp. 498–510. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(08)70182-8. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18652996>.
- [89] H. E. Jenkins et al. "Mortality in children diagnosed with tuberculosis: a systematic review and meta-analysis". In: *Lancet Infect Dis* 17.3 (2017), pp. 285–295. ISSN: 1474-4457. DOI: 10.1016/S1473-3099(16)30474-1. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27964822>.
- [90] L. Martinez et al. "The risk of tuberculosis in children after close exposure: a systematic review and individual-participant meta-analysis". In: *Lancet* 395.10228 (2020), pp. 973–984. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(20)30166-5. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32199484>.
- [91] P. J. Dodd et al. "The global burden of tuberculosis mortality in children: a mathematical modelling study". In: *Lancet Glob Health* 5.9 (2017), e898–e906. ISSN: 2214-109X. DOI: 10.1016/S2214-109X(17)30289-9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28807188>.
- [92] B. J. Vonasek et al. "Tuberculosis in Children Living With HIV: Ongoing Progress and Challenges". In: *J Pediatric Infect Dis Soc* 11.Suppl 3 (2022), S72–S78. ISSN: 2048-7207. DOI: 10.1093/jpids/piac060. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36314545>.
- [93] World Health Organization. "BCG vaccines: WHO position paper". In: *Wkly Epidemiol Rec* 93.8 (2018), pp. 73–96. ISSN: 0049-8114. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29474026>.
- [94] S. Sahu, E. Wandwalo, and N. Arinaminpathy. "Exploring the Impact of the COVID-19 Pandemic on Tuberculosis Care and Prevention". In: *J Pediatric Infect Dis Soc* 11.Suppl 3 (2022), S67–S71. ISSN: 2048-7207. DOI: 10.1093/jpids/piac102. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36314548>.
- [95] P. Goussard et al. "COVID-19 in a child with tuberculous airway compression". In: *Pediatr Pulmonol* 55.9 (2020), pp. 2201–2203. ISSN: 1099-0496. DOI: 10.1002/ppul.24927. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32662947>.
- [96] B. J. Freij et al. "Fatal central nervous system co-infection with SARS-CoV-2 and tuberculosis in a healthy child". In: *BMC Pediatr* 20.1 (2020), p. 429. ISSN: 1471-2431. DOI: 10.1186/s12887-020-02308-1. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32907595>.
- [97] T. G. Connell, H. J. Zar, and M. P. Nicol. "Advances in the diagnosis of pulmonary tuberculosis in HIV-infected and HIV-uninfected children". In: *J Infect Dis* 204 Suppl 4 (2011), S1151–8. ISSN: 1537-6613. DOI: 10.1093/infdis/jir413. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21996697>.
- [98] J. Gaensbauer and R. Broadhurst. "Recent Innovations in Diagnosis and Treatment of Pediatric Tuberculosis". In: *Curr Infect Dis Rep* 21.1 (2019), p. 4. ISSN: 1523-3847. DOI: 10.1007/s11908-019-0662-0. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30767077>.
- [99] H. J. Zar et al. "Tuberculosis Diagnosis in Children Using Xpert Ultra on Different Respiratory Specimens". In: *Am J Respir Crit Care Med* 200.12 (2019), pp. 1531–1538. ISSN: 1535-4970. DOI: 10.1164/rccm.201904-0772OC. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31381861>.
- [100] W. Zheng et al. "Diagnosis of paediatric tuberculosis by optically detecting two virulence factors on extracellular vesicles in blood samples". In: *Nat Biomed Eng* 6.8 (2022), pp. 979–991. ISSN: 2157-846X. DOI: 10.1038/s41551-022-00922-1. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35986185>.
- [101] F. Crick. "Central dogma of molecular biology". In: *Nature* 227.5258 (1970), pp. 561–3. ISSN: 0028-0836. DOI: 10.1038/227561a0. URL: <https://www.ncbi.nlm.nih.gov/pubmed/4913914>.
- [102] M. Morange. "What history tells us XIII. Fifty years of the Central Dogma". In: *J Biosci* 33.2 (2008), pp. 171–5. ISSN: 0250-5991. DOI: 10.1007/s12038-008-0034-7. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18535351>.
- [103] J. M. Coffin and H. Fan. "The Discovery of Reverse Transcriptase". In: *Annu Rev Virol* 3.1 (2016), pp. 29–51. ISSN: 2327-0578. DOI: 10.1146/annurev-virology-110615-035556. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27482900>.
- [104] A. Travers and G. Muskhelishvili. "DNA structure and function". In: *FEBS J* 282.12 (2015), pp. 2279–95. ISSN: 1742-4658. DOI: 10.1111/febs.13307. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25903461>.
- [105] Harvey F. Lodish. *Molecular cell biology*. Ninth edition. Austin: Macmillan Learning, 2021, 1 volume (various pagings). ISBN: 9781319208523.
- [106] S. Clancy and W. Brown. "Translation: DNA to mRNA to Protein". In: *Nature Education* (2008).

- [107] World Health Organization. *Biomarkers in risk assessment: validity and validation*. Report. 2001.
- [108] A. Conesa and S. Beck. "Making multi-omics data accessible to researchers". In: *Sci Data* 6.1 (2019), p. 251. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0258-4. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31672978>.
- [109] Y. Hasin, M. Seldin, and A. Lusi. "Multi-omics approaches to disease". In: *Genome Biol* 18.1 (2017), p. 83. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1215-1. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28476144>.
- [110] R. P. J. Lai et al. "Transcriptomic Characterization of Tuberculous Sputum Reveals a Host Warburg Effect and Microbial Cholesterol Catabolism". In: *mBio* 12.6 (2021), e0176621. ISSN: 2150-7511. DOI: 10.1128/mBio.01766-21. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34872348>.
- [111] A. Deffur, R. J. Wilkinson, and A. K. Coussens. "Tricks to translating TB transcriptomics". In: *Ann Transl Med* 3.Suppl 1 (2015), S43. ISSN: 2305-5839. DOI: 10.3978/j.issn.2305-5839.2015.04.12. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26046091>.
- [112] O. Ramilo et al. "Gene expression patterns in blood leukocytes discriminate patients with acute infections". In: *Blood* 109.5 (2007), pp. 2066–77. ISSN: 0006-4971. DOI: 10.1182/blood-2006-02-002477. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17105821>.
- [113] J. Aaroe et al. "Gene expression profiling of peripheral blood cells for early detection of breast cancer". In: *Breast Cancer Res* 12.1 (2010), R7. ISSN: 1465-542X. DOI: 10.1186/bcr2472. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20078854>.
- [114] T. Zander et al. "Blood-based gene expression signatures in non-small cell lung cancer". In: *Clin Cancer Res* 17.10 (2011), pp. 3360–7. ISSN: 1557-3265. DOI: 10.1158/1078-0432.CCR-10-0533. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21558400>.
- [115] D. H. Chang et al. "The effect of lung cancer on cytokine expression in peripheral blood mononuclear cells". In: *PLoS One* 8.6 (2013), e64456. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0064456. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23762239>.
- [116] M. Gurevich et al. "Prediction of acute multiple sclerosis relapses by transcription levels of peripheral blood cells". In: *BMC Med Genomics* 2 (2009), p. 46. ISSN: 1755-8794. DOI: 10.1186/1755-8794-2-46. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19624813>.
- [117] A. Calon et al. "Stromal gene expression defines poor-prognosis subtypes in colorectal cancer". In: *Nat Genet* 47.4 (2015), pp. 320–9. ISSN: 1546-1718. DOI: 10.1038/ng.3225. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25706628>.
- [118] T. Lequerre et al. "Gene profiling in white blood cells predicts infliximab responsiveness in rheumatoid arthritis". In: *Arthritis Res Ther* 8.4 (2006), R105. ISSN: 1478-6362. DOI: 10.1186/ar1990. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16817978>.
- [119] Richard P Horgan and Louise C Kenny. "Omic technologies: genomics, transcriptomics, proteomics and metabolomics". In: *The Obstetrician & Gynaecologist* 13.3 (2011), pp. 189–195. DOI: <https://doi.org/10.1576/toag.13.3.189.27672>. eprint: <https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1576/toag.13.3.189.27672>. URL: <https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1576/toag.13.3.189.27672>.
- [120] A. Menke et al. "Peripheral blood gene expression: it all boils down to the RNA collection tubes". In: *BMC Res Notes* 5 (2012), p. 1. ISSN: 1756-0500. DOI: 10.1186/1756-0500-5-1. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22214347>.
- [121] T. Nikula et al. "Genome-wide comparison of two RNA-stabilizing reagents for transcriptional profiling of peripheral blood". In: *Transl Res* 161.3 (2013), pp. 181–8. ISSN: 1878-1810. DOI: 10.1016/j.trsl.2012.10.003. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23138105>.
- [122] A. H. Skogholt et al. "Gene expression differences between PAXgene and Tempus blood RNA tubes are highly reproducible between independent samples and biobanks". In: *BMC Res Notes* 10.1 (2017), p. 136. ISSN: 1756-0500. DOI: 10.1186/s13104-017-2455-6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28335817>.
- [123] J. A. Carrillo-Avila et al. "Evaluation of RNA purification methods by using different blood stabilization tubes: identification of key features for epidemiological studies". In: *BMC Res Notes* 13.1 (2020), p. 77. ISSN: 1756-0500. DOI: 10.1186/s13104-020-04943-4. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32070402>.
- [124] M. Puchta, M. Boczkowska, and J. Groszyk. "Low RIN Value for RNA-Seq Library Construction from Long-Term Stored Seeds: A Case Study of Barley Seeds". In: *Genes (Basel)* 11.10 (2020). ISSN: 2073-4425. DOI: 10.3390/genes11101190. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33066221>.
- [125] C. Broderick et al. "Host transcriptional response to TB preventive therapy differentiates two sub-groups of IGRA-positive individuals". In: *Tuberculosis (Edinb)* 127 (2021), p. 102033. ISSN: 1873-281X. DOI: 10.1016/j.tube.2020.102033. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33524936>.
- [126] J. C. Venter et al. "The sequence of the human genome". In: *Science* 291.5507 (2001), pp. 1304–51. ISSN: 0036-8075. DOI: 10.1126/science.1058040. URL: <https://www.ncbi.nlm.nih.gov/pubmed/11181995>.

- [127] E. S. Lander et al. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (2001), pp. 860–921. ISSN: 0028-0836. DOI: 10.1038/35057062. URL: <https://www.ncbi.nlm.nih.gov/pubmed/11237011>.
- [128] T. W. Chang. "Binding of cells to matrixes of distinct antibodies coated on solid surface". In: *J Immunol Methods* 65.1-2 (1983), pp. 217–23. ISSN: 0022-1759. DOI: 10.1016/0022-1759(83)90318-6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/6606681>.
- [129] P. A. Bryant et al. "Chips with everything: DNA microarrays in infectious diseases". In: *Lancet Infect Dis* 4.2 (2004), pp. 100–11. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(04)00930-2. URL: <https://www.ncbi.nlm.nih.gov/pubmed/14871635>.
- [130] G. Stefani and F. J. Slack. "Small non-coding RNAs in animal development". In: *Nat Rev Mol Cell Biol* 9.3 (2008), pp. 219–30. ISSN: 1471-0080. DOI: 10.1038/nrm2347. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18270516>.
- [131] K. R. Kukurba and S. B. Montgomery. "RNA Sequencing and Analysis". In: *Cold Spring Harb Protoc* 2015.11 (2015), pp. 951–69. ISSN: 1559-6095. DOI: 10.1101/pdb.top084970. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25870306>.
- [132] C. W. Fuller et al. "The challenges of sequencing by synthesis". In: *Nat Biotechnol* 27.11 (2009), pp. 1013–23. ISSN: 1546-1696. DOI: 10.1038/nbt.1585. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19898456>.
- [133] A. Modi et al. "The Illumina Sequencing Protocol and the NovaSeq 6000 System". In: *Methods Mol Biol* 2242 (2021), pp. 15–42. ISSN: 1940-6029. DOI: 10.1007/978-1-0716-1099-2_2. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33961215>.
- [134] Z. Wang, M. Gerstein, and M. Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nat Rev Genet* 10.1 (2009), pp. 57–63. ISSN: 1471-0064. DOI: 10.1038/nrg2484. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19015660>.
- [135] L. M. Verhagen et al. "A predictive signature gene set for discriminating active from latent tuberculosis in Warao Amerindian children". In: *BMC Genomics* 14 (2013), p. 74. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-74. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23375113>.
- [136] S. Dhanasekaran et al. "Identification of biomarkers for *Mycobacterium tuberculosis* infection and disease in BCG-vaccinated young children in Southern India". In: *Genes Immun* 14.6 (2013), pp. 356–64. ISSN: 1476-5470. DOI: 10.1038/gene.2013.26. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23676757>.
- [137] S. T. Anderson et al. "Diagnosis of childhood tuberculosis and host RNA expression in Africa". In: *N Engl J Med* 370.18 (2014), pp. 1712–1723. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1303657. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24785206>.
- [138] Q. Li et al. "Increased IL-9 mRNA expression as a biomarker to diagnose childhood tuberculosis in a high burden settings". In: *J Infect* 71.2 (2015), pp. 273–6. ISSN: 1532-2742. DOI: 10.1016/j.jinf.2015.04.002. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25869536>.
- [139] J. X. Wang et al. "Diagnostic values of microRNA-31 in peripheral blood mononuclear cells for pediatric pulmonary tuberculosis in Chinese patients". In: *Genet Mol Res* 14.4 (2015), pp. 17235–43. ISSN: 1676-5680. DOI: 10.4238/2015.December.16.23. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26681217>.
- [140] D. E. Zak et al. "A blood RNA signature for tuberculosis disease risk: a prospective cohort study". In: *Lancet* 387.10035 (2016), pp. 2312–2322. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(15)01316-1. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27017310>.
- [141] H. A. Fletcher et al. "Human newborn bacille Calmette-Guerin vaccination and risk of tuberculosis disease: a case-control study". In: *BMC Med* 14 (2016), p. 76. ISSN: 1741-7015. DOI: 10.1186/s12916-016-0617-3. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27183822>.
- [142] S. Jenum et al. "Approaching a diagnostic point-of-care test for pediatric tuberculosis through evaluation of immune biomarkers across the clinical disease spectrum". In: *Sci Rep* 6 (2016), p. 18520. ISSN: 2045-2322. DOI: 10.1038/srep18520. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26725873>.
- [143] M. Zhou et al. "Circulating microRNAs as biomarkers for the early diagnosis of childhood tuberculosis infection". In: *Mol Med Rep* 13.6 (2016), pp. 4620–6. ISSN: 1791-3004. DOI: 10.3892/mmr.2016.5097. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27082104>.
- [144] J. E. Gjoen et al. "Novel transcriptional signatures for sputum-independent diagnostics of tuberculosis in children". In: *Sci Rep* 7.1 (2017), p. 5839. ISSN: 2045-2322. DOI: 10.1038/s41598-017-05057-x. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28724962>.
- [145] C. Hemingway et al. "Childhood tuberculosis is associated with decreased abundance of T cell gene transcripts and impaired T cell function". In: *PLoS One* 12.11 (2017), e0185973. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0185973. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29140996>.
- [146] U. K. Rohlwick et al. "Tuberculous meningitis in children is characterized by compartmentalized immune responses and neural excitotoxicity". In: *Nat Commun* 10.1 (2019), p. 3767. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11783-9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31434901>.

- [147] A. Penn-Nicholson et al. "RISK6, a 6-gene transcriptomic signature of TB disease risk, diagnosis and treatment response". In: *Sci Rep* 10.1 (2020), p. 8629. ISSN: 2045-2322. DOI: 10.1038/s41598-020-65043-8. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32451443>.
- [148] J. A. Tornheim et al. "Transcriptomic Profiles of Confirmed Pediatric Tuberculosis Patients and Household Contacts Identifies Active Tuberculosis, Infection, and Treatment Response Among Indian Children". In: *J Infect Dis* 221.10 (2020), pp. 1647–1658. ISSN: 1537-6613. DOI: 10.1093/infdis/jiz639. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31796955>.
- [149] W. E. Johnson et al. "Comparing tuberculosis gene signatures in malnourished individuals using the TBSignatureProfiler". In: *BMC Infect Dis* 21.1 (2021), p. 106. ISSN: 1471-2334. DOI: 10.1186/s12879-020-05598-z. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33482742>.
- [150] M. P. Berry et al. "An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis". In: *Nature* 466.7309 (2010), pp. 973–7. ISSN: 1476-4687. DOI: 10.1038/nature09247. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20725040>.
- [151] J. Maertzdorf et al. "Human gene expression profiles of susceptibility and resistance in tuberculosis". In: *Genes Immun* 12.1 (2011), pp. 15–22. ISSN: 1476-5470. DOI: 10.1038/gene.2010.51. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20861863>.
- [152] M. C. Haks et al. "Focused human gene expression profiling using dual-color reverse transcriptase multiplex ligation-dependent probe amplification". In: *Vaccine* 33.40 (2015), pp. 5282–8. ISSN: 1873-2518. DOI: 10.1016/j.vaccine.2015.04.054. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25917681>.
- [153] V. Roth. "The generalized LASSO". In: *IEEE Trans Neural Netw* 15.1 (2004), pp. 16–28. ISSN: 1045-9227. DOI: 10.1109/TNN.2003.809398. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15387244>.
- [154] T. E. Sweeney et al. "Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis". In: *Lancet Respir Med* 4.3 (2016), pp. 213–24. ISSN: 2213-2619. DOI: 10.1016/S2213-2600(16)00048-5. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26907218>.
- [155] J. S. Sutherland et al. "Diagnostic Accuracy of the Cepheid 3-gene Host Response Fingerstick Blood Test in a Prospective, Multi-site Study: Interim Results". In: *Clin Infect Dis* 74.12 (2022), pp. 2136–2141. ISSN: 1537-6591. DOI: 10.1093/cid/ciab839. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34550342>.
- [156] L. Berteloot et al. "Value of chest X-ray in TB diagnosis in HIV-infected children living in resource-limited countries: the ANRS 12229-PAANTHER 01 study". In: *Int J Tuberc Lung Dis* 22.8 (2018), pp. 844–850. ISSN: 1815-7920. DOI: 10.5588/ijtld.18.0122. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29991391>.
- [157] S. Machingaidze 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". In: *Pediatr Infect Dis J* 30.8 (2011), pp. 694–700. ISSN: 1532-0987. DOI: 10.1097/INF.0b013e318214b915. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21427627>.
- [158] A. W. Kay et al. "Xpert MTB/RIF and Xpert MTB/RIF Ultra assays for active tuberculosis and rifampicin resistance in children". In: *Cochrane Database Syst Rev* 8 (2020), p. CD013359. ISSN: 1469-493X. DOI: 10.1002/14651858.CD013359.pub2. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32853411>.
- [159] R. Bayaa et al. "Multi-country evaluation of RISK6, a 6-gene blood transcriptomic signature, for tuberculosis diagnosis and treatment monitoring". In: *Sci Rep* 11.1 (2021), p. 13646. ISSN: 2045-2322. DOI: 10.1038/s41598-021-93059-1. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34211042>.
- [160] D. J. Horne et al. "Sputum monitoring during tuberculosis treatment for predicting outcome: systematic review and meta-analysis". In: *Lancet Infect Dis* 10.6 (2010), pp. 387–94. ISSN: 1474-4457. DOI: 10.1016/S1473-3099(10)70071-2. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20510279>.
- [161] E. G. Thompson et al. "Host blood RNA signatures predict the outcome of tuberculosis treatment". In: *Tuberculosis (Edinb)* 107 (2017), pp. 48–58. ISSN: 1873-281X. DOI: 10.1016/j.tube.2017.08.004. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29050771>.
- [162] C. I. Bloom et al. "Detectable changes in the blood transcriptome are present after two weeks of antituberculosis therapy". In: *PLoS One* 7.10 (2012), e46191. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0046191. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23056259>.
- [163] J. M. Cliff et al. "Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response". In: *J Infect Dis* 207.1 (2013), pp. 18–29. ISSN: 1537-6613. DOI: 10.1093/infdis/jis499. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22872737>.
- [164] M. Jacobsen et al. "Novel strategies to identify biomarkers in tuberculosis". In: *Biol Chem* 389.5 (2008), pp. 487–95. ISSN: 1431-6730. DOI: 10.1515/bc.2008.053. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18953715>.
- [165] S. Ray et al. "Proteomic technologies for the identification of disease biomarkers in serum: advances and challenges ahead". In: *Proteomics* 11.11 (2011), pp. 2139–61. ISSN: 1615-9861. DOI: 10.1002/pmic.201000460. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21548090>.

- [166] S. Lista, F. Faltraco, and H. Hampel. "Biological and methodical challenges of blood-based proteomics in the field of neurological research". In: *Prog Neurobiol* 101-102 (2013), pp. 18–34. ISSN: 1873-5118. DOI: 10.1016/j.pneurobio.2012.06.006. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22743551>.
- [167] J. L. Luque-Garcia and T. A. Neubert. "Sample preparation for serum/plasma profiling and biomarker identification by mass spectrometry". In: *J Chromatogr A* 1153.1-2 (2007), pp. 259–76. ISSN: 0021-9673. DOI: 10.1016/j.chroma.2006.11.054. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17166507>.
- [168] A. H. Ren, E. P. Diamandis, and V. Kulasingam. "Uncovering the Depths of the Human Proteome: Antibody-based Technologies for Ultrasensitive Multiplexed Protein Detection and Quantification". In: *Mol Cell Proteomics* 20 (2021), p. 100155. ISSN: 1535-9484. DOI: 10.1016/j.mcpro.2021.100155. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34597790>.
- [169] K. Shah and P. Maghsoudlou. "Enzyme-linked immunosorbent assay (ELISA): the basics". In: *Br J Hosp Med (Lond)* 77.7 (2016), pp. C98–101. ISSN: 1750-8460. DOI: 10.12968/hmed.2016.77.7.C98. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27388394>.
- [170] T. Koudriavtseva et al. "Coagulation/Complement Activation and Cerebral Hypoperfusion in Relapsing-Remitting Multiple Sclerosis". In: *Front Immunol* 11 (2020), p. 548604. ISSN: 1664-3224. DOI: 10.3389/fimmu.2020.548604. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33193314>.
- [171] J. C. Rohloff et al. "Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents". In: *Mol Ther Nucleic Acids* 3 (2014), e201. ISSN: 2162-2531. DOI: 10.1038/mtna.2014.49. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25291143>.
- [172] V. Thiviyanathan and D. G. Gorenstein. "Aptamers and the next generation of diagnostic reagents". In: *Proteomics Clin Appl* 6.11-12 (2012), pp. 563–73. ISSN: 1862-8354. DOI: 10.1002/prca.201200042. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23090891>.
- [173] J. Sharifi et al. "Improving monoclonal antibody pharmacokinetics via chemical modification". In: *Q J Nucl Med* 42.4 (1998), pp. 242–9. ISSN: 1125-0135. URL: <https://www.ncbi.nlm.nih.gov/pubmed/9973839>.
- [174] A. Constantinou, C. Chen, and M. P. Deonarain. "Modulating the pharmacokinetics of therapeutic antibodies". In: *Biotechnol Lett* 32.5 (2010), pp. 609–22. ISSN: 1573-6776. DOI: 10.1007/s10529-010-0214-z. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20131077>.
- [175] R. Jacobs et al. "Identification of novel host biomarkers in plasma as candidates for the immunodiagnosis of tuberculosis disease and monitoring of tuberculosis treatment response". In: *Oncotarget* 7.36 (2016), pp. 57581–57592. ISSN: 1949-2553. DOI: 10.18632/oncotarget.11420. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27557501>.
- [176] N. N. Chegou et al. "Diagnostic performance of a seven-marker serum protein biosignature for the diagnosis of active TB disease in African primary healthcare clinic attendees with signs and symptoms suggestive of TB". In: *Thorax* 71.9 (2016), pp. 785–94. ISSN: 1468-3296. DOI: 10.1136/thoraxjnl-2015-207999. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27146200>.
- [177] N. N. Chegou et al. "Africa-wide evaluation of host biomarkers in QuantiFERON supernatants for the diagnosis of pulmonary tuberculosis". In: *Sci Rep* 8.1 (2018), p. 2675. ISSN: 2045-2322. DOI: 10.1038/s41598-018-20855-7. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29422548>.
- [178] N. Alsleben et al. "Interferon-gamma inducible protein 10 as a biomarker for active tuberculosis and latent tuberculosis infection in children: a case-control study". In: *Scand J Infect Dis* 44.4 (2012), pp. 256–62. ISSN: 1651-1980. DOI: 10.3109/00365548.2011.632644. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22103555>.
- [179] N. N. Chegou et al. "Utility of host markers detected in Quantiferon supernatants for the diagnosis of tuberculosis in children in a high-burden setting". In: *PLoS One* 8.5 (2013), e64226. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0064226. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23691173>.
- [180] N. Pavan Kumar et al. "Circulating biomarkers of pulmonary and extrapulmonary tuberculosis in children". In: *Clin Vaccine Immunol* 20.5 (2013), pp. 704–11. ISSN: 1556-679X. DOI: 10.1128/CVI.00038-13. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23486418>.
- [181] I. Latorre et al. "IP-10 is an accurate biomarker for the diagnosis of tuberculosis in children". In: *J Infect* 69.6 (2014), pp. 590–9. ISSN: 1532-2742. DOI: 10.1016/j.jinf.2014.06.013. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24975172>.
- [182] M. Tebruegge et al. "Mycobacteria-Specific Cytokine Responses Detect Tuberculosis Infection and Distinguish Latent from Active Tuberculosis". In: *Am J Respir Crit Care Med* 192.4 (2015), pp. 485–99. ISSN: 1535-4970. DOI: 10.1164/rccm.201501-00590C. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26030187>.
- [183] E. L. Sudbury et al. "Mycobacterium tuberculosis-specific cytokine biomarkers for the diagnosis of childhood TB in a TB-endemic setting". In: *J Clin Tuberc Other Mycobact Dis* 16 (2019), p. 100102. ISSN: 2405-5794. DOI: 10.1016/j.jctube.2019.100102. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31720428>.

- [184] C. M. Manyelo et al. "Potential of Host Serum Protein Biomarkers in the Diagnosis of Tuberculous Meningitis in Children". In: *Front Pediatr* 7 (2019), p. 376. ISSN: 2296-2360. DOI: 10.3389/fped.2019.00376. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31612118>.
- [185] V. V. S. Albuquerque et al. "Plasma levels of C-reactive protein, matrix metalloproteinase-7 and lipopolysaccharide-binding protein distinguish active pulmonary or extrapulmonary tuberculosis from uninfected controls in children". In: *Cytokine* 123 (2019), p. 154773. ISSN: 1096-0023. DOI: 10.1016/j.cyto.2019.154773. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31299414>.
- [186] T. Togun et al. "A three-marker protein biosignature distinguishes tuberculosis from other respiratory diseases in Gambian children". In: *EBioMedicine* 58 (2020), p. 102909. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2020.102909. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32711253>.
- [187] M. Druszczyńska et al. "Cytokine Biosignature of Active and Latent *Mycobacterium tuberculosis* Infection in Children". In: *Pathogens* 10.5 (2021). ISSN: 2076-0817. DOI: 10.3390/pathogens10050517. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33923293>.
- [188] N. P. Kumar et al. "Plasma chemokines as immune biomarkers for diagnosis of pediatric tuberculosis". In: *BMC Infect Dis* 21.1 (2021), p. 1055. ISSN: 1471-2334. DOI: 10.1186/s12879-021-06749-6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34635070>.
- [189] A. Nehme. "The Fundamentals for Studying Local Renin-Angiotensin System: Experience from Multi-omics Studies on Atherosclerosis". In: *EC Cardiology* (2020). eprint: <https://www.eicon.com/eccy/pdf/ECCY-06-00462.pdf>.
- [190] C. X. Li et al. "Integration of multi-omics datasets enables molecular classification of COPD". In: *Eur Respir J* 51.5 (2018). ISSN: 1399-3003. DOI: 10.1183/13993003.01930-2017. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29545283>.
- [191] J. C. Venter, H. O. Smith, and M. D. Adams. "The Sequence of the Human Genome". In: *Clin Chem* 61.9 (2015), pp. 1207–8. ISSN: 1530-8561. DOI: 10.1373/clinchem.2014.237016. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26185218>.
- [192] B. E. Bernstein, A. Meissner, and E. S. Lander. "The mammalian epigenome". In: *Cell* 128.4 (2007), pp. 669–81. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.01.033. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17320505>.
- [193] R. Jaenisch and A. Bird. "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals". In: *Nat Genet* 33 Suppl (2003), pp. 245–54. ISSN: 1061-4036. DOI: 10.1038/ng1089. URL: <https://www.ncbi.nlm.nih.gov/pubmed/12610534>.
- [194] M. Wilhelm et al. "Mass-spectrometry-based draft of the human proteome". In: *Nature* 509.7502 (2014), pp. 582–7. ISSN: 1476-4687. DOI: 10.1038/nature13319. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24870543>.
- [195] S. G. Villas-Boas et al. "Mass spectrometry in metabolome analysis". In: *Mass Spectrom Rev* 24.5 (2005), pp. 613–46. ISSN: 0277-7037. DOI: 10.1002/mas.20032. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15389842>.
- [196] Y. F. Sun and Y. J. Hu. "Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases". In: *Adv Genet* 93 (2016), pp. 147–90. ISSN: 0065-2660. DOI: 10.1016/bs.adgen.2015.11.004. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26915271>.
- [197] M. Chagoyen, J. A. G. Ranea, and F. Pazos. "Applications of molecular networks in biomedicine". In: *Biol Methods Protoc* 4.1 (2019), bpz012. ISSN: 2396-8923. DOI: 10.1093/biomethods/bpz012. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32395629>.
- [198] S. Boccaletti et al. "The structure and dynamics of multilayer networks". In: *Phys Rep* 544.1 (2014), pp. 1–122. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2014.07.001. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32834429>.
- [199] M. Zitnik and J. Leskovec. "Predicting multicellular function through multi-layer tissue networks". In: *Bioinformatics* 33.14 (July 2017), pp. i190–i198. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx252. eprint: <https://academic.oup.com/bioinformatics/article-pdf/33/14/i190/25157097/btx252.pdf>. URL: <https://doi.org/10.1093/bioinformatics/btx252>.
- [200] F. McGee et al. "The State of the Art in Multilayer Network Visualization". In: *Computer Graphics Forum* 38.6 (2019), pp. 125–149. DOI: <https://doi.org/10.1111/cgf.13610>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13610>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13610>.
- [201] V. N. Kristensen et al. "Principles and methods of integrative genomic analyses in cancer". In: *Nat Rev Cancer* 14.5 (2014), pp. 299–313. ISSN: 1474-1768. DOI: 10.1038/nrc3721. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24759209>.
- [202] M. Bersanelli et al. "Methods for the integration of multi-omics data: mathematical aspects". In: *BMC Bioinformatics* 17 Suppl 2 (2016), p. 15. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0857-9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26821531>.
- [203] F. Rohart et al. "mixOmics: An R package for 'omics feature selection and multiple data integration". In: *PLoS Comput Biol* 13.11 (2017), e1005752. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005752. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29099853>.

- [204] I. Subramanian et al. “Multi-omics Data Integration, Interpretation, and Its Application”. In: *Bioinform Biol Insights* 14 (2020), p. 1177932219899051. ISSN: 1177-9322. DOI: 10.1177/1177932219899051. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32076369>.
- [205] Chen Y. et al. “Similarity-based Classification: Concepts and Algorithms”. In: *Journal of Machine Learning Research* 10.27 (2009), pp. 747–776. URL: <http://jmlr.org/papers/v10/chen09a.html>.
- [206] L. S. de Araujo et al. “Transcriptomic Biomarkers for Tuberculosis: Validation of NPC2 as a Single mRNA Biomarker to Diagnose TB, Predict Disease Progression, and Monitor Treatment Response”. In: *Cells* 10.10 (2021). ISSN: 2073-4409. DOI: 10.3390/cells10102704. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34685683>.
- [207] P. A. Harris et al. “Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support”. In: *J Biomed Inform* 42.2 (2009), pp. 377–81. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2008.08.010. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18929686>.
- [208] S. M. Graham et al. “Clinical Case Definitions for Classification of Intrathoracic Tuberculosis in Children: An Update”. In: *Clin Infect Dis* 61Suppl 3.Suppl 3 (2015), S179–87. ISSN: 1537-6591. DOI: 10.1093/cid/civ581. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26409281>.
- [209] U. Egere et al. “Identifying children with tuberculosis among household contacts in The Gambia”. In: *Int J Tuberc Lung Dis* 21.1 (2017), pp. 46–52. ISSN: 1815-7920. DOI: 10.5588/ijtld.16.0289. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28157464>.
- [210] D. Jaganath and E. Mupere. “Childhood tuberculosis and malnutrition”. In: *J Infect Dis* 206.12 (2012), pp. 1809–15. ISSN: 1537-6613. DOI: 10.1093/infdis/jis608. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23033147>.
- [211] A. A. Price et al. “Utilization of Current Diagnostic Indicators to Characterize Pediatric Undernutrition among US Children”. In: *Nutrients* 12.5 (2020). ISSN: 2072-6643. DOI: 10.3390/nu12051409. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32422888>.
- [212] WHO. *Weight-for-age: Charts*. URL: <https://www.who.int/tools/child-growth-standards/standards/weight-for-age>. (accessed: 07.12.2022).
- [213] “Malawi 2004: results from the Demographic and Health Survey”. In: *Stud Fam Plann* 38.1 (2007), pp. 61–66. ISSN: 0039-3665. DOI: 10.1111/j.1728-4465.2007.00117.x. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17385384>.
- [214] F. W. Kalembo et al. “Prevalence and factors associated with emotional and behavioural difficulties among children living with HIV in Malawi: a cross-sectional study”. In: *BMC Psychiatry* 19.1 (2019), p. 60. ISSN: 1471-244X. DOI: 10.1186/s12888-019-2046-2. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30736758>.
- [215] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: <https://www.R-project.org/>.
- [216] S. A. Bustin et al. “The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments”. In: *Clin Chem* 55.4 (2009), pp. 611–22. ISSN: 1530-8561. DOI: 10.1373/clinchem.2008.112797. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19246619>.
- [217] A. Turabelidze, S. Guo, and L. A. DiPietro. “Importance of housekeeping gene selection for accurate reverse transcription-quantitative polymerase chain reaction in a wound healing model”. In: *Wound Repair Regen* 18.5 (2010), pp. 460–6. ISSN: 1524-475X. DOI: 10.1111/j.1524-475X.2010.00611.x. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20731795>.
- [218] Y. Cao et al. “Digital PCR as an Emerging Tool for Monitoring of Microbial Biodegradation”. In: *Molecules* 25.3 (2020). ISSN: 1420-3049. DOI: 10.3390/molecules25030706. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32041334>.
- [219] A. Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts635. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23104886>.
- [220] D. Kim, B. Langmead, and S. L. Salzberg. “HISAT: a fast spliced aligner with low memory requirements”. In: *Nat Methods* 12.4 (2015), pp. 357–60. ISSN: 1548-7105. DOI: 10.1038/nmeth.3317. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25751142>.
- [221] Y. Liao, G. K. Smyth, and W. Shi. “The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote”. In: *Nucleic Acids Res* 41.10 (2013), e108. ISSN: 1362-4962. DOI: 10.1093/nar/gkt214. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23558742>.
- [222] S. Andrews. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. 2010. URL: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. (accessed: 07.12.2022).
- [223] P. Ewels et al. “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19 (2016), pp. 3047–8. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw354. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27312411>.
- [224] A. Dobin and T. R. Gingeras. “Mapping RNA-seq Reads with STAR”. In: *Curr Protoc Bioinformatics* 51 (2015), pp. 11.14.1–11.14.19. ISSN: 1934-340X. DOI: 10.1002/0471250953.bi1114s51. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26334920>.
- [225] Y. Liao, G. K. Smyth, and W. Shi. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. In: *Bioinformatics* 30.7 (2014), pp. 923–30.

- ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt656. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24227677>.
- [226] D. A. Vignali. “Multiplexed particle-based flow cytometric assays”. In: *J Immunol Methods* 243.1-2 (2000), pp. 243–55. ISSN: 0022-1759. DOI: 10.1016/s0022-1759(00)00238-6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/10986418>.
- [227] M. B. O’Rourke et al. “What is Normalization? The Strategies Employed in Top-Down and Bottom-Up Proteome Analysis Workflows”. In: *Proteomes* 7.3 (2019). ISSN: 2227-7382. DOI: 10.3390/proteomes7030029. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31443461>.
- [228] M. Ringnér. “What is principal component analysis?” In: *Nature Biotechnology* 26 (2008). DOI: 10.1038/nbt0308-303. URL: <https://www.nature.com/articles/nbt0308-303>.
- [229] J. T. Leek et al. “Tackling the widespread and critical impact of batch effects in high-throughput data”. In: *Nat Rev Genet* 11.10 (2010), pp. 733–9. ISSN: 1471-0064. DOI: 10.1038/nrg2825. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20838408>.
- [230] X. Liu et al. “Normalization Methods for the Analysis of Unbalanced Transcriptome Data: A Review”. In: *Front Bioeng Biotechnol* 7 (2019), p. 358. ISSN: 2296-4185. DOI: 10.3389/fbioe.2019.00358. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32039167>.
- [231] W. E. Johnson, C. Li, and A. Rabinovic. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8.1 (2007), pp. 118–27. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj037. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16632515>.
- [232] J. T. Leek et al. “The sva package for removing batch effects and other unwanted variation in high-throughput experiments”. In: *Bioinformatics* 28.6 (2012), pp. 882–3. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts034. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22257669>.
- [233] W. W. Goh and L. Wong. “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects — a case study in clinical proteomics”. In: *BMC Genomics* 18.Suppl 2 (2017), p. 142. ISSN: 1471-2164. DOI: 10.1186/s12864-017-3490-3. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28361693>.
- [234] Q. Liu et al. “Addressing the batch effect issue for LC/MS metabolomics data in data preprocessing”. In: *Sci Rep* 10.1 (2020), p. 13856. ISSN: 2045-2322. DOI: 10.1038/s41598-020-70850-0. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32807888>.
- [235] Y. Zhang, G. Parmigiani, and W. E. Johnson. “ComBat-seq: batch effect adjustment for RNA-seq count data”. In: *NAR Genom Bioinform* 2.3 (2020), lqaa078. ISSN: 2631-9268. DOI: 10.1093/nargab/lqaa078. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33015620>.
- [236] M. E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Res* 43.7 (2015), e47. ISSN: 1362-4962. DOI: 10.1093/nar/gkv007. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25605792>.
- [237] V. Nygaard, E. A. Rodland, and E. Hovig. “Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses”. In: *Biostatistics* 17.1 (2016), pp. 29–39. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxv027. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26272994>.
- [238] M. G. Hong et al. “Multidimensional Normalization to Minimize Plate Effects of Suspension Bead Array Data”. In: *J Proteome Res* 15.10 (2016), pp. 3473–3480. ISSN: 1535-3907. DOI: 10.1021/acs.jproteome.5b01131. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27570895>.
- [239] K. J. Livak and T. D. Schmittgen. “Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method”. In: *Methods* 25.4 (2001), pp. 402–8. ISSN: 1046-2023. DOI: 10.1006/meth.2001.1262. URL: <https://www.ncbi.nlm.nih.gov/pubmed/11846609>.
- [240] H. Dvinge and P. Bertone. “HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R”. In: *Bioinformatics* 25.24 (2009), pp. 3325–6. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp578. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19808880>.
- [241] F.M. Hassan et al. “XIST and RPS4Y1 long non-coding RNA transcriptome as sex biomarkers in different body fluids”. In: *Egypt J Forensic Sci* 9.16 (2019). DOI: 10.1186/s41935-019-0122-x. URL: <https://ejfs.springeropen.com/articles/10.1186/s41935-019-0122-x>.
- [242] C. W. Law et al. “RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR”. In: *F1000Res* 5 (2016). ISSN: 2046-1402. DOI: 10.12688/f1000research.9005.3. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27441086>.
- [243] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2010), pp. 139–40. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19910308>.
- [244] B. M. Bolstad et al. “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”. In: *Bioinformatics* 19.2 (2003), pp. 185–93. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/19.2.185. URL: <https://www.ncbi.nlm.nih.gov/pubmed/12538238>.

- [245] M. D. Robinson and A. Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data”. In: *Genome Biol* 11.3 (2010), R25. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-3-r25. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20196867>.
- [246] A. Penn-Nicholson et al. “Discovery and validation of a prognostic proteomic signature for tuberculosis progression: A prospective cohort study”. In: *PLoS Med* 16.4 (2019), e1002781. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002781. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30990820>.
- [247] M. I. Love, W. Huber, and S. Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biol* 15.12 (2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25516281>.
- [248] H. B. Mann and D. R. Whitney. “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other”. In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60. DOI: 10.1214/aoms/1177730491. URL: <https://doi.org/10.1214/aoms/1177730491>.
- [249] Y. Benjamini and Hochberg Y. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300. URL: <https://www.jstor.org/stable/2346101>.
- [250] T. J. Hardcastle and K. A. Kelly. “baySeq: empirical Bayesian methods for identifying differential expression in sequence count data”. In: *BMC Bioinformatics* 11 (2010), p. 422. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-422. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20698981>.
- [251] N. Leng et al. “EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments”. In: *Bioinformatics* 29.8 (2013), pp. 1035–43. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt087. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23428641>.
- [252] C. W. Law et al. “voom: Precision weights unlock linear model analysis tools for RNA-seq read counts”. In: *Genome Biol* 15.2 (2014), R29. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-2-r29. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24485249>.
- [253] R. Liu et al. “Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses”. In: *Nucleic Acids Res* 43.15 (2015), e97. ISSN: 1362-4962. DOI: 10.1093/nar/gkv412. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25925576>.
- [254] G. K. Smyth. “Linear models and empirical bayes methods for assessing differential expression in microarray experiments”. In: *Stat Appl Genet Mol Biol* 3 (2004), Article3. ISSN: 1544-6115. DOI: 10.2202/1544-6115.1027. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16646809>.
- [255] G. K. Smyth, J. Michaud, and H. S. Scott. “Use of within-array replicate spots for assessing differential expression in microarray experiments”. In: *Bioinformatics* 21.9 (2005), pp. 2067–75. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti270. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15657102>.
- [256] M. A. Garcia-Campos, J. Espinal-Enriquez, and E. Hernandez-Lemus. “Pathway Analysis: State of the Art”. In: *Front Physiol* 6 (2015), p. 383. ISSN: 1664-042X. DOI: 10.3389/fphys.2015.00383. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26733877>.
- [257] J. Reimand et al. “Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap”. In: *Nat Protoc* 14.2 (2019), pp. 482–517. ISSN: 1750-2799. DOI: 10.1038/s41596-018-0103-9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30664679>.
- [258] U. Raudvere et al. “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)”. In: *Nucleic Acids Res* 47.W1 (2019), W191–W198. ISSN: 1362-4962. DOI: 10.1093/nar/gkz369. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31066453>.
- [259] L. Kolberg et al. “gprofiler2 – an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler”. In: *F1000Res* 9 (2020). ISSN: 2046-1402. DOI: 10.12688/f1000research.24956.2. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33564394>.
- [260] G. Yu et al. “clusterProfiler: an R package for comparing biological themes among gene clusters”. In: *OMICS* 16.5 (2012), pp. 284–7. ISSN: 1557-8100. DOI: 10.1089/omi.2011.0118. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22455463>.
- [261] T. Wu et al. “clusterProfiler 4.0: A universal enrichment tool for interpreting omics data”. In: *Innovation (Camb)* 2.3 (2021), p. 100141. ISSN: 2666-6758. DOI: 10.1016/j.xinn.2021.100141. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34557778>.
- [262] J. A. Herberg et al. “Diagnostic Test Accuracy of a 2-Transcript Host RNA Signature for Discriminating Bacterial vs Viral Infection in Febrile Children”. In: *JAMA* 316.8 (2016), pp. 835–45. ISSN: 1538-3598. DOI: 10.1001/jama.2016.11236. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27552617>.
- [263] H. D. Gliddon et al. “Identification of Reduced Host Transcriptomic Signatures for Tuberculosis Disease and Digital PCR-Based Validation and Quantification”. In: *Front Immunol* 12 (2021), p. 637164. ISSN: 1664-3224. DOI: 10.3389/fimmu.2021.637164. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33763081>.
- [264] W. J. Youden. “Index for rating diagnostic tests”. In: *Cancer* 3.1 (1950), pp. 32–5. ISSN: 0008-543X. DOI: 10.1002/1097-0142(1950)3:1<32::aid-cnrcr2820030106>3.0.co;2-3. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15405679>.

- [265] A. Singh et al. "DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays". In: *Bioinformatics* 35.17 (2019), pp. 3055–3062. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bty1054. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30657866>.
- [266] F. Rohart et al. "MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms". In: *BMC Bioinformatics* 18.1 (2017), p. 128. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1553-8. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28241739>.
- [267] A. Tenenhaus et al. "Variable selection for generalized canonical correlation analysis". In: *Biostatistics* 15.3 (2014), pp. 569–83. ISSN: 1468-4357 (Electronic) 1465-4644 (Linking). DOI: 10.1093/biostatistics/kxu001. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24550197>.
- [268] A. L. Garcia-Basteiro et al. "Adolescents and young adults: a neglected population group for tuberculosis surveillance". In: *Eur Respir J* 51.2 (2018). ISSN: 1399-3003. DOI: 10.1183/13993003.00176-2018. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29467211>.
- [269] S. Lamble et al. "Improved workflows for high throughput library preparation using the transposome-based Nextera system". In: *BMC Biotechnol* 13 (2013), p. 104. ISSN: 1472-6750. DOI: 10.1186/1472-6750-13-104. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24256843>.
- [270] Y. Yang et al. "Recent advances in the mechanisms of NLRP3 inflammasome activation and its inhibitors". In: *Cell Death Dis* 10.2 (2019), p. 128. ISSN: 2041-4889. DOI: 10.1038/s41419-019-1413-8. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30755589>.
- [271] V. Hornung et al. "AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC". In: *Nature* 458.7237 (2009), pp. 514–8. ISSN: 1476-4687. DOI: 10.1038/nature07725. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19158675>.
- [272] A. R. Shenoy et al. "GBP5 promotes NLRP3 inflammasome assembly and immunity in mammals". In: *Science* 336.6080 (2012), pp. 481–5. ISSN: 1095-9203. DOI: 10.1126/science.1217141. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22461501>.
- [273] E. Meunier et al. "Guanylate-binding proteins promote activation of the AIM2 inflammasome during infection with *Francisella novicida*". In: *Nat Immunol* 16.5 (2015), pp. 476–484. ISSN: 1529-2916. DOI: 10.1038/ni.3119. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25774716>.
- [274] M. Kaforou et al. "Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study". In: *PLoS Med* 10.10 (2013), e1001538. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001538. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24167453>.
- [275] J. Maertzdorf et al. "Common patterns and disease-related signatures in tuberculosis and sarcoidosis". In: *Proc Natl Acad Sci U S A* 109.20 (2012), pp. 7853–8. ISSN: 1091-6490. DOI: 10.1073/pnas.1121072109. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22547807>.
- [276] A. Besson, S. F. Dowdy, and J. M. Roberts. "CDK inhibitors: cell cycle regulators and beyond". In: *Dev Cell* 14.2 (2008), pp. 159–69. ISSN: 1534-5807. DOI: 10.1016/j.devcel.2008.01.013. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18267085>.
- [277] J. A. Mahoney et al. "Cloning and characterization of CPVL, a novel serine carboxypeptidase, from human macrophages". In: *Genomics* 72.3 (2001), pp. 243–51. ISSN: 0888-7543. DOI: 10.1006/geno.2000.6484. URL: <https://www.ncbi.nlm.nih.gov/pubmed/11401439>.
- [278] J. Harris et al. "A vitellogenic-like carboxypeptidase expressed by human macrophages is localized in endoplasmic reticulum and membrane ruffles". In: *Int J Exp Pathol* 87.1 (2006), pp. 29–39. ISSN: 0959-9673. DOI: 10.1111/j.0959-9673.2006.00450.x. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16436111>.
- [279] L. A. Stanton et al. "Immunophenotyping of macrophages in human pulmonary tuberculosis and sarcoidosis". In: *Int J Exp Pathol* 84.6 (2003), pp. 289–304. ISSN: 0959-9673. DOI: 10.1111/j.0959-9673.2003.00365.x. URL: <https://www.ncbi.nlm.nih.gov/pubmed/14748748>.
- [280] C. Vilcheze et al. "*Mycobacterium tuberculosis* is extraordinarily sensitive to killing by a vitamin C-induced Fenton reaction". In: *Nat Commun* 4 (2013), p. 1881. ISSN: 2041-1723. DOI: 10.1038/ncomms2898. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23695675>.
- [281] C. Vilcheze, J. Kim, and Jr. Jacobs W. R. "Vitamin C Potentiates the Killing of *Mycobacterium tuberculosis* by the First-Line Tuberculosis Drugs Isoniazid and Rifampin in Mice". In: *Antimicrob Agents Chemother* 62.3 (2018). ISSN: 1098-6596. DOI: 10.1128/AAC.02165-17. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29298757>.
- [282] N. Andreu et al. "Primary macrophages and J774 cells respond differently to infection with *Mycobacterium tuberculosis*". In: *Sci Rep* 7 (2017), p. 42225. ISSN: 2045-2322. DOI: 10.1038/srep42225. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28176867>.
- [283] L. S. Meena. "Interrelation of Ca²⁺ and PE_PGRS proteins during *Mycobacterium tuberculosis* pathogenesis". In: *J Biosci* 44.1 (2019). ISSN: 0973-7138. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30837375>.
- [284] O. Tabone et al. "Blood transcriptomics reveal the evolution and resolution of the immune response in tuberculosis". In: *J Exp Med* 218.10 (2021). ISSN: 1540-9538. DOI: 10.1084/jem.20210915. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34491266>.

- [285] N. P. Long et al. "A 10-gene biosignature of tuberculosis treatment monitoring and treatment outcome prediction". In: *Tuberculosis (Edinb)* 131 (2021), p. 102138. ISSN: 1873-281X. DOI: 10.1016/j.tube.2021.102138. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34801869>.
- [286] N. Rockwood et al. "Assessment of treatment response in tuberculosis". In: *Expert Rev Respir Med* 10.6 (2016), pp. 643–54. ISSN: 1747-6356. DOI: 10.1586/17476348.2016.1166960. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27030924>.
- [287] The Sentinel Project for Pediatric Drug-Resistant Tuberculosis. *Management of Multidrug-Resistant Tuberculosis in Children: A Field Guide*. Report. 2019.
- [288] World Health Organization. *WHO consolidated guidelines on tuberculosis: Module 5: Management of tuberculosis in children and adolescents*. Report. 2022.
- [289] World Health Organization. *Companion Handbook to the WHO Guidelines for the Programmatic Management of Drug-Resistant Tuberculosis*. Report. 2014.
- [290] T. Domaszewska et al. "Concordant and discordant gene expression patterns in mouse strains identify best-fit animal model for human tuberculosis". In: *Sci Rep* 7.1 (2017), p. 12094. ISSN: 2045-2322. DOI: 10.1038/s41598-017-11812-x. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28935874>.
- [291] M. Kumar et al. "Activating transcription factor 3 modulates the macrophage immune response to *Mycobacterium tuberculosis* infection via reciprocal regulation of inflammatory genes and lipid body formation". In: *Cell Microbiol* 22.3 (2020), e13142. ISSN: 1462-5822. DOI: 10.1111/cmi.13142. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31709711>.
- [292] E. I. Boyle et al. "GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes". In: *Bioinformatics* 20.18 (2004), pp. 3710–5. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bth456. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15297299>.
- [293] E. Guirado, L. S. Schlesinger, and G. Kaplan. "Macrophages in tuberculosis: friend or foe". In: *Semin Immunopathol* 35.5 (2013), pp. 563–83. ISSN: 1863-2300. DOI: 10.1007/s00281-013-0388-2. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23864058>.
- [294] T. Villaseñor et al. "Activation of the Wnt Pathway by *Mycobacterium tuberculosis*: A Wnt-Wnt Situation". In: *Front Immunol* 8 (2017), p. 50. ISSN: 1664-3224. DOI: 10.3389/fimmu.2017.00050. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28203237>.
- [295] X. Wu et al. "Wnt/beta-catenin signaling reduces *Bacillus Calmette-Guerin*-induced macrophage necrosis through a ROS-mediated PARP/AIF-dependent pathway". In: *BMC Immunol* 16 (2015), p. 16. ISSN: 1471-2172. DOI: 10.1186/s12865-015-0080-5. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25887795>.
- [296] S. R. Weinberger et al. "Surface-enhanced laser desorption-ionization retentate chromatography mass spectrometry (SELDI-RC-MS): a new method for rapid development of process chromatography conditions". In: *J Chromatogr B Analyt Technol Biomed Life Sci* 782.1-2 (2002), pp. 307–16. ISSN: 1570-0232. DOI: 10.1016/S1570-0232(02)00564-0. URL: <https://www.ncbi.nlm.nih.gov/pubmed/12458014>.
- [297] L. Gold et al. "Aptamer-based multiplexed proteomic technology for biomarker discovery". In: *PLoS One* 5.12 (2010), e15004. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0015004. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21165148>.
- [298] E. Cassol et al. "Macrophage polarization and HIV-1 infection". In: *J Leukoc Biol* 87.4 (2010), pp. 599–608. ISSN: 1938-3673. DOI: 10.1189/jlb.1009673. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20042468>.
- [299] Z. Gu, R. Eils, and M. Schlesner. "Complex heatmaps reveal patterns and correlations in multidimensional genomic data". In: *Bioinformatics* 32.18 (2016), pp. 2847–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw313. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27207943>.
- [300] M. Stastna and J. E. Van Eyk. "Secreted proteins as a fundamental source for biomarker discovery". In: *Proteomics* 12.4-5 (2012), pp. 722–35. ISSN: 1615-9861. DOI: 10.1002/pmic.201100346. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22247067>.
- [301] M. D. Miller and M. S. Krangel. "The human cytokine I-309 is a monocyte chemoattractant". In: *Proc Natl Acad Sci U S A* 89.7 (1992), pp. 2950–4. ISSN: 0027-8424. DOI: 10.1073/pnas.89.7.2950. URL: <https://www.ncbi.nlm.nih.gov/pubmed/1557400>.
- [302] C. Trebst et al. "CC chemokine receptor 8 in the central nervous system is associated with phagocytic macrophages". In: *Am J Pathol* 162.2 (2003), pp. 427–38. ISSN: 0002-9440. DOI: 10.1016/S0002-9440(10)63837-0. URL: <https://www.ncbi.nlm.nih.gov/pubmed/12547701>.
- [303] M. Wei et al. "Regulation network of serum cytokines induced by tuberculosis-specific antigens reveals biomarkers for tuberculosis diagnosis". In: *Genet Mol Res* 14.4 (2015), pp. 17182–92. ISSN: 1676-5680. DOI: 10.4238/2015.December.16.18. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26681212>.
- [304] P. M. Manngo et al. "Prospective evaluation of host biomarkers other than interferon gamma in QuantiFERON Plus supernatants as candidates for the diagnosis of tuberculosis in symptomatic individuals". In: *J Infect* 79.3 (2019), pp. 228–235. ISSN: 1532-2742. DOI: 10.1016/j.jinf.2019.07.007. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31319143>.

- [305] E. M. Delemarre et al. "Serum Biomarker Profile Including CCL1, CXCL10, VEGF, and Adenosine Deaminase Activity Distinguishes Active From Remotely Acquired Latent Tuberculosis". In: *Front Immunol* 12 (2021), p. 725447. ISSN: 1664-3224. DOI: 10.3389/fimmu.2021.725447. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34691031>.
- [306] U. K. Rohlwick et al. "Matrix Metalloproteinases in Pulmonary and Central Nervous System Tuberculosis-A Review". In: *Int J Mol Sci* 20.6 (2019). ISSN: 1422-0067. DOI: 10.3390/ijms20061350. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30889803>.
- [307] S. Brilha et al. "Early Secretory Antigenic Target-6 Drives Matrix Metalloproteinase-10 Gene Expression and Secretion in Tuberculosis". In: *Am J Respir Cell Mol Biol* 56.2 (2017), pp. 223–232. ISSN: 1535-4989. DOI: 10.1165/rcmb.2016-01620C. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27654284>.
- [308] S. Krishnan et al. "Integrative Multi-Omics Reveals Serum Markers of Tuberculosis in Advanced HIV". In: *Front Immunol* 12 (2021), p. 676980. ISSN: 1664-3224. DOI: 10.3389/fimmu.2021.676980. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34168648>.
- [309] K. G. Smith and M. R. Clatworthy. "FcγRIIb in autoimmunity and infection: evolutionary and therapeutic implications". In: *Nat Rev Immunol* 10.5 (2010), pp. 328–43. ISSN: 1474-1741. DOI: 10.1038/nri2762. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20414206>.
- [310] N. Ahamad, S. Gupta, and D. Parashar. "Using Omics to Study Leprosy, Tuberculosis, and Other Mycobacterial Diseases". In: *Front Cell Infect Microbiol* 12 (2022), p. 792617. ISSN: 2235-2988 (Electronic) 2235-2988 (Linking). DOI: 10.3389/fcimb.2022.792617. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35281437>.
- [311] B. H. Kim et al. "IFN-inducible GTPases in host cell defense". In: *Cell Host Microbe* 12.4 (2012), pp. 432–44. ISSN: 1934-6069. DOI: 10.1016/j.chom.2012.09.007. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23084913>.
- [312] H. Mulenga et al. "Performance of diagnostic and predictive host blood transcriptomic signatures for Tuberculosis disease: A systematic review and meta-analysis". In: *PLoS One* 15.8 (2020), e0237574. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0237574. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32822359>.
- [313] N. N. Chegou et al. "Beyond the IFN-gamma horizon: biomarkers for immunodiagnosis of infection with *Mycobacterium tuberculosis*". In: *Eur Respir J* 43.5 (2014), pp. 1472–86. ISSN: 1399-3003. DOI: 10.1183/09031936.00151413. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24311770>.
- [314] M. Ruhwald et al. "CXCL10/IP-10 release is induced by incubation of whole blood from tuberculosis patients with ESAT-6, CFP10 and TB7.7". In: *Microbes Infect* 9.7 (2007), pp. 806–12. ISSN: 1286-4579. DOI: 10.1016/j.micinf.2007.02.021. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17533146>.
- [315] D. Chen et al. "Interleukin-7 Biology and Its Effects on Immune Cells: Mediator of Generation, Differentiation, Survival, and Homeostasis". In: *Front Immunol* 12 (2021), p. 747324. ISSN: 1664-3224. DOI: 10.3389/fimmu.2021.747324. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34925323>.
- [316] C. Lundtoft et al. "Aberrant plasma IL-7 and soluble IL-7 receptor levels indicate impaired T-cell response to IL-7 in human tuberculosis". In: *PLoS Pathog* 13.6 (2017), e1006425. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1006425. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28582466>.
- [317] M. A. Poritz et al. "FilmArray, an automated nested multiplex PCR system for multi-pathogen detection: development and application to respiratory tract infection". In: *PLoS One* 6.10 (2011), e26047. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0026047. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22039434>.
- [318] M. Kodani and J. M. Winchell. "Engineered combined-positive-control template for real-time reverse transcription-PCR in multiple-pathogen-detection assays". In: *J Clin Microbiol* 50.3 (2012), pp. 1057–60. ISSN: 1098-660X. DOI: 10.1128/JCM.05987-11. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22170926>.

APPENDIX A

CASE DEFINITIONS FOR EXTRAPULMONARY TUBERCULOSIS IN CHILDREN

microbiologically confirmed TB: Bacteriological confirmation requires *Mycobacterium tuberculosis* (*Mtb*) to be confirmed (culture, Xpert MTB/RIF assay or Xpert MTB/RIF Ultra assay) from at least one bodily fluid or tissue specimen, obtained from the site of Extra-Pulmonary Tuberculosis (EPTB); for example cerebrospinal fluid (tuberculous meningitis), lymph node tissue or fine needle aspirate (tuberculous lymphadenitis), abdominal organ tissue or ascites fluid (abdominal TB), other organ tissue (eye, bone, spine or joint TB), joint fluid (joint TB), or pleural fluid or tissue (tuberculous pleural effusion)

Unconfirmed TB: Bacteriological confirmation NOT obtained AND at least 2 of the following criteria, including at least one criterion at the site of Extra-Pulmonary Tuberculosis (EPTB):

- Symptoms suggestive of TB
- Close TB exposure or immunologic evidence of Latent Tuberculosis Infection (LTBI)
- Positive response to TB treatment (requires documented positive clinical response on TB treatment – no time duration specified)
- Criteria at the site of EPTB (refer below)

Unconfirmed TB should be subdivided into cases (a) with immunologic evidence of *Mtb* infection i.e. Tuberculin Skin Test (TST) and/or Interferon-Gamma Release Assay (IGRA) positive and (b) without immunologic evidence of *Mtb* infection.

Criteria at the site of EPTB

1. Tuberculous pleural effusion:
 - (a) Radiological features: a pleural effusion together with chest radiographic features consistent with PTB
 - (b) Consistent features on pleural fluid analysis: straw colour, protein content >30 g/L, pH <7.3, lactate dehydrogenase (LDH) >200 U/L, lymphocytosis is typical, but a minority of patients have a predominantly polymorphonuclear cellular infiltrate
2. Central nervous system TB, including tuberculous meningitis and central nervous system tuberculoma
 - (a) Typical neuroimaging features: hydrocephalus, basal meningeal enhancement, infarcts and/or tuberculomas
 - (b) Consistent features on cerebrospinal fluid analysis: raised opening pressure, a raised protein concentration, hypoglycorrhachia or cerebrospinal fluid to plasma glucose ratio of less

than 50% and leucocyte counts vary from 10 to 2000 leucocytes/ μ L with lymphocyte predominance (>50%) in 65-75% of patients

3. Tuberculous lymphadenitis

- (a) Consistent signs of tuberculous lymphadenitis: painless, non-tender lymph node enlargement, frequently matted, with or without sinus formation
- (b) Consistent features on lymph node histology: tuberculous granulomata, acid-fast bacilli and caseating necrosis

4. Abdominal TB

- (a) Consistent signs include hepatosplenomegaly, detectable ascites, lymph node mass located in the right lower quadrant and a doughy abdomen
- (b) Consistent abdominal ultrasound findings of enlarged mesenteric and para-aortic lymph nodes, ascites with/without loculation with/without debris, organomegaly (liver and/or spleen), and focal lesions in the liver and/or spleen
- (c) Consistent features of ascites fluid analysis: elevated protein concentration (>25 g/L), lymphocytosis is typical but in a minority of patients neutrophils may be dominant
- (d) Consistent features on organ histology: tuberculous granulomata, acid-fast bacilli, caseating necrosis

5. Osteoarticular TB (includes bone, spinal, joint TB)

- (a) Kyphosis indicative of spinal TB
- (b) Typical radiological features: bone destruction, cold abscess formation, hypertrophy of articular membranes, loss of intravertebral spaces and gibbus formation in spinal TB
- (c) Consistent features on histology: tuberculous granulomata, acid-fast bacilli, caseating necrosis
- (d) Lymphocytosis on synovial fluid

6. Ocular TB

- (a) Consistent features on histology of ocular tissue: tuberculous granulomata, acid-fast bacilli, caseating necrosis

Declaration of originality for Appendix A:

The case definitions for extrapulmonary tuberculosis have been drafted by Prof. Brian Eley (Red Cross War Memorial Children's Hospital, Cape Town, South Africa).

APPENDIX B

SDE GENES IN THE DISCOVERY RNA-SEQ DATA WHEN COMPARING MICROBIOLOGICALLY CONFIRMED TB VERSUS UNLIKELY TB

Table B.1: SDE genes in the discovery RNA-Seq data when comparing microbiologically confirmed TB vs. unlikely TB. *SDE* = Significantly Differentially Expressed; *FC* = Fold Change; *BH* = Benjamini-Hochberg

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
GBP5	ENSG00000154451	1.6	1.4x10 ⁻⁵⁶
POLB	ENSG00000070501	0.7	2.17x10 ⁻⁴⁹
GBP6	ENSG00000183347	1.9	1.32x10 ⁻⁴⁶
P2RY14	ENSG00000174944	1.68	1.92x10 ⁻⁴⁴
CALHM6	ENSG00000188820	1.29	3.79x10 ⁻⁴³
ANKRD22	ENSG00000152766	1.91	8.56x10 ⁻⁴²
GBP2	ENSG00000162645	0.77	9.57x10 ⁻⁴¹
CYB561	ENSG00000008283	0.8	9.57x10 ⁻⁴¹
CD274	ENSG00000120217	1.42	1.67x10 ⁻³⁹
GBP1	ENSG00000117228	1.3	1.24x10 ⁻³⁸
LINC02555	ENSG00000260943	1.51	7.7x10 ⁻³⁷
BATF2	ENSG00000168062	1.66	3.67x10 ⁻³⁶
WDFY1	ENSG00000085449	0.57	1.19x10 ⁻³⁵
ETV7	ENSG0000010030	1.74	7.98x10 ⁻³⁵
LHFPL2	ENSG00000145685	0.94	2.24x10 ⁻³⁴
PDCD1LG2	ENSG00000197646	1.44	2.24x10 ⁻³⁴
SCARF1	ENSG00000074660	1.19	6.9x10 ⁻³⁴
FCGR1CP	ENSG00000265531	1.48	2.44x10 ⁻³³
FCGR1A	ENSG00000150337	1.38	3.43x10 ⁻³³
IRF1	ENSG00000125347	0.57	1.11x10 ⁻³²
FGL2	ENSG00000127951	0.68	1.68x10 ⁻³²
NPC2	ENSG00000119655	0.63	2.54x10 ⁻³²
PSMB9	ENSG00000240065	0.57	4.2x10 ⁻³²
ALDH1A1	ENSG00000165092	0.97	5.46x10 ⁻³²
VAMP5	ENSG00000168899	0.77	1.19x10 ⁻³¹
FAS	ENSG00000026103	0.63	1.31x10 ⁻³¹
SEPTIN4	ENSG00000108387	1.7	3.86x10 ⁻³⁰
SERPING1	ENSG00000149131	1.62	4.56x10 ⁻³⁰
GBP4	ENSG00000162654	0.86	6.84x10 ⁻³⁰
APOL1	ENSG00000100342	0.89	1.01x10 ⁻²⁹
APOL2	ENSG00000128335	0.64	4.5x10 ⁻²⁹
LINC01232	ENSG00000280734	0.5	7.45x10 ⁻²⁹
WARS1	ENSG00000140105	0.89	7.97x10 ⁻²⁹
PSME2	ENSG00000100911	0.53	1.27x10 ⁻²⁸
FCGR1B	ENSG00000198019	1.13	1.37x10 ⁻²⁷
APOL6	ENSG00000221963	0.7	8.2x10 ⁻²⁷
LAP3	ENSG00000002549	0.97	3.18x10 ⁻²⁶
DHRS9	ENSG00000073737	0.85	5.65x10 ⁻²⁶
LRRK2-DT	ENSG00000225342	0.67	2.38x10 ⁻²⁵
TNFAIP6	ENSG00000123610	1.15	5.07x10 ⁻²⁵
PSTPIP2	ENSG00000152229	0.78	7.33x10 ⁻²⁵
CASP1	ENSG00000137752	0.56	2.08x10 ⁻²⁴
TIFA	ENSG00000145365	0.69	5.95x10 ⁻²⁴
SMCO4	ENSG00000166002	0.56	7.48x10 ⁻²⁴
PRRG4	ENSG00000135378	0.82	3.51x10 ⁻²³
DUSP3	ENSG00000108861	0.58	9.23x10 ⁻²³

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
GBP3	ENSG00000117226	0.58	9.48x10 ⁻²³
STAT1	ENSG00000115415	0.65	3.68x10 ⁻²²
PLAAT4	ENSG00000133321	0.54	5.17x10 ⁻²²
CARD17	ENSG00000255221	1.33	9.48x10 ⁻²²
UBE2L6	ENSG00000156587	0.69	1.85x10 ⁻²¹
SECTM1	ENSG00000141574	0.79	5.44x10 ⁻²¹
GRAMD1B	ENSG00000023171	0.66	6.2x10 ⁻²¹
IGSF6	ENSG00000140749	0.63	7.07x10 ⁻²¹
MSRB2	ENSG00000148450	0.57	3.08x10 ⁻²⁰
TAP1	ENSG00000168394	0.58	3.24x10 ⁻²⁰
AIM2	ENSG00000163568	0.79	5.03x10 ⁻²⁰
ICAM1	ENSG00000090339	0.63	6.81x10 ⁻²⁰
ODF3B	ENSG00000177989	0.69	1.2x10 ⁻¹⁹
GCH1	ENSG00000131979	0.51	1.74x10 ⁻¹⁹
KCNJ2	ENSG00000123700	0.77	1.74x10 ⁻¹⁹
PTGES3P1	ENSG00000234518	0.52	1.96x10 ⁻¹⁹
ENSG00000273272	ENSG00000273272	0.79	2.79x10 ⁻¹⁹
RMI2	ENSG00000175643	0.63	3.58x10 ⁻¹⁹
MYOF	ENSG00000138119	0.79	3.76x10 ⁻¹⁹
GK	ENSG00000198814	0.75	7.52x10 ⁻¹⁹
ENSG00000261438	ENSG00000261438	0.51	1.43x10 ⁻¹⁸
F2RL1	ENSG00000164251	0.72	2.59x10 ⁻¹⁸
TIMM10	ENSG00000134809	0.61	3.68x10 ⁻¹⁸
CAPNS2	ENSG00000256812	0.62	6x10 ⁻¹⁸
HCAR3	ENSG00000255398	0.79	1.07x10 ⁻¹⁷
FRMD3	ENSG00000172159	0.6	1.19x10 ⁻¹⁷
CARD16	ENSG00000204397	0.53	3.14x10 ⁻¹⁷
C1QA	ENSG00000173372	1.14	5.23x10 ⁻¹⁷
TNFSF10	ENSG00000121858	0.6	5.98x10 ⁻¹⁷
HCAR2	ENSG00000182782	0.77	1.11x10 ⁻¹⁶
C1QB	ENSG00000173369	1.3	1.31x10 ⁻¹⁶
DHRS12	ENSG00000102796	0.51	1.4x10 ⁻¹⁶
ENSG00000261997	ENSG00000261997	0.6	1.77x10 ⁻¹⁶
LRRK2	ENSG00000188906	0.55	2.68x10 ⁻¹⁶
SMARCD3	ENSG00000082014	0.67	2.87x10 ⁻¹⁶
EFCAB2	ENSG00000203666	0.57	3.16x10 ⁻¹⁶
KCNJ2-AS1	ENSG00000267365	0.73	5.17x10 ⁻¹⁶
CASP5	ENSG00000137757	0.9	6.17x10 ⁻¹⁶
PLAUR	ENSG00000011422	0.56	1.74x10 ⁻¹⁵
ENSG00000255801	ENSG00000255801	0.95	4.14x10 ⁻¹⁵
FBXO6	ENSG00000116663	0.61	4.26x10 ⁻¹⁵
CPVL	ENSG00000106066	0.53	6.35x10 ⁻¹⁵
RTP4	ENSG00000136514	0.69	7.88x10 ⁻¹⁵
IL13RA1	ENSG00000131724	0.51	1.17x10 ⁻¹⁴
NOD2	ENSG00000167207	0.53	1.34x10 ⁻¹⁴
NCF1	ENSG00000158517	0.51	1.77x10 ⁻¹⁴
CACNA1E	ENSG00000198216	0.94	4.47x10 ⁻¹⁴
OR52K3P	ENSG00000225101	0.54	4.47x10 ⁻¹⁴
FLVCR2	ENSG00000119686	0.65	6.42x10 ⁻¹⁴
SOD2	ENSG00000112096	0.57	9.47x10 ⁻¹⁴
EPSTI1	ENSG00000133106	0.66	9.48x10 ⁻¹⁴
ENSG00000270120	ENSG00000270120	0.54	9.67x10 ⁻¹⁴
RAB20	ENSG00000139832	0.67	1.31x10 ⁻¹³
ENSG00000241886	ENSG00000241886	0.69	1.5x10 ⁻¹³
KCNJ15	ENSG00000157551	0.71	2.5x10 ⁻¹³
SORT1	ENSG00000134243	0.56	6.49x10 ⁻¹³
IDO1	ENSG00000131203	1.12	2.05x10 ⁻¹²
PLD4	ENSG00000166428	-0.52	5.35x10 ⁻¹²
ENSG00000233461	ENSG00000233461	0.63	5.38x10 ⁻¹²
DOCK4	ENSG00000128512	0.58	6.26x10 ⁻¹²
SAMD9L	ENSG00000177409	0.52	6.6x10 ⁻¹²
IL1B	ENSG00000125538	0.58	1.03x10 ⁻¹¹

Continues on next page

APPENDIX B. LIST OF SDE GENES

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
DDX10P1	ENSG00000237135	0.58	1.29x10 ⁻¹¹
CDKN1C	ENSG00000129757	-0.79	1.38x10 ⁻¹¹
FCER1G	ENSG00000158869	0.59	1.91x10 ⁻¹¹
IFIT3	ENSG00000119917	0.95	2.2x10 ⁻¹¹
LRRC4	ENSG00000128594	0.58	4.9x10 ⁻¹¹
MSR1	ENSG00000038945	0.73	6.83x10 ⁻¹¹
SOX5	ENSG00000134532	-0.62	1x10 ⁻¹⁰
FFAR2	ENSG00000126262	0.56	1.23x10 ⁻¹⁰
PROS1	ENSG00000184500	0.83	1.56x10 ⁻¹⁰
CLEC4E	ENSG00000166523	0.57	4.51x10 ⁻¹⁰
CEACAM1	ENSG00000079385	0.57	5.48x10 ⁻¹⁰
IFITM3	ENSG00000142089	0.63	6.15x10 ⁻¹⁰
GYPC	ENSG00000136732	0.62	6.59x10 ⁻¹⁰
PF4	ENSG00000163737	0.62	9.47x10 ⁻¹⁰
ADM	ENSG00000148926	0.61	1.49x10 ⁻⁰⁹
IFIT2	ENSG00000119922	0.7	1.71x10 ⁻⁰⁹
FPR2	ENSG00000171049	0.54	4.19x10 ⁻⁰⁹
MIR3945HG	ENSG00000251230	0.63	6.12x10 ⁻⁰⁹
EGF	ENSG00000138798	0.59	1.52x10 ⁻⁰⁸
GNG11	ENSG00000127920	0.5	1.52x10 ⁻⁰⁸
TREML1	ENSG00000161911	0.65	1.8x10 ⁻⁰⁸
ALAS2	ENSG00000158578	0.88	2.42x10 ⁻⁰⁸
PLEK2	ENSG00000100558	0.73	2.7x10 ⁻⁰⁸
SLAH2	ENSG00000181788	0.6	2.8x10 ⁻⁰⁸
TGM2	ENSG00000198959	0.63	5.53x10 ⁻⁰⁸
HBA2	ENSG00000188536	0.85	7.11x10 ⁻⁰⁸
PPBP	ENSG00000163736	0.61	7.25x10 ⁻⁰⁸
FPR1	ENSG00000171051	0.5	8.65x10 ⁻⁰⁸
SH3BGRL2	ENSG00000198478	0.56	9.91x10 ⁻⁰⁸
SELP	ENSG00000174175	0.52	1.05x10 ⁻⁰⁷
PF4V1	ENSG00000109272	0.6	1.19x10 ⁻⁰⁷
FCGR3B	ENSG00000162747	0.64	1.63x10 ⁻⁰⁷
GP9	ENSG00000169704	0.56	2.29x10 ⁻⁰⁷
UBXN6	ENSG00000167671	0.52	2.39x10 ⁻⁰⁷
TCN2	ENSG00000185339	0.55	2.62x10 ⁻⁰⁷
ITGA2B	ENSG00000005961	0.7	3.05x10 ⁻⁰⁷
ENSG00000280287	ENSG00000280287	-0.52	3.5x10 ⁻⁰⁷
MARCO	ENSG00000019169	0.57	6.36x10 ⁻⁰⁷
CMTM5	ENSG00000166091	0.56	6.51x10 ⁻⁰⁷
KREMEN1	ENSG00000183762	0.63	6.63x10 ⁻⁰⁷
ITGB5	ENSG00000082781	0.51	9.5x10 ⁻⁰⁷
KLC3	ENSG00000104892	0.77	1.47x10 ⁻⁰⁶
SPARC	ENSG00000113140	0.51	1.6x10 ⁻⁰⁶
FAM210B	ENSG00000124098	0.58	3.2x10 ⁻⁰⁶
GLDC	ENSG00000178445	-0.54	3.36x10 ⁻⁰⁶
SLC4A1	ENSG000000004939	0.65	7.7x10 ⁻⁰⁶
MMRN1	ENSG00000138722	0.51	8.64x10 ⁻⁰⁶
SELENBP1	ENSG00000143416	0.69	1.03x10 ⁻⁰⁵
IGF2BP2	ENSG00000073792	0.52	1.3x10 ⁻⁰⁵
SPTB	ENSG00000070182	0.62	1.37x10 ⁻⁰⁵
KRT1	ENSG00000167768	0.7	1.73x10 ⁻⁰⁵
ITGB3	ENSG00000259207	0.55	2.01x10 ⁻⁰⁵
TMOD1	ENSG00000136842	0.51	2.02x10 ⁻⁰⁵
EPB42	ENSG00000166947	0.66	2.07x10 ⁻⁰⁵
SLC26A8	ENSG00000112053	0.51	2.15x10 ⁻⁰⁵
TRIM10	ENSG00000204613	0.56	3.21x10 ⁻⁰⁵
SLC25A39	ENSG00000013306	0.51	3.45x10 ⁻⁰⁵
FAM20A	ENSG00000108950	0.64	3.6x10 ⁻⁰⁵
SLC6A8	ENSG00000130821	0.52	3.76x10 ⁻⁰⁵
IGHV3-30	ENSG00000270550	-0.56	4.44x10 ⁻⁰⁵
RSAD2	ENSG00000134321	0.69	6.34x10 ⁻⁰⁵
IGHG3	ENSG00000211897	-0.53	8.26x10 ⁻⁰⁵

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
GMPR	ENSG00000137198	0.52	1.05x10 ⁻⁰⁴
TRIM58	ENSG00000162722	0.53	1.07x10 ⁻⁰⁴
KLF1	ENSG00000105610	0.5	2.81x10 ⁻⁰⁴
AHSP	ENSG00000169877	0.6	6.37x10 ⁻⁰⁴
OSBP2	ENSG00000184792	0.52	6.54x10 ⁻⁰⁴
RAP1GAP	ENSG00000076864	0.82	6.86x10 ⁻⁰⁴
OR2W3	ENSG00000238243	0.53	1.15x10 ⁻⁰³
DEFA3	ENSG00000239839	0.69	3.1x10 ⁻⁰³

APPENDIX C

SDE GENES IN THE RNA-SEQ DATA WHEN COMPARING MICROBIOLOGICALLY CONFIRMED TB VERSUS UNLIKELY TB AT BASELINE

Table C.1: SDE genes in the RNA-Seq data when comparing microbiologically confirmed TB vs. unlikely TB at baseline. *SDE* = Significantly Differentially Expressed; *FC* = Fold Change; *BH* = Benjamini-Hochberg

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
GBP5	ENSG00000154451	1.9	9.54x10 ⁻¹⁶
CYB561	ENSG00000008283	1.03	2.39x10 ⁻¹⁵
P2RY14	ENSG00000174944	1.99	5.52x10 ⁻¹⁴
SCARF1	ENSG00000074660	1.69	4.27x10 ⁻¹³
CALHM6	ENSG00000188820	1.63	7.71x10 ⁻¹³
GBP6	ENSG00000183347	2.03	5.01x10 ⁻¹²
GBP1P1	ENSG00000225492	2.55	5.01x10 ⁻¹²
APOL4	ENSG00000100336	2.64	1.18x10 ⁻¹¹
LHFPL2	ENSG00000145685	1.20	1.94x10 ⁻¹¹
GBP4	ENSG00000162654	1.15	4.21x10 ⁻¹¹
CD274	ENSG00000120217	1.80	4.21x10 ⁻¹¹
ANKRD22	ENSG00000152766	2.41	4.21x10 ⁻¹¹
PDCD1LG2	ENSG00000197646	1.82	1.90x10 ⁻¹⁰
SERPING1	ENSG00000149131	2.19	1.90x10 ⁻¹⁰
SEPTIN4	ENSG00000108387	2.1	1.90x10 ⁻¹⁰
GBP1	ENSG00000117228	1.51	2.26x10 ⁻¹⁰
APOL1	ENSG00000100342	1.13	2.91x10 ⁻¹⁰
BATF2	ENSG00000168062	2.01	4.24x10 ⁻¹⁰
LINC02555	ENSG00000260943	1.74	7.85x10 ⁻¹⁰
CDCP1	ENSG00000163814	1.45	8.31x10 ⁻¹⁰
ETV7	ENSG00000010030	2.06	1.63x10 ⁻⁰⁹
SIAH2	ENSG00000181788	1.13	2.86x10 ⁻⁰⁹
FCGR1CP	ENSG00000265531	1.97	5.13x10 ⁻⁰⁹
ETV7-AS1	ENSG00000224666	2.06	5.56x10 ⁻⁰⁹
TGM2	ENSG00000198959	1.29	5.56x10 ⁻⁰⁹
FCGR1A	ENSG00000150337	1.77	2.62x10 ⁻⁰⁸
ENSG00000224789	ENSG00000224789	1.33	2.71x10 ⁻⁰⁸
ACOD1	ENSG00000102794	1.87	2.71x10 ⁻⁰⁸
WARS1	ENSG00000140105	1.03	3.7x10 ⁻⁰⁸
ALDH1A1	ENSG00000165092	1.02	4.7x10 ⁻⁰⁸
RAB20	ENSG00000139832	1.12	5.01x10 ⁻⁰⁸
GLRX5	ENSG00000182512	1.22	5.01x10 ⁻⁰⁸
SLAMF8	ENSG00000158714	1.22	5.03x10 ⁻⁰⁸
LINC02471	ENSG00000223914	1.87	5.18x10 ⁻⁰⁸
SECTM1	ENSG00000141574	1.14	5.59x10 ⁻⁰⁸
IL31RA	ENSG00000164509	1.55	5.91x10 ⁻⁰⁸
TRIM10	ENSG00000204613	1.36	6.78x10 ⁻⁰⁸
AHSP	ENSG00000169877	1.81	8.45x10 ⁻⁰⁸
FLVCR2	ENSG00000119686	1.01	9.65x10 ⁻⁰⁸
ENSG00000237372	ENSG00000237372	1.5	1.55x10 ⁻⁰⁷
GK-IT1	ENSG00000229331	1.2	1.58x10 ⁻⁰⁷
ALAS2	ENSG00000158578	1.63	1.58x10 ⁻⁰⁷
FCGR1B	ENSG00000198019	1.48	1.83x10 ⁻⁰⁷
SELENBP1	ENSG00000143416	1.59	2.12x10 ⁻⁰⁷

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
KLC3	ENSG00000104892	1.68	2.21x10 ⁻⁰⁷
SDC3	ENSG00000162512	1.49	2.85x10 ⁻⁰⁷
RHCE	ENSG00000188672	1.25	3.03x10 ⁻⁰⁷
SLC4A1	ENSG00000004939	1.47	3.47x10 ⁻⁰⁷
GK	ENSG00000198814	1.06	5.61x10 ⁻⁰⁷
LINC02528	ENSG00000226004	1.89	8.53x10 ⁻⁰⁷
ATF3	ENSG00000162772	1.37	1x10 ⁻⁰⁶
MFSD2B	ENSG00000205639	1.07	1.08x10 ⁻⁰⁶
ENSG00000241886	ENSG00000241886	1.07	1.18x10 ⁻⁰⁶
SPTB	ENSG00000070182	1.39	1.29x10 ⁻⁰⁶
ARHGEF37	ENSG00000183111	1.04	1.62x10 ⁻⁰⁶
CARD17	ENSG00000255221	1.65	1.69x10 ⁻⁰⁶
ENSG00000228302	ENSG00000228302	1.86	1.69x10 ⁻⁰⁶
EPB42	ENSG00000166947	1.51	1.74x10 ⁻⁰⁶
ENSG00000254602	ENSG00000254602	1.75	1.75x10 ⁻⁰⁶
RN7SKP26	ENSG00000222179	1.12	1.93x10 ⁻⁰⁶
KLF1	ENSG00000105610	1.33	2.11x10 ⁻⁰⁶
TRIM58	ENSG00000162722	1.29	2.17x10 ⁻⁰⁶
KRT1	ENSG00000167768	1.52	2.29x10 ⁻⁰⁶
RBM38	ENSG00000132819	1	3.42x10 ⁻⁰⁶
TMEM217	ENSG00000172738	1.06	3.45x10 ⁻⁰⁶
BPGM	ENSG00000172331	1.12	3.45x10 ⁻⁰⁶
LAP3	ENSG00000002549	1.02	3.57x10 ⁻⁰⁶
RAB6B	ENSG00000154917	1.37	3.66x10 ⁻⁰⁶
RHD	ENSG00000187010	1.13	4.12x10 ⁻⁰⁶
C1QA	ENSG00000173372	1.41	4.32x10 ⁻⁰⁶
PLEK2	ENSG00000100558	1.31	4.36x10 ⁻⁰⁶
OSBP2	ENSG00000184792	1.37	4.36x10 ⁻⁰⁶
ENSG00000259182	ENSG00000259182	1.4	4.43x10 ⁻⁰⁶
SLC6A9	ENSG00000196517	1.14	4.49x10 ⁻⁰⁶
SLC6A19	ENSG00000174358	2	4.89x10 ⁻⁰⁶
PROS1	ENSG00000184500	1.31	5.07x10 ⁻⁰⁶
CXCL9	ENSG00000138755	1.41	5.79x10 ⁻⁰⁶
RN7SL124P	ENSG00000242113	1.22	5.91x10 ⁻⁰⁶
TNFAIP6	ENSG00000123610	1.34	6.57x10 ⁻⁰⁶
C1QB	ENSG00000173369	1.58	6.57x10 ⁻⁰⁶
TNS1	ENSG00000079308	1.27	6.91x10 ⁻⁰⁶
CAPN11	ENSG00000137225	1.39	7.09x10 ⁻⁰⁶
GMPR	ENSG00000137198	1.17	8.08x10 ⁻⁰⁶
ERFE	ENSG00000178752	2.38	8.21x10 ⁻⁰⁶
NRN1	ENSG00000124785	1.61	9.13x10 ⁻⁰⁶
FAM210B	ENSG00000124098	1.07	9.36x10 ⁻⁰⁶
CACNA1E	ENSG00000198216	1.33	9.65x10 ⁻⁰⁶
ABCG2	ENSG00000118777	1.16	1.33x10 ⁻⁰⁵
ANKRD9	ENSG00000156381	1.04	1.44x10 ⁻⁰⁵
ANK1	ENSG00000029534	1.01	1.72x10 ⁻⁰⁵
PKD1L3	ENSG00000277481	1.14	1.76x10 ⁻⁰⁵
MYL9	ENSG00000101335	1.13	1.78x10 ⁻⁰⁵
IGF2BP2	ENSG00000073792	1.01	1.87x10 ⁻⁰⁵
GRIN3A	ENSG00000198785	1.15	1.9x10 ⁻⁰⁵
RUNDC3A-AS1	ENSG00000267750	1.22	1.9x10 ⁻⁰⁵
TMOD1	ENSG00000136842	1.03	1.94x10 ⁻⁰⁵
IGHV4-31	ENSG00000231475	-1.59	2.17x10 ⁻⁰⁵
GK4P	ENSG00000178146	1.01	2.29x10 ⁻⁰⁵
DCAF12	ENSG00000198876	1.05	2.43x10 ⁻⁰⁵
SLC6A8	ENSG00000130821	1.08	2.81x10 ⁻⁰⁵
SLC25A39	ENSG00000013306	1.04	3.53x10 ⁻⁰⁵
SLC2A4	ENSG00000181856	1.17	3.72x10 ⁻⁰⁵
ITGA2B	ENSG00000005961	1.17	3.73x10 ⁻⁰⁵
SMIM24	ENSG00000095932	1.16	3.9x10 ⁻⁰⁵
IL21-AS1	ENSG00000227145	-1.02	3.96x10 ⁻⁰⁵
ABCC13	ENSG00000243064	1.13	5.21x10 ⁻⁰⁵

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
CORIN	ENSG00000145244	1.41	7.49x10 ⁻⁰⁵
HBA2	ENSG00000188536	1.26	7.49x10 ⁻⁰⁵
IFIT1B	ENSG00000204010	1.34	8.07x10 ⁻⁰⁵
STRADB	ENSG00000082146	1.01	8.3x10 ⁻⁰⁵
LINC00570	ENSG00000224177	1.07	8.32x10 ⁻⁰⁵
C1QC	ENSG00000159189	1.66	9.9x10 ⁻⁰⁵
GK3P	ENSG00000229894	1.05	1.11x10 ⁻⁰⁴
LINC01151	ENSG00000253819	1.15	1.31x10 ⁻⁰⁴
CASP5	ENSG00000137757	1.17	1.37x10 ⁻⁰⁴
C15orf48	ENSG00000166920	1.02	1.66x10 ⁻⁰⁴
SNCA	ENSG00000145335	1	1.78x10 ⁻⁰⁴
ACHE	ENSG00000087085	1.29	1.89x10 ⁻⁰⁴
OR2W3	ENSG00000238243	1.28	2.01x10 ⁻⁰⁴
IL27	ENSG00000197272	1.21	2.01x10 ⁻⁰⁴
ROR1	ENSG00000185483	-1.05	2.09x10 ⁻⁰⁴
COL10A1	ENSG00000123500	1.47	2.25x10 ⁻⁰⁴
ENSG00000249244	ENSG00000249244	1.04	2.27x10 ⁻⁰⁴
ENSG00000258831	ENSG00000258831	-1.2	2.44x10 ⁻⁰⁴
HEPACAM2	ENSG00000188175	1.01	2.7x10 ⁻⁰⁴
CA1	ENSG00000133742	1.33	2.8x10 ⁻⁰⁴
CMBL	ENSG00000164237	1.38	2.93x10 ⁻⁰⁴
AQP10	ENSG00000143595	1.03	4.24x10 ⁻⁰⁴
KREMEN1	ENSG00000183762	1.17	4.28x10 ⁻⁰⁴
VSIG4	ENSG00000155659	1.05	4.41x10 ⁻⁰⁴
LINC02772	ENSG00000271736	1.03	4.95x10 ⁻⁰⁴
ENSG00000273573	ENSG00000273573	1.18	5.4x10 ⁻⁰⁴
TSP02	ENSG00000112212	1.08	5.4x10 ⁻⁰⁴
ENSG00000268903	ENSG00000268903	1.03	6.16x10 ⁻⁰⁴
EXOC3L1	ENSG00000179044	1.11	6.77x10 ⁻⁰⁴
IDO1	ENSG00000131203	1.31	6.99x10 ⁻⁰⁴
UICLM	ENSG00000233392	-1.24	7.7x10 ⁻⁰⁴
ENSG00000253838	ENSG00000253838	1.33	7.9x10 ⁻⁰⁴
ENSG00000280270	ENSG00000280270	1.18	9.27x10 ⁻⁰⁴
TEK	ENSG00000120156	-1.19	9.97x10 ⁻⁰⁴
TRIM51BP	ENSG00000204455	1.14	1.14x10 ⁻⁰³
CLEC2L	ENSG00000236279	1.15	1.19x10 ⁻⁰³
BCAM	ENSG00000187244	1.3	1.32x10 ⁻⁰³
GSTA7P	ENSG00000237674	1.74	1.45x10 ⁻⁰³
IGHV3-30	ENSG00000270550	-1.02	1.47x10 ⁻⁰³
XCR1	ENSG00000173578	-1.06	1.89x10 ⁻⁰³
SGIP1	ENSG00000118473	1.06	1.93x10 ⁻⁰³
NECTIN2	ENSG00000130202	1	2.04x10 ⁻⁰³
SAMD14	ENSG00000167100	1.06	2.07x10 ⁻⁰³
MT1L	ENSG00000260549	1.01	2.14x10 ⁻⁰³
CD300LD	ENSG00000204345	1.09	2.53x10 ⁻⁰³
EDNRB	ENSG00000136160	1.12	2.53x10 ⁻⁰³
SOCS3-DT	ENSG00000266970	1.01	2.67x10 ⁻⁰³
CNN1	ENSG00000130176	1.07	3.07x10 ⁻⁰³
TMPRSS9	ENSG00000178297	1.07	3.35x10 ⁻⁰³
FAM83A	ENSG00000147689	1.42	3.53x10 ⁻⁰³
ENSG00000256948	ENSG00000256948	1.18	3.55x10 ⁻⁰³
IGF2	ENSG00000167244	1.14	3.69x10 ⁻⁰³
ENSG00000251139	ENSG00000251139	1.04	4.26x10 ⁻⁰³
LINC01671	ENSG00000225431	-1.03	4.49x10 ⁻⁰³
RAP1GAP	ENSG00000076864	1.68	5.02x10 ⁻⁰³
THEM5	ENSG00000196407	1.05	6.9x10 ⁻⁰³
FAM210CP	ENSG00000213770	1.11	8.4x10 ⁻⁰³
FAM20A	ENSG00000108950	1.08	9.48x10 ⁻⁰³
CXCL10	ENSG00000169245	1.16	1.04x10 ⁻⁰²
C4BPA	ENSG00000123838	1.17	1.22x10 ⁻⁰²
OR2T8	ENSG00000177462	1	1.32x10 ⁻⁰²
MEG3	ENSG00000214548	-1.07	1.6x10 ⁻⁰²

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
CLEC4F	ENSG00000152672	-1.28	1.81x10 ⁻⁰²
TUBB2A	ENSG00000137267	1.07	2.18x10 ⁻⁰²

APPENDIX D

SDE GENES IN THE RNA-SEQ DATA WHEN COMPARING MICROBIOLOGICALLY CONFIRMED TB VERSUS UNLIKELY TB AT FOLLOW-UP VISIT

Table D.1: SDE genes in the RNA-Seq data when comparing microbiologically confirmed TB vs. unlikely TB at baseline. *SDE* = Significantly Differentially Expressed; *FC* = Fold Change; *BH* = Benjamini-Hochberg

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
ATP5ME	ENSG00000169020	1.09	1.44x10 ⁻⁰⁹
PRR13P2	ENSG00000254884	1.14	3.93x10 ⁻⁰⁸
ENSG00000260895	ENSG00000260895	-1.14	1.05x10 ⁻⁰⁵
MT-ND6	ENSG00000198695	1.05	1.7x10 ⁻⁰⁴
TLR9	ENSG00000239732	-1.04	2.19x10 ⁻⁰⁴
LINC01270	ENSG00000203999	-1.02	2.52x10 ⁻⁰⁴
IGLV6-57	ENSG00000211640	-1.19	3.59x10 ⁻⁰⁴
ENSG00000258878	ENSG00000258878	-1.06	6.62x10 ⁻⁰⁴
IGLV3-25	ENSG00000211659	-1.08	7.46x10 ⁻⁰⁴
LINC02596	ENSG00000233431	-1	8.08x10 ⁻⁰⁴
IGHV3-30	ENSG00000270550	-1.1	1.03x10 ⁻⁰³
INSC	ENSG00000188487	-1.04	1.75x10 ⁻⁰³
IGHV4-31	ENSG00000231475	-1.27	1.9x10 ⁻⁰³
BICDL2	ENSG00000162069	-1.16	3.21x10 ⁻⁰³
ENSG00000243181	ENSG00000243181	1.09	3.47x10 ⁻⁰³
FAM149A	ENSG00000109794	-1.01	4.62x10 ⁻⁰³
ENSG00000279942	ENSG00000279942	-1.19	4.83x10 ⁻⁰³
ENSG00000241494	ENSG00000241494	1.08	7.23x10 ⁻⁰³
RPL26P19	ENSG00000226221	1.02	1.28x10 ⁻⁰²
RNF182	ENSG00000180537	-1.65	1.35x10 ⁻⁰²
RPL34P18	ENSG00000240509	1.02	1.67x10 ⁻⁰²
ENSG00000254789	ENSG00000254789	-1.12	2.33x10 ⁻⁰²
ENSG00000261616	ENSG00000261616	-1.02	2.51x10 ⁻⁰²
CHST6	ENSG00000183196	-1.02	3.17x10 ⁻⁰²
SNX7	ENSG00000162627	1.22	3.78x10 ⁻⁰²
CAPN13	ENSG00000162949	-1.04	4.73x10 ⁻⁰²

APPENDIX E

SDE GENES IN THE RNA-SEQ DATA WHEN COMPARING MICROBIOLOGICALLY CONFIRMED TB AFTER ANTI-TB TREATMENT VERSUS MICROBIOLOGICALLY CONFIRMED TB BEFORE ANTI-TB TREATMENT

Table E.1: SDE genes in the RNA-Seq data when comparing microbiologically confirmed TB after anti-TB treatment vs. microbiologically confirmed TB before anti-TB treatment. *SDE* = Significantly Differentially Expressed; *FC* = Fold Change; *BH* = Benjamini-Hochberg

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted p-value
SMARCD3	ENSG00000082014	-1.49	2.11x10 ⁻²⁵
TIFA	ENSG00000145365	-1.27	5.36x10 ⁻²⁵
STOM	ENSG00000148175	-1.09	5.36x10 ⁻²⁵
C1QB	ENSG00000173369	-2.91	8.18x10 ⁻²⁵
C1QA	ENSG00000173372	-2.51	9.53x10 ⁻²⁵
CCR2	ENSG00000121807	-1.02	2.52x10 ⁻²³
CREG1	ENSG00000143162	-1.24	3.19x10 ⁻²³
METTL9	ENSG00000197006	-1.02	3.19x10 ⁻²³
DRAM1	ENSG00000136048	-1.16	4.19x10 ⁻²³
MMP8	ENSG00000118113	-3.25	2.01x10 ⁻²²
RAB13	ENSG00000143545	-1.38	2.02x10 ⁻²²
S100A8	ENSG00000143546	-2.24	2.83x10 ⁻²²
VAMP5	ENSG00000168899	-1.2	4.05x10 ⁻²²
FZD5	ENSG00000163251	-1.09	4.25x10 ⁻²²
AIM2	ENSG00000163568	-1.68	8.72x10 ⁻²²
DSE	ENSG00000111817	-1.02	9.56x10 ⁻²²
ADCY3	ENSG00000138031	-1.39	1.11x10 ⁻²¹
ABCA13	ENSG00000179869	-2.24	1.14x10 ⁻²¹
FLVCR2	ENSG00000119686	-1.41	1.56x10 ⁻²¹
DUSP3	ENSG00000108861	-1.04	3.54x10 ⁻²¹
CD274	ENSG00000120217	-2.01	4.54x10 ⁻²¹
MAP3K20	ENSG00000091436	-1.08	9.75x10 ⁻²¹
CALHM6	ENSG00000188820	-1.66	9.75x10 ⁻²¹
DOCK1	ENSG00000150760	-1.33	1.93x10 ⁻²⁰
S100A9	ENSG00000163220	-1.8	2.01x10 ⁻²⁰
ENSG00000217275	ENSG00000217275	-1.72	2.48x10 ⁻²⁰
LY96	ENSG00000154589	-1.37	2.6x10 ⁻²⁰
DDX10P1	ENSG00000237135	-1.52	3.38x10 ⁻²⁰
CARD16	ENSG00000204397	-1.15	4.14x10 ⁻²⁰
RPL7P18	ENSG00000243806	-1.39	6.58x10 ⁻²⁰
FAM20A	ENSG00000108950	-2.82	6.81x10 ⁻²⁰
HORMAD1	ENSG00000143452	-1.36	7.24x10 ⁻²⁰
GBP6	ENSG00000183347	-2.06	1.94x10 ⁻¹⁹
GASK1B	ENSG00000164125	-1.2	1.95x10 ⁻¹⁹
PLD4	ENSG00000166428	1.12	2.7x10 ⁻¹⁹
HPSE	ENSG00000173083	-1.19	3.94x10 ⁻¹⁹
CCRL2	ENSG00000121797	-1.09	4.14x10 ⁻¹⁹
LMNB1	ENSG00000113368	-1.24	4.14x10 ⁻¹⁹
MSR1	ENSG00000038945	-1.83	4.31x10 ⁻¹⁹
LHFPL2	ENSG00000145685	-1.24	4.57x10 ⁻¹⁹
MROCK1	ENSG00000227502	-2.33	4.57x10 ⁻¹⁹

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
CAMP	ENSG00000164047	-1.86	5.36x10 ⁻¹⁹
PDCD1LG2	ENSG00000197646	-1.97	6.81x10 ⁻¹⁹
EVI2A	ENSG00000126860	-1.47	7.63x10 ⁻¹⁹
TNFAIP6	ENSG00000123610	-1.98	1.16x10 ⁻¹⁸
IFI30	ENSG00000216490	-1.01	1.27x10 ⁻¹⁸
EFCAB2	ENSG00000203666	-1.04	1.61x10 ⁻¹⁸
LINC02009	ENSG00000283646	-2.24	1.79x10 ⁻¹⁸
CD36	ENSG00000135218	-1.01	1.83x10 ⁻¹⁸
FCGR1A	ENSG00000150337	-2.12	1.88x10 ⁻¹⁸
TCN2	ENSG00000185339	-1.56	1.88x10 ⁻¹⁸
C1QC	ENSG00000159189	-2.95	1.91x10 ⁻¹⁸
LRRK2-DT	ENSG00000225342	-1.09	2.36x10 ⁻¹⁸
ANKRD22	ENSG00000152766	-2.47	3.05x10 ⁻¹⁸
ICAM1	ENSG00000090339	-1.03	3.31x10 ⁻¹⁸
CARNS1	ENSG00000172508	1.03	3.4x10 ⁻¹⁸
PLOD2	ENSG00000152952	-1.55	3.57x10 ⁻¹⁸
WDFY3-AS2	ENSG00000180769	-1.94	3.61x10 ⁻¹⁸
FCGR1CP	ENSG00000265531	-2.25	3.78x10 ⁻¹⁸
ZNF415P1	ENSG00000266127	-1.12	4.47x10 ⁻¹⁸
CFAP58-DT	ENSG00000231233	-1.82	4.73x10 ⁻¹⁸
KCNJ2-AS1	ENSG00000267365	-1.47	5.2x10 ⁻¹⁸
MED12L	ENSG00000144893	-1.27	5.21x10 ⁻¹⁸
ENSG00000253214	ENSG00000253214	-1.48	5.38x10 ⁻¹⁸
ENSG00000254420	ENSG00000254420	-3.34	6.83x10 ⁻¹⁸
CASP5	ENSG00000137757	-1.98	7.07x10 ⁻¹⁸
G0S2	ENSG00000123689	-1.91	7.56x10 ⁻¹⁸
ZNF438	ENSG00000183621	-1.18	8.59x10 ⁻¹⁸
METTTL7B	ENSG00000170439	-2.67	9.49x10 ⁻¹⁸
SRGN	ENSG00000122862	-1.25	1.09x10 ⁻¹⁷
SEPTIN4	ENSG00000108387	-2.19	1.17x10 ⁻¹⁷
UICLM	ENSG00000233392	2.25	1.42x10 ⁻¹⁷
SCARF1	ENSG00000074660	-1.52	1.87x10 ⁻¹⁷
HMGA2-AS1	ENSG00000197301	-1.39	1.91x10 ⁻¹⁷
RAB20	ENSG00000139832	-1.33	2.25x10 ⁻¹⁷
GYG1	ENSG00000163754	-1.09	2.34x10 ⁻¹⁷
OMG	ENSG00000126861	-1.28	2.56x10 ⁻¹⁷
CCN3	ENSG00000136999	1.75	2.61x10 ⁻¹⁷
LINC01013	ENSG00000228495	1.48	2.82x10 ⁻¹⁷
SIGLEC8	ENSG00000105366	2.82	3.01x10 ⁻¹⁷
ENSG00000233461	ENSG00000233461	-1.34	4.6x10 ⁻¹⁷
NLRC4	ENSG00000091106	-1.23	4.84x10 ⁻¹⁷
S100A12	ENSG00000163221	-2.13	5.27x10 ⁻¹⁷
ANXA3	ENSG00000138772	-2.15	8.33x10 ⁻¹⁷
ENSG00000283399	ENSG00000283399	1.47	1.45x10 ⁻¹⁶
PTGDR2	ENSG00000183134	1.73	1.45x10 ⁻¹⁶
AK4	ENSG00000162433	-1.06	1.63x10 ⁻¹⁶
ANKRD9	ENSG00000156381	-1.48	1.79x10 ⁻¹⁶
SAT1	ENSG00000130066	-1.02	1.82x10 ⁻¹⁶
ENSG00000251139	ENSG00000251139	-2.18	2.12x10 ⁻¹⁶
FCGR1B	ENSG00000198019	-1.77	2.14x10 ⁻¹⁶
GPR84	ENSG00000139572	-2.06	2.31x10 ⁻¹⁶
SIAH2	ENSG00000181788	-1.18	3.2x10 ⁻¹⁶
FCER1G	ENSG00000158869	-1.18	3.36x10 ⁻¹⁶
XK	ENSG00000047597	-1.36	3.54x10 ⁻¹⁶
PRSS33	ENSG00000103355	2.78	3.67x10 ⁻¹⁶
FCGBP	ENSG00000275395	1.2	3.72x10 ⁻¹⁶
OLIG2	ENSG00000205927	2.6	3.72x10 ⁻¹⁶
NKD1	ENSG00000140807	1.16	3.73x10 ⁻¹⁶
TLR5	ENSG00000187554	-1.39	3.78x10 ⁻¹⁶
CIBAR1	ENSG00000188343	-1.21	4.41x10 ⁻¹⁶
SOCS3	ENSG00000184557	-1.62	4.76x10 ⁻¹⁶
HP	ENSG00000257017	-2.4	4.9x10 ⁻¹⁶

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
CACNA1E	ENSG00000198216	-1.84	4.95x10 ⁻¹⁶
TLR1	ENSG00000174125	-1.04	5.22x10 ⁻¹⁶
EDNRB	ENSG00000136160	-2.2	5.33x10 ⁻¹⁶
MARCKS	ENSG00000277443	-1.03	6.35x10 ⁻¹⁶
SLC26A8	ENSG00000112053	-1.93	6.36x10 ⁻¹⁶
TFEC	ENSG00000105967	-1.18	7.28x10 ⁻¹⁶
ADM	ENSG00000148926	-1.64	9.34x10 ⁻¹⁶
MMRN1	ENSG00000138722	-1.33	1.01x10 ⁻¹⁵
GBP1P1	ENSG00000225492	-2.25	1.04x10 ⁻¹⁵
NAIPP1	ENSG00000250687	-1.89	1.11x10 ⁻¹⁵
FRMD3	ENSG00000172159	-1.04	1.3x10 ⁻¹⁵
HRK	ENSG00000135116	1.48	1.33x10 ⁻¹⁵
PSTPIP2	ENSG00000152229	-1.08	1.35x10 ⁻¹⁵
ITGA7	ENSG00000135424	-1.9	1.55x10 ⁻¹⁵
ADM-DT	ENSG00000254554	-1.45	1.6x10 ⁻¹⁵
CARD17	ENSG00000255221	-2.07	1.62x10 ⁻¹⁵
RARA-AS1	ENSG00000265666	-1.05	1.76x10 ⁻¹⁵
ERFE	ENSG00000178752	-3.3	1.82x10 ⁻¹⁵
LINC02555	ENSG00000260943	-1.71	1.97x10 ⁻¹⁵
GNG11	ENSG00000127920	-1.06	2.44x10 ⁻¹⁵
ASGR2	ENSG00000161944	-1.22	2.66x10 ⁻¹⁵
CTNNAL1	ENSG00000119326	-1.24	2.88x10 ⁻¹⁵
PKD1L3	ENSG00000277481	-1.57	3.07x10 ⁻¹⁵
KREMEN1	ENSG00000183762	-1.93	3.07x10 ⁻¹⁵
MCEMP1	ENSG00000183019	-2.17	3.44x10 ⁻¹⁵
ALOX15	ENSG00000161905	2.62	4.49x10 ⁻¹⁵
NAIP	ENSG00000249437	-1.53	4.59x10 ⁻¹⁵
TGM2	ENSG00000198959	-1.31	4.59x10 ⁻¹⁵
ARG2	ENSG00000081181	-1.29	5.23x10 ⁻¹⁵
VNN1	ENSG00000112299	-1.73	5.6x10 ⁻¹⁵
RPS3P7	ENSG00000243101	1.27	7.27x10 ⁻¹⁵
GK	ENSG00000198814	-1.23	7.37x10 ⁻¹⁵
NRXN1	ENSG00000179915	1.88	7.59x10 ⁻¹⁵
ARHGEF37	ENSG00000183111	-1.26	7.59x10 ⁻¹⁵
VSIG4	ENSG00000155659	-1.69	8.95x10 ⁻¹⁵
LINC01270	ENSG00000203999	-1.49	9.17x10 ⁻¹⁵
ENSG00000254602	ENSG00000254602	-2.16	1.27x10 ⁻¹⁴
CCR3	ENSG00000183625	1.55	1.31x10 ⁻¹⁴
TOX2	ENSG00000124191	1.11	1.48x10 ⁻¹⁴
CCR5AS	ENSG00000223552	-1.24	1.64x10 ⁻¹⁴
ENSG00000279693	ENSG00000279693	1.92	1.74x10 ⁻¹⁴
ENSG00000258101	ENSG00000258101	-1.25	1.77x10 ⁻¹⁴
CSTA	ENSG00000121552	-1.23	1.78x10 ⁻¹⁴
HCAR3	ENSG00000255398	-1.34	2.1x10 ⁻¹⁴
ALDH1A2	ENSG00000128918	-1.47	2.1x10 ⁻¹⁴
ENSG00000268734	ENSG00000268734	-1.27	2.16x10 ⁻¹⁴
FLT3	ENSG00000122025	-1.14	2.19x10 ⁻¹⁴
LINC01800	ENSG00000234572	1.01	2.21x10 ⁻¹⁴
PROS1	ENSG00000184500	-1.63	2.33x10 ⁻¹⁴
PLBD1	ENSG00000121316	-1.18	2.86x10 ⁻¹⁴
KCNJ2	ENSG00000123700	-1.26	2.86x10 ⁻¹⁴
ACOD1	ENSG00000102794	-1.93	2.91x10 ⁻¹⁴
LCN2	ENSG00000148346	-1.45	2.93x10 ⁻¹⁴
ENSG00000273320	ENSG00000273320	-1.15	2.96x10 ⁻¹⁴
ENSG00000237927	ENSG00000237927	-1.58	3x10 ⁻¹⁴
ADAM23	ENSG00000114948	1.21	3.03x10 ⁻¹⁴
DPRXP2	ENSG00000271304	-1.92	3.05x10 ⁻¹⁴
IL27	ENSG00000197272	-1.84	3.41x10 ⁻¹⁴
NFE4	ENSG00000230257	-1.36	3.84x10 ⁻¹⁴
SEZ6L	ENSG00000100095	1.49	3.86x10 ⁻¹⁴
ENSG00000274213	ENSG00000274213	-1.21	4.31x10 ⁻¹⁴
CACNG8	ENSG00000142408	2.01	4.44x10 ⁻¹⁴

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
SLC22A4	ENSG00000197208	-1.19	5.03x10 ⁻¹⁴
PLSCR4	ENSG00000114698	-1.47	5.13x10 ⁻¹⁴
PPBP	ENSG00000163736	-1.28	5.26x10 ⁻¹⁴
CLEC4D	ENSG00000166527	-1.41	5.34x10 ⁻¹⁴
ENSG00000272567	ENSG00000272567	-1.11	5.6x10 ⁻¹⁴
CEACAM8	ENSG00000124469	-1.78	5.74x10 ⁻¹⁴
TLR2	ENSG00000137462	-1.04	6.13x10 ⁻¹⁴
MYO10	ENSG00000145555	-1.68	6.95x10 ⁻¹⁴
LINC01281	ENSG00000235304	1.28	7.43x10 ⁻¹⁴
CLIC2	ENSG00000155962	-1.07	7.59x10 ⁻¹⁴
CLU	ENSG00000120885	-1.23	1.05x10 ⁻¹³
NAIP3	ENSG00000248477	-1.63	1.08x10 ⁻¹³
OSM	ENSG00000099985	-1.37	1.09x10 ⁻¹³
IL18R1	ENSG00000115604	-1.08	1.11x10 ⁻¹³
SMPD3	ENSG00000103056	1.56	1.13x10 ⁻¹³
ADGRE4P	ENSG00000268758	1.58	1.14x10 ⁻¹³
FCAR	ENSG00000186431	-1.22	1.43x10 ⁻¹³
RN7SL124P	ENSG00000242113	-1.48	1.48x10 ⁻¹³
CLEC1B	ENSG00000165682	-1.17	1.49x10 ⁻¹³
ENSG00000259268	ENSG00000259268	-1.74	1.54x10 ⁻¹³
TNFSF10	ENSG00000121858	-1.08	1.57x10 ⁻¹³
IL5RA	ENSG00000091181	1.85	1.82x10 ⁻¹³
ENSG00000270030	ENSG00000270030	-1.58	1.83x10 ⁻¹³
ENSG00000249309	ENSG00000249309	-1.37	1.91x10 ⁻¹³
CACNG6	ENSG00000130433	2.2	1.98x10 ⁻¹³
GK4P	ENSG00000178146	-1.26	2.14x10 ⁻¹³
COL17A1	ENSG00000065618	-2.29	2.29x10 ⁻¹³
ODF3B	ENSG00000177989	-1.09	2.3x10 ⁻¹³
CD177	ENSG00000204936	-3.95	2.58x10 ⁻¹³
ENSG00000272914	ENSG00000272914	-1.18	2.84x10 ⁻¹³
SUCNR1	ENSG00000198829	-1.18	2.98x10 ⁻¹³
ENSG00000224789	ENSG00000224789	-1.31	3.08x10 ⁻¹³
TMEM45A	ENSG00000181458	-1.32	3.08x10 ⁻¹³
CA1	ENSG00000133742	-1.94	3.08x10 ⁻¹³
EGF	ENSG00000138798	-1.2	3.15x10 ⁻¹³
BASP1-AS1	ENSG00000215196	-1.82	3.44x10 ⁻¹³
SLC2A3	ENSG00000059804	-1.03	3.52x10 ⁻¹³
ENSG00000267737	ENSG00000267737	-1.4	3.63x10 ⁻¹³
BCL2A1	ENSG00000140379	-1.25	4.29x10 ⁻¹³
CACNB4	ENSG00000182389	1.09	4.52x10 ⁻¹³
ENSG00000235321	ENSG00000235321	-1.23	4.57x10 ⁻¹³
AHSP	ENSG00000169877	-1.83	4.87x10 ⁻¹³
CREB5	ENSG00000146592	-1.23	5.65x10 ⁻¹³
RNASE2CP	ENSG00000136315	-1.6	5.66x10 ⁻¹³
ENSG00000259182	ENSG00000259182	-1.63	5.89x10 ⁻¹³
ENSG00000228302	ENSG00000228302	-2.09	6.21x10 ⁻¹³
ENSG00000234436	ENSG00000234436	-1.4	6.51x10 ⁻¹³
SLC29A1	ENSG00000112759	1.3	7.59x10 ⁻¹³
ENSG00000254639	ENSG00000254639	-1.38	8.35x10 ⁻¹³
LTF	ENSG00000012223	-1.74	8.69x10 ⁻¹³
ENSG00000260861	ENSG00000260861	-1.07	8.73x10 ⁻¹³
ENSG00000275393	ENSG00000275393	1.25	9.34x10 ⁻¹³
GK-IT1	ENSG00000229331	-1.22	9.76x10 ⁻¹³
GADD45G	ENSG00000130222	-1.06	1x10 ⁻¹²
LINC02217	ENSG00000248455	-1.97	1.05x10 ⁻¹²
TMEM119	ENSG00000183160	-1.86	1.06x10 ⁻¹²
GBP5	ENSG00000154451	-1.25	1.1x10 ⁻¹²
SLPI	ENSG00000124107	-1.76	1.15x10 ⁻¹²
CLEC4E	ENSG00000166523	-1.11	1.19x10 ⁻¹²
LILRA5	ENSG00000187116	-1.33	1.2x10 ⁻¹²
RHD	ENSG00000187010	-1.27	2.01x10 ⁻¹²
IL34	ENSG00000157368	1.99	2.23x10 ⁻¹²

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
ENSG00000230492	ENSG00000230492	-1.81	2.28x10 ⁻¹²
OPLAH	ENSG00000178814	-1.36	2.33x10 ⁻¹²
MYL9	ENSG00000101335	-1.36	2.54x10 ⁻¹²
ADAMTS7P1	ENSG00000274376	1.54	2.74x10 ⁻¹²
BMX	ENSG00000102010	-1.7	2.83x10 ⁻¹²
DYRK3	ENSG00000143479	-1.13	2.91x10 ⁻¹²
ENSG00000258732	ENSG00000258732	1.12	3.24x10 ⁻¹²
RNU6-917P	ENSG00000252103	-1.26	3.24x10 ⁻¹²
MARCO	ENSG00000019169	-1.43	3.27x10 ⁻¹²
ENSG00000283602	ENSG00000283602	-1.07	3.59x10 ⁻¹²
FXSD2	ENSG00000137731	1.06	3.59x10 ⁻¹²
LIPC-AS1	ENSG00000259293	-1.59	3.62x10 ⁻¹²
IFITM3	ENSG00000142089	-1.29	3.68x10 ⁻¹²
CEACAM6	ENSG00000086548	-1.89	3.74x10 ⁻¹²
LRG1	ENSG00000171236	-1.3	4.21x10 ⁻¹²
ENSG00000280190	ENSG00000280190	1.01	4.59x10 ⁻¹²
ENSG00000253986	ENSG00000253986	-1.4	4.67x10 ⁻¹²
ENSG00000273573	ENSG00000273573	-1.71	4.68x10 ⁻¹²
ENSG00000273272	ENSG00000273272	-1.19	4.77x10 ⁻¹²
OR8R1P	ENSG00000256347	1.58	4.99x10 ⁻¹²
RNU6-226P	ENSG00000200648	-1.08	5.32x10 ⁻¹²
IL18RAP	ENSG00000115607	-1.13	5.34x10 ⁻¹²
ENSG00000214708	ENSG00000214708	-1.44	5.34x10 ⁻¹²
SPTB	ENSG00000070182	-1.46	5.64x10 ⁻¹²
ENSG00000249244	ENSG00000249244	-1.36	5.8x10 ⁻¹²
TMEM144	ENSG00000164124	-1.07	5.88x10 ⁻¹²
ENSG00000254859	ENSG00000254859	-1.1	6.46x10 ⁻¹²
PLSCR1	ENSG00000188313	-1.12	6.63x10 ⁻¹²
ENSG00000223401	ENSG00000223401	-1.24	6.7x10 ⁻¹²
OLFM4	ENSG00000102837	-2.48	6.85x10 ⁻¹²
GBP1	ENSG00000117228	-1.22	6.93x10 ⁻¹²
GALNT14	ENSG00000158089	-1.57	7.27x10 ⁻¹²
GSEC	ENSG00000280832	-1.18	7.46x10 ⁻¹²
ENSG00000261211	ENSG00000261211	-1.22	7.93x10 ⁻¹²
GRIN3A	ENSG00000198785	-1.35	7.99x10 ⁻¹²
BPGM	ENSG00000172331	-1.21	8.27x10 ⁻¹²
FGF13	ENSG00000129682	-1.92	8.58x10 ⁻¹²
MAOA	ENSG00000189221	-1.98	9.62x10 ⁻¹²
NAMPT	ENSG00000105835	-1.23	9.82x10 ⁻¹²
ENSG00000256433	ENSG00000256433	-1.11	9.83x10 ⁻¹²
NRN1	ENSG00000124785	-1.82	1.05x10 ⁻¹¹
SPNS3	ENSG00000182557	1.1	1.08x10 ⁻¹¹
TMPRSS15	ENSG00000154646	1.82	1.14x10 ⁻¹¹
CRISP3	ENSG00000096006	-1.77	1.16x10 ⁻¹¹
ENSG00000253115	ENSG00000253115	-1.59	1.23x10 ⁻¹¹
SMIM10L2A	ENSG00000178947	1.01	1.26x10 ⁻¹¹
GK3P	ENSG00000229894	-1.3	1.27x10 ⁻¹¹
SERPING1	ENSG00000149131	-1.74	1.36x10 ⁻¹¹
TMEM217	ENSG00000172738	-1.14	1.37x10 ⁻¹¹
APOL4	ENSG00000100336	-1.93	1.52x10 ⁻¹¹
SLC51A	ENSG00000163959	-1.88	1.52x10 ⁻¹¹
RAB6B	ENSG00000154917	-1.47	1.53x10 ⁻¹¹
SECTM1	ENSG00000141574	-1.04	1.57x10 ⁻¹¹
KCNJ15	ENSG00000157551	-1.31	1.72x10 ⁻¹¹
AJAP1	ENSG00000196581	1.72	1.72x10 ⁻¹¹
SCN1B	ENSG00000105711	-1.07	1.78x10 ⁻¹¹
FPR2	ENSG00000171049	-1.08	2.01x10 ⁻¹¹
ITGA2B	ENSG00000005961	-1.38	2.05x10 ⁻¹¹
MGAM2	ENSG00000257743	-1.42	2.07x10 ⁻¹¹
MYOF	ENSG00000138119	-1.01	2.07x10 ⁻¹¹
LAP3	ENSG00000002549	-1.08	2.21x10 ⁻¹¹
IL1B	ENSG00000125538	-1.1	2.26x10 ⁻¹¹

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
HCAR2	ENSG00000182782	-1.16	2.32x10 ⁻¹¹
NUDT19P5	ENSG00000251073	-1.16	2.39x10 ⁻¹¹
SLC1A3	ENSG00000079215	-1.45	3.57x10 ⁻¹¹
LINC02528	ENSG00000226004	-1.87	3.57x10 ⁻¹¹
ENSG00000241886	ENSG00000241886	-1.07	3.6x10 ⁻¹¹
AQP9	ENSG00000103569	-1.16	3.8x10 ⁻¹¹
FAM225A	ENSG00000231528	-1.61	4.12x10 ⁻¹¹
HEPACAM2	ENSG00000188175	-1.32	4.12x10 ⁻¹¹
PRMT5-DT	ENSG00000257285	-1.03	4.27x10 ⁻¹¹
CYSTM1	ENSG00000120306	-1.15	4.49x10 ⁻¹¹
ENSG00000280115	ENSG00000280115	1.19	4.59x10 ⁻¹¹
LINC02363	ENSG00000180712	-1.05	4.6x10 ⁻¹¹
ENSG00000214797	ENSG00000214797	1	5.05x10 ⁻¹¹
ADAMTS2	ENSG00000087116	-2.58	5.06x10 ⁻¹¹
CDCP1	ENSG00000163814	-1.15	5.07x10 ⁻¹¹
ROR1	ENSG00000185483	1.3	5.19x10 ⁻¹¹
PLIN5	ENSG00000214456	-1.18	5.29x10 ⁻¹¹
ANKRD34B	ENSG00000189127	-1.72	5.5x10 ⁻¹¹
FFAR3	ENSG00000185897	-1.91	6.04x10 ⁻¹¹
NAMPTP1	ENSG00000229644	-1.29	6.05x10 ⁻¹¹
GASK1B-AS1	ENSG00000248429	-1.02	6.2x10 ⁻¹¹
VNN2	ENSG00000112303	-1.08	6.55x10 ⁻¹¹
TCN1	ENSG00000134827	-1.09	6.59x10 ⁻¹¹
FAP	ENSG00000078098	-1.66	7.28x10 ⁻¹¹
PLIN4	ENSG00000167676	-1.2	7.49x10 ⁻¹¹
OLIG1	ENSG00000184221	1.1	7.63x10 ⁻¹¹
RHCE	ENSG00000188672	-1.17	8.52x10 ⁻¹¹
ERG	ENSG00000157554	-1.3	8.66x10 ⁻¹¹
NLRP9P1	ENSG00000256581	-1.2	8.82x10 ⁻¹¹
ENSG00000254789	ENSG00000254789	-2.05	8.87x10 ⁻¹¹
DAZL	ENSG00000092345	1.55	9.32x10 ⁻¹¹
MS4A4A	ENSG00000110079	-1.34	9.55x10 ⁻¹¹
OLR1	ENSG00000173391	-1.63	1.03x10 ⁻¹⁰
ENSG00000259641	ENSG00000259641	1.16	1.11x10 ⁻¹⁰
AQP1	ENSG00000240583	-1.3	1.17x10 ⁻¹⁰
CLEC6A	ENSG00000205846	-1.09	1.19x10 ⁻¹⁰
P4HA2	ENSG00000072682	-1.02	1.19x10 ⁻¹⁰
KLHDC8A	ENSG00000162873	-1.38	1.24x10 ⁻¹⁰
PCOLCE2	ENSG00000163710	-2.3	1.28x10 ⁻¹⁰
ENSG00000261549	ENSG00000261549	-1.12	1.29x10 ⁻¹⁰
MPO	ENSG00000005381	-1.23	1.31x10 ⁻¹⁰
BATF2	ENSG00000168062	-1.54	1.31x10 ⁻¹⁰
P2RY14	ENSG00000174944	-1.25	1.35x10 ⁻¹⁰
ARG1	ENSG00000118520	-1.73	1.41x10 ⁻¹⁰
ENSG00000249138	ENSG00000249138	-1.12	1.51x10 ⁻¹⁰
SEMA6B	ENSG00000167680	-1.36	1.54x10 ⁻¹⁰
ENSG00000279569	ENSG00000279569	1.61	1.66x10 ⁻¹⁰
PLAAT5	ENSG00000168004	1.5	1.82x10 ⁻¹⁰
SLC6A12	ENSG00000111181	-1.16	1.82x10 ⁻¹⁰
RHAG	ENSG00000112077	-1.18	1.86x10 ⁻¹⁰
CDK15	ENSG00000138395	1.83	1.88x10 ⁻¹⁰
INHBA	ENSG00000122641	-1.41	1.96x10 ⁻¹⁰
HOXA9	ENSG00000078399	-1.39	1.97x10 ⁻¹⁰
NPTX1	ENSG00000171246	1.38	2.12x10 ⁻¹⁰
CACNA2D3	ENSG00000157445	1.03	2.13x10 ⁻¹⁰
ACSL1	ENSG00000151726	-1.24	2.17x10 ⁻¹⁰
BEND7	ENSG00000165626	-1.54	2.17x10 ⁻¹⁰
ACHE	ENSG00000087085	-1.58	2.22x10 ⁻¹⁰
DYNLT5	ENSG00000152760	-1.5	2.33x10 ⁻¹⁰
BCL6	ENSG00000113916	-1.01	2.37x10 ⁻¹⁰
GLRX5	ENSG00000182512	-1.04	2.42x10 ⁻¹⁰
CCDC73	ENSG00000186714	-1.1	2.61x10 ⁻¹⁰

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
MIR3945HG	ENSG00000251230	-1.36	2.72x10 ⁻¹⁰
SPTA1	ENSG00000163554	-1.11	2.81x10 ⁻¹⁰
ALPL	ENSG00000162551	-1.44	2.84x10 ⁻¹⁰
CD177P1	ENSG00000204933	-2.69	2.94x10 ⁻¹⁰
GPRI1	ENSG00000164850	-1.51	3x10 ⁻¹⁰
STEAP4	ENSG00000127954	-1.02	3.02x10 ⁻¹⁰
NRCAM	ENSG00000091129	1.19	3.05x10 ⁻¹⁰
LINC01222	ENSG00000233410	1.19	3.08x10 ⁻¹⁰
DBH-AS1	ENSG00000225756	1.11	3.16x10 ⁻¹⁰
SIGLEC5	ENSG00000105501	-1.08	3.23x10 ⁻¹⁰
SPATA20P1	ENSG00000231123	-1.62	3.36x10 ⁻¹⁰
SERPINB10	ENSG00000242550	-1.66	3.69x10 ⁻¹⁰
NMRAL2P	ENSG00000171658	-1.53	3.83x10 ⁻¹⁰
ALAS2	ENSG00000158578	-1.43	4.1x10 ⁻¹⁰
CYP1B1	ENSG00000138061	-1.21	4.21x10 ⁻¹⁰
ENSG00000243273	ENSG00000243273	-1.98	4.38x10 ⁻¹⁰
NRXN3	ENSG00000021645	-1.48	4.8x10 ⁻¹⁰
TDRD9	ENSG00000156414	-1.22	4.89x10 ⁻¹⁰
ANOS1	ENSG00000011201	-1.84	5.32x10 ⁻¹⁰
SIPA1L2	ENSG00000116991	-1.01	5.39x10 ⁻¹⁰
FPR1	ENSG00000171051	-1.01	6.06x10 ⁻¹⁰
LIPH	ENSG00000163898	-1.08	6.17x10 ⁻¹⁰
LINC02471	ENSG00000223914	-1.56	6.24x10 ⁻¹⁰
CDH2	ENSG00000170558	1.39	6.49x10 ⁻¹⁰
EMID1	ENSG00000186998	-1.26	6.54x10 ⁻¹⁰
QPCT	ENSG00000115828	-1.03	6.61x10 ⁻¹⁰
RAB44	ENSG00000255587	1.06	7.3x10 ⁻¹⁰
DEFA4	ENSG00000164821	-2.02	7.38x10 ⁻¹⁰
ENSG00000254288	ENSG00000254288	-1.21	7.93x10 ⁻¹⁰
PLAU	ENSG00000122861	-1.13	7.95x10 ⁻¹⁰
RN7SKP26	ENSG00000222179	-1.05	8.27x10 ⁻¹⁰
FGFR2	ENSG00000066468	1.32	8.7x10 ⁻¹⁰
ANO5	ENSG00000171714	-1.23	8.92x10 ⁻¹⁰
ELOVL3	ENSG00000119915	-1.13	9.6x10 ⁻¹⁰
ENSG00000250696	ENSG00000250696	1.67	9.81x10 ⁻¹⁰
DUSP13	ENSG00000079393	-1.55	1.05x10 ⁻⁰⁹
H1-6	ENSG00000187475	-1.04	1.08x10 ⁻⁰⁹
FFAR2	ENSG00000126262	-1.03	1.13x10 ⁻⁰⁹
TRIM6	ENSG00000121236	-1.32	1.24x10 ⁻⁰⁹
ENSG00000237372	ENSG00000237372	-1.28	1.28x10 ⁻⁰⁹
MMP9	ENSG00000100985	-1.54	1.32x10 ⁻⁰⁹
RHOXF1P1	ENSG00000234493	1.65	1.43x10 ⁻⁰⁹
ENSG00000260859	ENSG00000260859	-1.18	1.48x10 ⁻⁰⁹
ENSG00000275210	ENSG00000275210	-1.09	1.53x10 ⁻⁰⁹
BPI	ENSG00000101425	-1.27	1.64x10 ⁻⁰⁹
PDE8B	ENSG00000113231	1.22	1.68x10 ⁻⁰⁹
LINC00900	ENSG00000246100	-1.17	1.68x10 ⁻⁰⁹
ENSG00000259600	ENSG00000259600	-1.07	1.7x10 ⁻⁰⁹
KCNMA1	ENSG00000156113	-1.31	1.73x10 ⁻⁰⁹
TMC5	ENSG00000103534	-1.16	1.73x10 ⁻⁰⁹
TREML1	ENSG00000161911	-1.07	1.8x10 ⁻⁰⁹
ENSG00000233214	ENSG00000233214	-1.31	2.25x10 ⁻⁰⁹
P2RX6	ENSG00000099957	1.06	2.28x10 ⁻⁰⁹
ENSG00000268833	ENSG00000268833	-1.34	2.38x10 ⁻⁰⁹
NECAB1	ENSG00000123119	-2.44	2.67x10 ⁻⁰⁹
ENSG00000260188	ENSG00000260188	-1.09	2.83x10 ⁻⁰⁹
ARPP21	ENSG00000172995	1.23	2.88x10 ⁻⁰⁹
CA4	ENSG00000167434	-1.17	2.98x10 ⁻⁰⁹
CLEC5A	ENSG00000258227	-1.05	3.12x10 ⁻⁰⁹
DDO	ENSG00000203797	-1.14	3.18x10 ⁻⁰⁹
CCN2	ENSG00000118523	1.4	3.22x10 ⁻⁰⁹
CETP	ENSG00000087237	-1.05	3.26x10 ⁻⁰⁹

Continues on next page

APPENDIX E. LIST OF SDE GENES

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
ENSG00000273419	ENSG00000273419	1.06	3.37x10 ⁻⁰⁹
KAZN	ENSG00000189337	-1.14	3.39x10 ⁻⁰⁹
SYCP2L	ENSG00000153157	1.47	3.39x10 ⁻⁰⁹
DEFA3	ENSG00000239839	-2.25	3.39x10 ⁻⁰⁹
MFSD6L	ENSG00000185156	1.07	3.58x10 ⁻⁰⁹
CAPN13	ENSG00000162949	-1.92	3.6x10 ⁻⁰⁹
SDR42E1P5	ENSG00000236785	-1.26	4.02x10 ⁻⁰⁹
LINC02915	ENSG00000175746	1.4	4.16x10 ⁻⁰⁹
ENSG00000269887	ENSG00000269887	-1.09	4.62x10 ⁻⁰⁹
SPX	ENSG00000134548	-1.07	4.62x10 ⁻⁰⁹
SOCS3-DT	ENSG00000266970	-1.39	4.8x10 ⁻⁰⁹
CALD1	ENSG00000122786	-1.05	5.68x10 ⁻⁰⁹
MTCO3P23	ENSG00000259651	-1.03	5.74x10 ⁻⁰⁹
ENSG00000279400	ENSG00000279400	-1.6	6.09x10 ⁻⁰⁹
PTPRU	ENSG00000060656	-1.18	6.37x10 ⁻⁰⁹
ENSG00000249763	ENSG00000249763	1.83	6.54x10 ⁻⁰⁹
BMP3	ENSG00000152785	1.22	6.74x10 ⁻⁰⁹
ENSG00000261537	ENSG00000261537	1.03	6.76x10 ⁻⁰⁹
CCNA1	ENSG00000133101	-1.4	6.9x10 ⁻⁰⁹
IFTM3P4	ENSG00000230191	-1.27	7.11x10 ⁻⁰⁹
ENSG00000278909	ENSG00000278909	1.07	7.26x10 ⁻⁰⁹
ANKUB1	ENSG00000206199	1.23	7.43x10 ⁻⁰⁹
COL4A1	ENSG00000187498	1.26	9.42x10 ⁻⁰⁹
ENSG00000280274	ENSG00000280274	1.2	1.09x10 ⁻⁰⁸
HSPA4L	ENSG00000164070	-1.27	1.1x10 ⁻⁰⁸
ENSG00000273972	ENSG00000273972	-1.05	1.15x10 ⁻⁰⁸
AGAP7P	ENSG00000264204	1.05	1.17x10 ⁻⁰⁸
PCYT1B	ENSG00000102230	-1.04	1.19x10 ⁻⁰⁸
DYSF	ENSG00000135636	-1.04	1.23x10 ⁻⁰⁸
CORO2B	ENSG00000103647	1.06	1.28x10 ⁻⁰⁸
RCBTB2P1	ENSG00000235150	-1.15	1.34x10 ⁻⁰⁸
SDC3	ENSG00000162512	-1.22	1.35x10 ⁻⁰⁸
ENSG00000233030	ENSG00000233030	-1.18	1.38x10 ⁻⁰⁸
LINC01226	ENSG00000284543	1.09	1.56x10 ⁻⁰⁸
TMCC2	ENSG00000133069	-1	1.66x10 ⁻⁰⁸
LINC01094	ENSG00000251442	-1.03	1.66x10 ⁻⁰⁸
PRTN3	ENSG00000196415	-1.64	1.67x10 ⁻⁰⁸
ENSG00000231794	ENSG00000231794	-1.46	1.73x10 ⁻⁰⁸
ENSG00000268615	ENSG00000268615	-1.05	1.74x10 ⁻⁰⁸
ENSG00000255328	ENSG00000255328	-1.19	1.77x10 ⁻⁰⁸
CAPN11	ENSG00000137225	-1.25	1.79x10 ⁻⁰⁸
C15orf48	ENSG00000166920	-1.09	1.85x10 ⁻⁰⁸
ADORA3	ENSG00000282608	1.11	1.88x10 ⁻⁰⁸
CMBL	ENSG00000164237	-1.51	2.06x10 ⁻⁰⁸
RETN	ENSG00000104918	-1.38	2.08x10 ⁻⁰⁸
LINC00534	ENSG00000253394	-1.05	2.12x10 ⁻⁰⁸
ENSG00000268170	ENSG00000268170	-1.22	2.25x10 ⁻⁰⁸
DISP3	ENSG00000204624	1.04	2.52x10 ⁻⁰⁸
RPL35AP26	ENSG00000243607	-1.03	2.61x10 ⁻⁰⁸
RAP1GAP	ENSG00000076864	-2.34	2.64x10 ⁻⁰⁸
PRSS41	ENSG00000215148	1.54	2.69x10 ⁻⁰⁸
NECTIN2	ENSG00000130202	-1.27	2.7x10 ⁻⁰⁸
CCL23	ENSG00000274736	1.85	3.06x10 ⁻⁰⁸
CROCC2	ENSG00000226321	1.28	3.75x10 ⁻⁰⁸
ENSG00000273812	ENSG00000273812	-1.01	3.86x10 ⁻⁰⁸
CYP19A1	ENSG00000137869	-1.76	3.88x10 ⁻⁰⁸
KNDC1	ENSG00000171798	1.2	4.18x10 ⁻⁰⁸
ENSG00000274373	ENSG00000274373	1.08	4.18x10 ⁻⁰⁸
ENSG00000274173	ENSG00000274173	-2.01	4.24x10 ⁻⁰⁸
ITGB3	ENSG00000259207	-1.08	4.4x10 ⁻⁰⁸
ENSG00000254522	ENSG00000254522	1.08	4.45x10 ⁻⁰⁸
LDC1P	ENSG00000260386	1.2	4.58x10 ⁻⁰⁸

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
MAOB	ENSG00000069535	-1.27	4.59x10 ⁻⁰⁸
ABCG2	ENSG00000118777	-1.05	5.17x10 ⁻⁰⁸
ATF3	ENSG00000162772	-1.1	5.21x10 ⁻⁰⁸
HOXA10	ENSG00000253293	-1.08	5.27x10 ⁻⁰⁸
COL23A1	ENSG00000050767	-1.06	5.47x10 ⁻⁰⁸
ENSG00000130489	ENSG00000130489	-1.04	5.53x10 ⁻⁰⁸
LINC01093	ENSG00000249173	-1.81	5.96x10 ⁻⁰⁸
RIMKLA	ENSG00000177181	1.18	5.97x10 ⁻⁰⁸
HRH4	ENSG00000134489	1.11	6.53x10 ⁻⁰⁸
ATP1B2	ENSG00000129244	-1.16	6.57x10 ⁻⁰⁸
ETV7	ENSG00000010030	-1.35	6.6x10 ⁻⁰⁸
TSPO2	ENSG00000112212	-1.2	6.94x10 ⁻⁰⁸
PRUNE2	ENSG00000106772	-1.06	7.48x10 ⁻⁰⁸
BACE2-IT1	ENSG00000224388	1.11	7.65x10 ⁻⁰⁸
RPS3AP32	ENSG00000240047	-1.07	7.84x10 ⁻⁰⁸
ENSG00000224029	ENSG00000224029	1.05	7.92x10 ⁻⁰⁸
ENAH	ENSG00000154380	1.04	8x10 ⁻⁰⁸
LINC02596	ENSG00000233431	-1.08	8.77x10 ⁻⁰⁸
PSMA6P1	ENSG00000215414	-1.09	8.87x10 ⁻⁰⁸
TMEM158	ENSG00000249992	-1.22	9.21x10 ⁻⁰⁸
IFIT1B	ENSG00000204010	-1.29	9.24x10 ⁻⁰⁸
SLC2A14	ENSG00000173262	-1.22	9.45x10 ⁻⁰⁸
EPB41L4B	ENSG00000095203	-2.02	9.88x10 ⁻⁰⁸
LHFPL6	ENSG00000183722	-1.2	1.04x10 ⁻⁰⁷
SHROOM4	ENSG00000158352	-1.09	1.08x10 ⁻⁰⁷
ATP2C2	ENSG00000064270	-1.51	1.12x10 ⁻⁰⁷
PROK2	ENSG00000163421	-1.12	1.13x10 ⁻⁰⁷
ENSG00000214772	ENSG00000214772	-1.17	1.27x10 ⁻⁰⁷
SLC44A3	ENSG00000143036	-1.06	1.31x10 ⁻⁰⁷
SIAH3	ENSG00000215475	1.26	1.52x10 ⁻⁰⁷
ZNF683	ENSG00000176083	1.03	1.52x10 ⁻⁰⁷
ENSG00000234358	ENSG00000234358	-1.06	1.59x10 ⁻⁰⁷
TRIM58	ENSG00000162722	-1.03	1.6x10 ⁻⁰⁷
ENSG00000275772	ENSG00000275772	1	1.66x10 ⁻⁰⁷
ENSG00000259755	ENSG00000259755	-1.05	1.81x10 ⁻⁰⁷
AQP10	ENSG00000143595	-1.08	1.86x10 ⁻⁰⁷
IL31RA	ENSG00000164509	-1.07	1.92x10 ⁻⁰⁷
SHISA4	ENSG00000198892	1.26	1.93x10 ⁻⁰⁷
LINC02073	ENSG00000267452	1.73	2.04x10 ⁻⁰⁷
FOLR3	ENSG00000110203	-1.36	2.08x10 ⁻⁰⁷
SRGAP1	ENSG00000196935	-1.05	2.2x10 ⁻⁰⁷
SLC6A19	ENSG00000174358	-1.62	2.34x10 ⁻⁰⁷
IL1RL1	ENSG00000115602	1	2.47x10 ⁻⁰⁷
ENSG00000225963	ENSG00000225963	1.02	2.52x10 ⁻⁰⁷
OLAH	ENSG00000152463	-1.87	2.66x10 ⁻⁰⁷
LYPD2	ENSG00000197353	1.48	3.33x10 ⁻⁰⁷
CLRN1-AS1	ENSG00000239265	-1.55	3.36x10 ⁻⁰⁷
LINC01151	ENSG00000253819	-1.08	3.44x10 ⁻⁰⁷
IFITM3P2	ENSG00000223722	-1.14	3.71x10 ⁻⁰⁷
CSF2RBP1	ENSG00000232254	-1.13	3.94x10 ⁻⁰⁷
ENSG00000250138	ENSG00000250138	-1.38	4.38x10 ⁻⁰⁷
SLC7A8	ENSG00000092068	1.22	4.48x10 ⁻⁰⁷
ETV7-AS1	ENSG00000224666	-1.29	4.64x10 ⁻⁰⁷
SFTPD	ENSG00000133661	1.03	4.77x10 ⁻⁰⁷
MIR3945	ENSG00000266698	-1.19	5.32x10 ⁻⁰⁷
TRMT9B	ENSG00000250305	-1.12	5.36x10 ⁻⁰⁷
SAMD14	ENSG00000167100	-1.2	5.62x10 ⁻⁰⁷
CHST8	ENSG00000124302	-1.24	5.66x10 ⁻⁰⁷
LAP3P2	ENSG00000213500	-1.08	5.71x10 ⁻⁰⁷
PPARG	ENSG00000132170	-1.03	6.42x10 ⁻⁰⁷
SLC4A1	ENSG00000004939	-1.05	6.71x10 ⁻⁰⁷
LINC02345	ENSG00000259225	1.2	6.86x10 ⁻⁰⁷

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
ENSG00000267568	ENSG00000267568	1	7.42x10 ⁻⁰⁷
ENSG00000253190	ENSG00000253190	-1.05	7.65x10 ⁻⁰⁷
ENSG00000238009	ENSG00000238009	-1.05	7.8x10 ⁻⁰⁷
CORIN	ENSG00000145244	-1.24	8.59x10 ⁻⁰⁷
CNTN4-AS1	ENSG00000237990	1.3	8.73x10 ⁻⁰⁷
SGIP1	ENSG00000118473	-1.18	8.82x10 ⁻⁰⁷
ENSG00000256813	ENSG00000256813	1.25	8.84x10 ⁻⁰⁷
ZDHHC19	ENSG00000163958	-1.31	9.17x10 ⁻⁰⁷
ENSG00000253483	ENSG00000253483	-1.11	9.86x10 ⁻⁰⁷
MEG3	ENSG00000214548	1.44	1.11x10 ⁻⁰⁶
SYTL5	ENSG00000147041	1.51	1.28x10 ⁻⁰⁶
ENSG00000258878	ENSG00000258878	1.01	1.51x10 ⁻⁰⁶
TUNAR	ENSG00000250366	1.08	1.52x10 ⁻⁰⁶
ENSG00000283155	ENSG00000283155	-1.04	1.59x10 ⁻⁰⁶
SELENBP1	ENSG00000143416	-1.07	1.59x10 ⁻⁰⁶
BTNL10	ENSG00000215811	-1.04	1.6x10 ⁻⁰⁶
LRFN2	ENSG00000156564	1.03	1.67x10 ⁻⁰⁶
CYP7B1	ENSG00000172817	1.06	1.71x10 ⁻⁰⁶
ENSG00000270972	ENSG00000270972	-1.08	1.84x10 ⁻⁰⁶
SHISA7	ENSG00000187902	-1.21	1.85x10 ⁻⁰⁶
NCAM2	ENSG00000154654	1.08	1.97x10 ⁻⁰⁶
ENPP7P8	ENSG00000255319	1.19	1.98x10 ⁻⁰⁶
RNASE1	ENSG00000129538	-1.36	2.03x10 ⁻⁰⁶
GPR42	ENSG00000126251	-1.42	2.1x10 ⁻⁰⁶
INSC	ENSG00000188487	-1.05	2.13x10 ⁻⁰⁶
ENSG00000261655	ENSG00000261655	1.12	2.13x10 ⁻⁰⁶
SCARA5	ENSG00000168079	1.09	2.37x10 ⁻⁰⁶
LINC02701	ENSG00000250508	-1.01	2.43x10 ⁻⁰⁶
ENSG00000228651	ENSG00000228651	-1.01	2.53x10 ⁻⁰⁶
SPP1	ENSG00000118785	1.35	2.77x10 ⁻⁰⁶
FAT1	ENSG00000083857	-1.02	3.55x10 ⁻⁰⁶
PMP22	ENSG00000109099	1.3	4.07x10 ⁻⁰⁶
SMIM1	ENSG00000235169	-1.08	5.05x10 ⁻⁰⁶
ENSG00000233785	ENSG00000233785	-1.06	5.66x10 ⁻⁰⁶
LMOD1	ENSG00000163431	1.5	5.98x10 ⁻⁰⁶
NAALAD2	ENSG00000077616	1.1	6.04x10 ⁻⁰⁶
ENSG00000280276	ENSG00000280276	-1.36	6.12x10 ⁻⁰⁶
ST18	ENSG00000147488	1.01	6.4x10 ⁻⁰⁶
MEIS2	ENSG00000134138	1.1	6.85x10 ⁻⁰⁶
GPR20	ENSG00000204882	1.07	7.39x10 ⁻⁰⁶
CNN1	ENSG00000130176	-1.1	8.23x10 ⁻⁰⁶
CTSG	ENSG00000100448	-1.4	8.64x10 ⁻⁰⁶
AP3B2	ENSG00000103723	-1.62	9.16x10 ⁻⁰⁶
IFI44	ENSG00000137965	-1.02	9.94x10 ⁻⁰⁶
IFIT3	ENSG00000119917	-1.09	1.02x10 ⁻⁰⁵
ENSG00000267749	ENSG00000267749	1	1.04x10 ⁻⁰⁵
MT1L	ENSG00000260549	-1.01	1.07x10 ⁻⁰⁵
TEK	ENSG00000120156	1.1	1.07x10 ⁻⁰⁵
FAM83A	ENSG00000147689	-1.48	1.11x10 ⁻⁰⁵
ENSG00000256948	ENSG00000256948	-1.22	1.14x10 ⁻⁰⁵
COL10A1	ENSG00000123500	-1.22	1.22x10 ⁻⁰⁵
RNU5E-1	ENSG00000199347	-1.03	1.24x10 ⁻⁰⁵
CXCL10	ENSG00000169245	-1.35	1.32x10 ⁻⁰⁵
ENSG00000280270	ENSG00000280270	-1.09	1.33x10 ⁻⁰⁵
PAQR9	ENSG00000188582	-1	2.35x10 ⁻⁰⁵
ELANE	ENSG00000197561	-1.06	2.36x10 ⁻⁰⁵
ENSG00000231412	ENSG00000231412	-1.38	2.56x10 ⁻⁰⁵
ADAMTS5	ENSG00000154736	1.11	2.69x10 ⁻⁰⁵
MYO3A	ENSG00000095777	1.02	2.81x10 ⁻⁰⁵
ART4	ENSG00000111339	-1.01	3.2x10 ⁻⁰⁵
SDC1	ENSG00000115884	-1.01	3.5x10 ⁻⁰⁵
RPS7P11	ENSG00000213326	-1.14	3.57x10 ⁻⁰⁵

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
RPL9P32	ENSG00000242100	-1.07	3.93x10 ⁻⁰⁵
MXRA8	ENSG00000162576	1.02	6.06x10 ⁻⁰⁵
TARM1	ENSG00000248385	-1.13	7.44x10 ⁻⁰⁵
KRT5	ENSG00000186081	1.02	7.66x10 ⁻⁰⁵
RHOXF1P3	ENSG00000282933	1	8.36x10 ⁻⁰⁵
ENSG00000180211	ENSG00000180211	-1	9.48x10 ⁻⁰⁵
LINC02068	ENSG00000223387	-1.21	9.8x10 ⁻⁰⁵
S100P	ENSG00000163993	-1.13	1.83x10 ⁻⁰⁴
RSAD2	ENSG00000134321	-1.09	2.01x10 ⁻⁰⁴
CLEC4F	ENSG00000152672	1.31	2.63x10 ⁻⁰⁴
BCAM	ENSG00000187244	-1.01	3.07x10 ⁻⁰⁴
RPS3AP25	ENSG00000232385	-1.02	4.29x10 ⁻⁰⁴
TFF3	ENSG00000160180	1.11	4.41x10 ⁻⁰⁴
RP1L1	ENSG00000183638	-1.13	5.39x10 ⁻⁰⁴
RPS3AP26	ENSG00000214389	-1.01	6.79x10 ⁻⁰⁴
LINC00958	ENSG00000251381	1.19	7.21x10 ⁻⁰⁴
IFI27	ENSG00000165949	-1.17	5.71x10 ⁻⁰³

APPENDIX F

SDE GENES IN THE RNA-SEQ DATA WHEN COMPARING MICROBIOLOGICALLY CONFIRMED TB VERSUS UNLIKELY TB IN THE MULTI-OMICS COHORT

Table F.1: SDE genes in the RNA-Seq data when comparing microbiologically confirmed TB vs. unlikely TB in the multi-omics cohort. *SDE* = Significantly Differentially Expressed; *FC* = Fold Change; *BH* = Benjamini-Hochberg

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
GBP6	ENSG00000183347	2.43	5.50x10 ⁻¹⁸
GBP5	ENSG00000154451	1.73	2.27x10 ⁻¹⁵
CD274	ENSG00000120217	1.83	8.24x10 ⁻¹⁵
SCARF1	ENSG00000074660	1.71	2.43x10 ⁻¹⁴
PDCD1LG2	ENSG00000197646	1.75	3.19x10 ⁻¹⁴
CYB561	ENSG00000008283	0.96	5.73x10 ⁻¹⁴
SEPTIN4	ENSG00000108387	2.25	2.06x10 ⁻¹³
P2RY14	ENSG00000174944	1.92	2.33x10 ⁻¹³
CALHM6	ENSG00000188820	1.29	1.33x10 ⁻¹²
POLB	ENSG00000070501	0.66	3.05x10 ⁻¹²
ETV7	ENSG00000010030	2.04	3.22x10 ⁻¹²
ANKRD22	ENSG00000152766	2.01	3.80x10 ⁻¹²
IRF1	ENSG00000125347	0.74	4.99x10 ⁻¹²
APOL2	ENSG00000128335	0.81	5.60x10 ⁻¹²
BATF2	ENSG00000168062	1.90	6.37x10 ⁻¹²
APOL1	ENSG00000100342	1.11	6.37x10 ⁻¹²
GBP2	ENSG00000162645	0.87	7.70x10 ⁻¹²
SIPA1L1	ENSG00000197555	0.63	8.50x10 ⁻¹²
SERPING1	ENSG00000149131	1.98	5.15x10 ⁻¹¹
ICAM1	ENSG00000090339	1.01	5.16x10 ⁻¹¹
LHFPL2	ENSG00000145685	1.12	1.12x10 ⁻¹⁰
PDE4B	ENSG00000184588	0.51	1.21x10 ⁻¹⁰
GBP1	ENSG00000117228	1.37	2.15x10 ⁻¹⁰
FCGR1A	ENSG00000150337	1.70	2.15x10 ⁻¹⁰
WARS1	ENSG00000140105	1.01	6.82x10 ⁻¹⁰
GAS8	ENSG00000141013	0.67	6.82x10 ⁻¹⁰
SLAMF8	ENSG00000158714	1.22	6.90x10 ⁻¹⁰
TIFA	ENSG00000145365	0.91	6.90x10 ⁻¹⁰
GBP4	ENSG00000162654	0.97	8.18x10 ⁻¹⁰
WDFY1	ENSG00000085449	0.59	8.95x10 ⁻¹⁰
F2RL1	ENSG00000164251	1.19	1.49x10 ⁻⁰⁹
FAS	ENSG00000026103	0.67	1.49x10 ⁻⁰⁹
ASPHD2	ENSG00000128203	0.56	2.38x10 ⁻⁰⁹
DUSP3	ENSG00000108861	0.78	3.22x10 ⁻⁰⁹
MYOF	ENSG00000138119	1.15	3.72x10 ⁻⁰⁹
C2	ENSG00000166278	0.84	4.42x10 ⁻⁰⁹
GK	ENSG00000198814	1.15	4.67x10 ⁻⁰⁹
SECTM1	ENSG00000141574	1.15	8.65x10 ⁻⁰⁹
PSTPIP2	ENSG00000152229	1.04	8.65x10 ⁻⁰⁹
SLC6A12	ENSG00000111181	1.45	8.71x10 ⁻⁰⁹
GRAMD1B	ENSG00000023171	0.86	1.14x10 ⁻⁰⁸
TAP1	ENSG00000168394	0.79	1.34x10 ⁻⁰⁸
AIM2	ENSG00000163568	1.01	2.18x10 ⁻⁰⁸

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
VAMP5	ENSG00000168899	0.77	2.18x10 ⁻⁰⁸
HCAR3	ENSG00000255398	1.28	2.24x10 ⁻⁰⁸
LMTK2	ENSG00000164715	0.63	2.57x10 ⁻⁰⁸
HCAR2	ENSG00000182782	1.28	2.81x10 ⁻⁰⁸
SMARCD3	ENSG00000082014	1.08	3.00x10 ⁻⁰⁸
FLVCR2	ENSG00000119686	1.05	3.00x10 ⁻⁰⁸
PSMB9	ENSG00000240065	0.54	3.82x10 ⁻⁰⁸
APOL6	ENSG00000221963	0.78	3.85x10 ⁻⁰⁸
EFCAB2	ENSG00000203666	0.83	4.00x10 ⁻⁰⁸
SMCO4	ENSG00000166002	0.59	4.20x10 ⁻⁰⁸
LAP3	ENSG00000002549	1.03	5.11x10 ⁻⁰⁸
DAPP1	ENSG00000070190	0.56	5.39x10 ⁻⁰⁸
DHRS12	ENSG00000102796	0.77	6.93x10 ⁻⁰⁸
VPS9D1	ENSG00000075399	0.81	8.00x10 ⁻⁰⁸
IL15	ENSG00000164136	0.64	8.14x10 ⁻⁰⁸
ERLIN1	ENSG00000107566	0.61	8.88x10 ⁻⁰⁸
ATF3	ENSG00000162772	1.2	9.22x10 ⁻⁰⁸
PRRG4	ENSG00000135378	0.99	9.22x10 ⁻⁰⁸
TRAFD1	ENSG00000135148	0.57	9.22x10 ⁻⁰⁸
ODF3B	ENSG00000177989	0.83	1.02x10 ⁻⁰⁷
ENSG00000261884	ENSG00000261884	0.51	1.18x10 ⁻⁰⁷
RAB20	ENSG00000139832	1.09	1.20x10 ⁻⁰⁷
SOCS1	ENSG00000185338	0.94	1.27x10 ⁻⁰⁷
CASP5	ENSG00000137757	1.28	1.76x10 ⁻⁰⁷
TAP2	ENSG00000204267	0.51	2.07x10 ⁻⁰⁷
MSRB2	ENSG00000148450	0.72	2.10x10 ⁻⁰⁷
DMXL2	ENSG00000104093	0.64	2.27x10 ⁻⁰⁷
KCNJ2	ENSG00000123700	1.10	2.40x10 ⁻⁰⁷
GCH1	ENSG00000131979	0.60	2.43x10 ⁻⁰⁷
DOCK4	ENSG00000128512	0.93	2.53x10 ⁻⁰⁷
GPR82	ENSG00000171657	-1.06	2.89x10 ⁻⁰⁷
TNFAIP6	ENSG00000123610	1.27	2.98x10 ⁻⁰⁷
CASP1	ENSG00000137752	0.62	2.98x10 ⁻⁰⁷
LRRK2	ENSG00000188906	0.8	3.33x10 ⁻⁰⁷
IL1B	ENSG00000125538	1.00	3.47x10 ⁻⁰⁷
KCNJ15	ENSG00000157551	1.18	3.92x10 ⁻⁰⁷
JAK2	ENSG00000096968	0.52	4.27x10 ⁻⁰⁷
CACNA1E	ENSG00000198216	1.43	4.76x10 ⁻⁰⁷
ERI1	ENSG00000104626	0.52	5.18x10 ⁻⁰⁷
STX11	ENSG00000135604	0.81	6.86x10 ⁻⁰⁷
LIMK2	ENSG00000182541	0.99	6.86x10 ⁻⁰⁷
DHRS9	ENSG00000073737	0.98	7.49x10 ⁻⁰⁷
GBP3	ENSG00000117226	0.68	7.67x10 ⁻⁰⁷
GM2A	ENSG00000196743	0.60	8.86x10 ⁻⁰⁷
STAT1	ENSG00000115415	0.72	1.29x10 ⁻⁰⁶
RAB24	ENSG00000169228	0.59	1.29x10 ⁻⁰⁶
CASP4	ENSG00000196954	0.56	1.38x10 ⁻⁰⁶
FAM241A	ENSG00000174749	0.52	1.47x10 ⁻⁰⁶
HLA-E	ENSG00000204592	0.51	1.91x10 ⁻⁰⁶
LACTB	ENSG00000103642	0.55	1.91x10 ⁻⁰⁶
NOD2	ENSG00000167207	0.80	1.92x10 ⁻⁰⁶
CFAP58	ENSG00000120051	0.78	2.32x10 ⁻⁰⁶
NPC2	ENSG00000119655	0.57	2.78x10 ⁻⁰⁶
ZNF438	ENSG00000183621	0.82	2.80x10 ⁻⁰⁶
UBE2L6	ENSG00000156587	0.74	3.02x10 ⁻⁰⁶
TRANK1	ENSG00000168016	0.51	3.08x10 ⁻⁰⁶
ADM	ENSG00000148926	1.09	3.16x10 ⁻⁰⁶
TRIM21	ENSG00000132109	0.51	3.33x10 ⁻⁰⁶
ALPK1	ENSG00000073331	0.70	4.01x10 ⁻⁰⁶
IL13RA1	ENSG00000131724	0.74	4.18x10 ⁻⁰⁶
TMEM140	ENSG00000146859	0.71	4.39x10 ⁻⁰⁶
ALDH1A1	ENSG00000165092	0.96	5.08x10 ⁻⁰⁶

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
CARD16	ENSG00000204397	0.63	5.08x10 ⁻⁰⁶
LRRC4	ENSG00000128594	0.92	5.74x10 ⁻⁰⁶
AIG1	ENSG00000146416	0.53	5.99x10 ⁻⁰⁶
CCRL2	ENSG00000121797	0.63	6.27x10 ⁻⁰⁶
IFITM3	ENSG00000142089	0.98	6.31x10 ⁻⁰⁶
FBXO6	ENSG00000116663	0.77	6.48x10 ⁻⁰⁶
FRMD3	ENSG00000172159	0.71	6.48x10 ⁻⁰⁶
TNFAIP2	ENSG00000185215	0.76	6.50x10 ⁻⁰⁶
SORT1	ENSG00000134243	0.82	8.12x10 ⁻⁰⁶
RALB	ENSG00000144118	0.69	8.12x10 ⁻⁰⁶
ATP1B1	ENSG00000143153	0.60	8.28x10 ⁻⁰⁶
SLC49A3	ENSG00000169026	0.65	9.80x10 ⁻⁰⁶
VSIG10	ENSG00000176834	0.67	1.18x10 ⁻⁰⁵
C1QB	ENSG00000173369	1.38	1.23x10 ⁻⁰⁵
NCF1	ENSG00000158517	0.71	1.30x10 ⁻⁰⁵
P2RY13	ENSG00000181631	0.76	1.31x10 ⁻⁰⁵
PLAUR	ENSG00000011422	0.72	1.35x10 ⁻⁰⁵
SNX10	ENSG00000086300	0.60	1.38x10 ⁻⁰⁵
FFAR2	ENSG00000126262	0.94	1.46x10 ⁻⁰⁵
IGSF6	ENSG00000140749	0.71	1.57x10 ⁻⁰⁵
LMNB1	ENSG00000113368	0.63	1.62x10 ⁻⁰⁵
BCL3	ENSG00000069399	0.67	1.72x10 ⁻⁰⁵
ZDHHC18	ENSG00000204160	0.66	1.75x10 ⁻⁰⁵
GADD45B	ENSG00000099860	0.64	1.78x10 ⁻⁰⁵
RNF149	ENSG00000163162	0.63	1.80x10 ⁻⁰⁵
CAPNS2	ENSG00000256812	0.74	1.87x10 ⁻⁰⁵
RNF19B	ENSG00000116514	0.54	1.90x10 ⁻⁰⁵
HRK	ENSG00000135116	-1.00	2.00x10 ⁻⁰⁵
TNFSF10	ENSG00000121858	0.67	2.02x10 ⁻⁰⁵
IRAK2	ENSG00000134070	0.53	2.18x10 ⁻⁰⁵
VSIG4	ENSG00000155659	1.44	2.24x10 ⁻⁰⁵
TIMM10	ENSG00000134809	0.71	2.28x10 ⁻⁰⁵
KLHL23	ENSG00000213160	-0.58	2.38x10 ⁻⁰⁵
FPR2	ENSG00000171049	0.93	2.41x10 ⁻⁰⁵
RBM47	ENSG00000163694	0.74	2.60x10 ⁻⁰⁵
CEACAM1	ENSG00000079385	0.93	2.71x10 ⁻⁰⁵
KREMEN1	ENSG00000183762	1.20	2.71x10 ⁻⁰⁵
NABP1	ENSG00000173559	0.58	2.73x10 ⁻⁰⁵
FGL2	ENSG00000127951	0.58	2.73x10 ⁻⁰⁵
REM2	ENSG00000139890	0.69	3.16x10 ⁻⁰⁵
CFAP92	ENSG00000114656	0.82	3.70x10 ⁻⁰⁵
STK3	ENSG00000104375	0.58	3.80x10 ⁻⁰⁵
MPZL2	ENSG00000149573	0.66	3.98x10 ⁻⁰⁵
ITPRIP	ENSG00000148841	0.51	4.45x10 ⁻⁰⁵
ADCY3	ENSG00000138031	0.90	4.53x10 ⁻⁰⁵
C1QA	ENSG00000173372	1.18	4.65x10 ⁻⁰⁵
ADGRE5	ENSG00000123146	0.63	4.73x10 ⁻⁰⁵
RHBDF2	ENSG00000129667	0.50	4.83x10 ⁻⁰⁵
SQOR	ENSG00000137767	0.54	5.14x10 ⁻⁰⁵
CSF2RB	ENSG00000100368	0.85	5.27x10 ⁻⁰⁵
FCGR3B	ENSG00000162747	0.96	5.34x10 ⁻⁰⁵
CD300A	ENSG00000167851	0.54	5.34x10 ⁻⁰⁵
AK4	ENSG00000162433	0.69	5.40x10 ⁻⁰⁵
CTSA	ENSG00000064601	0.62	6.92x10 ⁻⁰⁵
BEST1	ENSG00000167995	0.74	6.97x10 ⁻⁰⁵
CRISPLD2	ENSG00000103196	0.95	7.05x10 ⁻⁰⁵
TLR1	ENSG00000174125	0.66	7.24x10 ⁻⁰⁵
FCER1G	ENSG00000158869	0.78	7.31x10 ⁻⁰⁵
ABHD2	ENSG00000140526	0.50	7.56x10 ⁻⁰⁵
SELL	ENSG00000188404	0.54	7.7x10 ⁻⁰⁵
SLC26A8	ENSG00000112053	1.03	8.54x10 ⁻⁰⁵
GAS7	ENSG00000007237	0.76	8.80x10 ⁻⁰⁵

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
MED12L	ENSG00000144893	0.79	8.82x10 ⁻⁰⁵
BRI3	ENSG00000164713	0.53	8.82x10 ⁻⁰⁵
LITAF	ENSG00000189067	0.55	8.84x10 ⁻⁰⁵
ATP7B	ENSG00000123191	0.51	9.31x10 ⁻⁰⁵
GAB2	ENSG00000033327	0.59	9.33x10 ⁻⁰⁵
FAM8A1	ENSG00000137414	0.51	9.37x10 ⁻⁰⁵
SLC22A1	ENSG00000175003	0.72	9.57x10 ⁻⁰⁵
IFITM2	ENSG00000185201	0.74	9.57x10 ⁻⁰⁵
RTP4	ENSG00000136514	0.76	9.65x10 ⁻⁰⁵
ASPRV1	ENSG00000244617	0.86	9.71x10 ⁻⁰⁵
CSF3R	ENSG00000119535	0.83	9.72x10 ⁻⁰⁵
MXD1	ENSG00000059728	0.74	9.72x10 ⁻⁰⁵
TMEM121B	ENSG00000183307	0.87	9.72x10 ⁻⁰⁵
OSM	ENSG00000099985	0.81	9.93x10 ⁻⁰⁵
GHRL	ENSG00000157017	0.8	9.94x10 ⁻⁰⁵
FNDC3B	ENSG00000075420	0.53	1.00x10 ⁻⁰⁴
HSD3B7	ENSG00000099377	1.12	1.00x10 ⁻⁰⁴
FCGR2A	ENSG00000143226	0.80	1.00x10 ⁻⁰⁴
CREB5	ENSG00000146592	0.86	1.00x10 ⁻⁰⁴
STX3	ENSG00000166900	0.66	1.10x10 ⁻⁰⁴
PIK3AP1	ENSG00000155629	0.55	1.10x10 ⁻⁰⁴
PLIN5	ENSG00000214456	0.99	1.20x10 ⁻⁰⁴
ENSG00000188897	ENSG00000188897	0.97	1.20x10 ⁻⁰⁴
SAT1	ENSG00000130066	0.54	1.30x10 ⁻⁰⁴
CPD	ENSG00000108582	0.58	1.30x10 ⁻⁰⁴
SOCS3	ENSG00000184557	1.02	1.40x10 ⁻⁰⁴
LPCAT2	ENSG00000087253	0.69	1.40x10 ⁻⁰⁴
MSRB1	ENSG00000198736	0.75	1.40x10 ⁻⁰⁴
TSEN34	ENSG00000170892	0.51	1.40x10 ⁻⁰⁴
AQP9	ENSG00000103569	0.88	1.40x10 ⁻⁰⁴
PLIN4	ENSG00000167676	1.07	1.40x10 ⁻⁰⁴
CXCR1	ENSG00000163464	0.88	1.40x10 ⁻⁰⁴
NAMPT	ENSG00000105835	0.86	1.50x10 ⁻⁰⁴
RMI2	ENSG00000175643	0.64	1.60x10 ⁻⁰⁴
SBNO2	ENSG00000064932	0.66	1.60x10 ⁻⁰⁴
ENSG00000284526	ENSG00000284526	0.81	1.70x10 ⁻⁰⁴
LTBR	ENSG00000111321	0.58	1.70x10 ⁻⁰⁴
DISC1	ENSG00000162946	0.51	1.80x10 ⁻⁰⁴
TMEM150B	ENSG00000180061	0.69	1.80x10 ⁻⁰⁴
PGS1	ENSG00000087157	0.59	1.90x10 ⁻⁰⁴
SLC31A2	ENSG00000136867	0.55	1.90x10 ⁻⁰⁴
LYN	ENSG00000254087	0.52	2.1x10 ⁻⁰⁴
TYMP	ENSG00000025708	0.66	2.10x10 ⁻⁰⁴
SIPA1L2	ENSG00000116991	0.78	2.10x10 ⁻⁰⁴
TRIM22	ENSG00000132274	0.54	2.20x10 ⁻⁰⁴
IFI30	ENSG00000216490	0.51	2.20x10 ⁻⁰⁴
PARP14	ENSG00000173193	0.5	2.20x10 ⁻⁰⁴
MPZ	ENSG00000158887	0.58	2.20x10 ⁻⁰⁴
RNF24	ENSG00000101236	0.68 2.40x10 ⁻⁰⁴	
MCTP1	ENSG00000175471	0.53	2.40x10 ⁻⁰⁴
DUSP1	ENSG00000120129	0.7	2.50x10 ⁻⁰⁴
TECPR2	ENSG00000196663	0.66	2.50x10 ⁻⁰⁴
TLR2	ENSG00000137462	0.72	2.50x10 ⁻⁰⁴
GPR27	ENSG00000170837	0.76	2.60x10 ⁻⁰⁴
LRG1	ENSG00000171236	0.98	2.70x10 ⁻⁰⁴
CXCR2	ENSG00000180871	0.83	2.70x10 ⁻⁰⁴
TCN2	ENSG00000185339	0.77	2.70x10 ⁻⁰⁴
CLEC7A	ENSG00000172243	0.57	2.70x10 ⁻⁰⁴
PPP1R3B	ENSG00000173281	0.71	2.70x10 ⁻⁰⁴
CFH	ENSG00000000971	1.21	2.90x10 ⁻⁰⁴
NFKBIA	ENSG00000100906	0.51	2.90x10 ⁻⁰⁴
INSL3	ENSG00000248099	0.74	2.90x10 ⁻⁰⁴

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
FPR1	ENSG00000171051	0.82	2.90x10 ⁻⁰⁴
TSHZ3	ENSG00000121297	0.73	3.00x10 ⁻⁰⁴
HPSE	ENSG00000173083	0.68	3.00x10 ⁻⁰⁴
TLR8	ENSG00000101916	0.7	3.10x10 ⁻⁰⁴
HSPA6	ENSG00000173110	0.71	3.20x10 ⁻⁰⁴
IFITM1	ENSG00000185885	0.58	3.20x10 ⁻⁰⁴
SCN1B	ENSG00000105711	0.81	3.20x10 ⁻⁰⁴
CADM4	ENSG00000105767	0.73	3.50x10 ⁻⁰⁴
JUNB	ENSG00000171223	0.55	3.50x10 ⁻⁰⁴
B4GALT5	ENSG00000158470	0.65	3.50x10 ⁻⁰⁴
FBXL5	ENSG00000118564	0.51	3.50x10 ⁻⁰⁴
PARP9	ENSG00000138496	0.52	3.60x10 ⁻⁰⁴
IL1RAP	ENSG00000196083	0.67	3.60x10 ⁻⁰⁴
KLHDC8B	ENSG00000185909	0.66	3.60x10 ⁻⁰⁴
AGTRAP	ENSG00000177674	0.55	3.70x10 ⁻⁰⁴
HORMAD1	ENSG00000143452	0.74	3.70x10 ⁻⁰⁴
DENND3	ENSG00000105339	0.53	3.80x10 ⁻⁰⁴
MYL9	ENSG00000101335	1.02	4.00x10 ⁻⁰⁴
DSC2	ENSG00000134755	1.01	4.10x10 ⁻⁰⁴
PHC2	ENSG00000134686	0.62	4.20x10 ⁻⁰⁴
MEFV	ENSG00000103313	0.6	4.40x10 ⁻⁰⁴
TLR6	ENSG00000174130	0.51	4.40x10 ⁻⁰⁴
MAPK14	ENSG00000112062	0.58	4.60x10 ⁻⁰⁴
PHEX	ENSG00000102174	-0.73	4.80x10 ⁻⁰⁴
SIRPD	ENSG00000125900	0.65	4.80x10 ⁻⁰⁴
RTN2	ENSG00000125744	0.58	4.90x10 ⁻⁰⁴
SHKBP1	ENSG00000160410	0.51	5.10x10 ⁻⁰⁴
TNNT3	ENSG00000130595	0.68	5.20x10 ⁻⁰⁴
EPB41L3	ENSG00000082397	0.51	5.40x10 ⁻⁰⁴
NIBAN1	ENSG00000135842	0.73	5.50x10 ⁻⁰⁴
ATF5	ENSG00000169136	0.55	5.50x10 ⁻⁰⁴
TIMD4	ENSG00000145850	-0.60	5.60x10 ⁻⁰⁴
NECTIN2	ENSG00000130202	0.98	5.60x10 ⁻⁰⁴
TNFRSF1A	ENSG00000067182	0.52	5.70x10 ⁻⁰⁴
DNAH7	ENSG00000118997	0.70	5.90x10 ⁻⁰⁴
ACTA2	ENSG00000107796	0.58	6.00x10 ⁻⁰⁴
FOS	ENSG00000170345	0.61	6.00x10 ⁻⁰⁴
TGFA	ENSG00000163235	0.76	6.1x10 ⁻⁰⁴
MNDA	ENSG00000163563	0.58	6.1x10 ⁻⁰⁴
PML	ENSG00000140464	0.50	6.20x10 ⁻⁰⁴
PTAFR	ENSG00000169403	0.65	6.30x10 ⁻⁰⁴
TLR4	ENSG00000136869	0.63	6.80x10 ⁻⁰⁴
CXCL16	ENSG00000161921	0.73	6.90x10 ⁻⁰⁴
CLEC4E	ENSG00000166523	0.77	7.30x10 ⁻⁰⁴
MBOAT7	ENSG00000125505	0.65	7.30x10 ⁻⁰⁴
EPSTI1	ENSG00000133106	0.69	7.30x10 ⁻⁰⁴
ALDH1A2	ENSG00000128918	0.89	7.40x10 ⁻⁰⁴
FZD2	ENSG00000180340	0.51	7.50x10 ⁻⁰⁴
TNFSF13B	ENSG00000102524	0.63	7.70x10 ⁻⁰⁴
NEURL1	ENSG00000107954	-0.70	7.70x10 ⁻⁰⁴
JAG1	ENSG00000101384	0.60	7.80x10 ⁻⁰⁴
ACSL1	ENSG00000151726	0.81	8.10x10 ⁻⁰⁴
DOCK1	ENSG00000150760	0.56	8.40x10 ⁻⁰⁴
DDX60L	ENSG00000181381	0.54	8.60x10 ⁻⁰⁴
OR52K1	ENSG00000196778	0.65	8.90x10 ⁻⁰⁴
ENSG00000282988	ENSG00000282988	0.7	8.90x10 ⁻⁰⁴
PLXDC2	ENSG00000120594	0.51	8.90x10 ⁻⁰⁴
AP5B1	ENSG00000254470	0.54	9.40x10 ⁻⁰⁴
DYSF	ENSG00000135636	0.82	9.70x10 ⁻⁰⁴
FLOT1	ENSG00000137312	0.5	9.90x10 ⁻⁰⁴
RARA	ENSG00000131759	0.52	1.00x10 ⁻⁰³
NAIP	ENSG00000249437	0.81	1.00x10 ⁻⁰³

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
B3GNT5	ENSG00000176597	0.63	1.00x10 ⁻⁰³
G0S2	ENSG00000123689	0.99	1.00x10 ⁻⁰³
OR52K2	ENSG00000181963	0.73	1.10x10 ⁻⁰³
NFAM1	ENSG00000235568	0.6	1.10x10 ⁻⁰³
BCL6	ENSG00000113916	0.76	1.10x10 ⁻⁰³
CEACAM3	ENSG00000170956	0.63	1.10x10 ⁻⁰³
FRAT2	ENSG00000181274	0.56	1.10x10 ⁻⁰³
IL18RAP	ENSG00000115607	0.7	1.10x10 ⁻⁰³
TLR5	ENSG00000187554	0.72	1.10x10 ⁻⁰³
PEAK3	ENSG00000188305	0.66	1.10x10 ⁻⁰³
TYROBP	ENSG00000011600	0.55	1.10x10 ⁻⁰³
SORL1	ENSG00000137642	0.57	1.20x10 ⁻⁰³
SPRY2	ENSG00000136158	-0.64	1.20x10 ⁻⁰³
WDFY3	ENSG00000163625	0.66	1.20x10 ⁻⁰³
MYBPC3	ENSG00000134571	0.64	1.20x10 ⁻⁰³
HLX	ENSG00000136630	0.56	1.30x10 ⁻⁰³
CEBPD	ENSG00000221869	0.63	1.30x10 ⁻⁰³
TIMP2	ENSG00000035862	0.56	1.30x10 ⁻⁰³
ST3GAL4	ENSG00000110080	0.6	1.30x10 ⁻⁰³
TRIM9	ENSG00000100505	0.72	1.30x10 ⁻⁰³
ROPN1L	ENSG00000145491	0.64	1.40x10 ⁻⁰³
CDKN1C	ENSG00000129757	-0.97	1.40x10 ⁻⁰³
PTGS2	ENSG00000073756	0.55	1.50x10 ⁻⁰³
IL18R1	ENSG00000115604	0.81	1.50x10 ⁻⁰³
CLEC6A	ENSG00000205846	0.78	1.50x10 ⁻⁰³
SPACA6	ENSG00000182310	0.51	1.50x10 ⁻⁰³
ABCA1	ENSG00000165029	0.61	1.60x10 ⁻⁰³
FLT3	ENSG00000122025	0.78	1.60x10 ⁻⁰³
IL1RN	ENSG00000136689	0.7	1.6x10 ⁻⁰³
ADGRG3	ENSG00000182885	0.76	1.7x10 ⁻⁰³
TNFRSF10C	ENSG00000173535	0.68	1.7x10 ⁻⁰³
IFIT2	ENSG00000119922	0.86	1.7x10 ⁻⁰³
MAP3K20	ENSG00000091436	0.54	1.8x10 ⁻⁰³
SAMD9L	ENSG00000177409	0.52	1.8x10 ⁻⁰³
CCR1	ENSG00000163823	0.64	1.9x10 ⁻⁰³
MSR1	ENSG00000038945	0.84	1.9x10 ⁻⁰³
ST6GALNAC2	ENSG00000070731	0.7	1.9x10 ⁻⁰³
SERPINA1	ENSG00000197249	0.57	2.0x10 ⁻⁰³
ZNF366	ENSG00000178175	0.71	2.0x10 ⁻⁰³
IDO1	ENSG00000131203	1.08	2.0x10 ⁻⁰³
LTA4H	ENSG00000111144	0.51	2.0x10 ⁻⁰³
TPPP3	ENSG00000159713	-0.77	2.1x10 ⁻⁰³
GZMK	ENSG00000113088	-0.61	2.1x10 ⁻⁰³
PILRA	ENSG00000085514	0.5	2.1x10 ⁻⁰³
ALOX5AP	ENSG00000132965	0.57	2.2x10 ⁻⁰³
ASGR2	ENSG00000161944	0.62	2.2x10 ⁻⁰³
NCF2	ENSG00000116701	0.51	2.3x10 ⁻⁰³
CEBPB	ENSG00000172216	0.53	2.3x10 ⁻⁰³
CC2D2B	ENSG00000188649	0.5	2.3x10 ⁻⁰³
NCF4	ENSG00000100365	0.53	2.4x10 ⁻⁰³
CR1	ENSG00000203710	0.67	2.4x10 ⁻⁰³
SIRPA	ENSG00000198053	0.62	2.4x10 ⁻⁰³
TMEM272	ENSG00000281106	0.55	2.4x10 ⁻⁰³
SELP	ENSG00000174175	0.75	2.4x10 ⁻⁰³
GASK1B	ENSG00000164125	0.53	2.4x10 ⁻⁰³
RAB31	ENSG00000168461	0.52	2.4x10 ⁻⁰³
EPHB1	ENSG00000154928	0.71	2.5x10 ⁻⁰³
PCLAF	ENSG00000166803	-0.6	2.5x10 ⁻⁰³
S100A11	ENSG00000163191	0.52	2.5x10 ⁻⁰³
MARCKS	ENSG00000277443	0.53	2.6x10 ⁻⁰³
PADI2	ENSG00000117115	0.79	2.6x10 ⁻⁰³
H4C1	ENSG00000278637	-0.52	2.6x10 ⁻⁰³

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
PLD4	ENSG00000166428	-0.55	2.7x10 ⁻⁰³
MMP25	ENSG00000008516	0.69	2.7x10 ⁻⁰³
ZNF467	ENSG00000181444	0.57	2.7x10 ⁻⁰³
GLT1D1	ENSG00000151948	0.64	2.8x10 ⁻⁰³
H1-6	ENSG00000187475	0.72	2.8x10 ⁻⁰³
STEAP4	ENSG00000127954	0.66	2.8x10 ⁻⁰³
TPST1	ENSG00000169902	0.79	2.8x10 ⁻⁰³
H2AC4	ENSG00000278463	-0.56	2.9x10 ⁻⁰³
RASSF6	ENSG00000169435	-0.87	2.9x10 ⁻⁰³
CLU	ENSG00000120885	0.69	3.0x10 ⁻⁰³
APOBEC3A	ENSG00000128383	0.55	3.0x10 ⁻⁰³
PLSCR1	ENSG00000188313	0.55	3.0x10 ⁻⁰³
DGAT2	ENSG00000062282	0.71	3.1x10 ⁻⁰³
IER3	ENSG00000137331	0.53	3.1x10 ⁻⁰³
LY96	ENSG00000154589	0.61	3.2x10 ⁻⁰³
LGALS3BP	ENSG00000108679	0.59	3.2x10 ⁻⁰³
GRINA	ENSG00000178719	0.53	3.2x10 ⁻⁰³
IFIH1	ENSG00000115267	0.51	3.3x10 ⁻⁰³
FAM20A	ENSG00000108950	0.94	3.4x10 ⁻⁰³
IFTT3	ENSG00000119917	1	3.5x10 ⁻⁰³
ALPL	ENSG00000162551	0.89	3.7x10 ⁻⁰³
SLC2A3	ENSG00000059804	0.57	3.8x10 ⁻⁰³
MYO10	ENSG00000145555	0.93	3.8x10 ⁻⁰³
TMEM252	ENSG00000181778	0.7	3.8x10 ⁻⁰³
C10orf105	ENSG00000214688	0.66	4.0x10 ⁻⁰³
CENPW	ENSG00000203760	-0.57	4.1x10 ⁻⁰³
SIRPB1	ENSG00000101307	0.5	4.2x10 ⁻⁰³
RASGRP4	ENSG00000171777	0.53	4.3x10 ⁻⁰³
TRIM6	ENSG00000121236	0.75	4.3x10 ⁻⁰³
TREML1	ENSG00000161911	0.83	4.3x10 ⁻⁰³
BASP1	ENSG00000176788	0.73	4.4x10 ⁻⁰³
FAM111B	ENSG00000189057	-0.52	4.4x10 ⁻⁰³
PLSCR2	ENSG00000163746	0.54	4.5x10 ⁻⁰³
GNG10	ENSG00000242616	0.58	4.6x10 ⁻⁰³
IFI35	ENSG00000068079	0.52	4.6x10 ⁻⁰³
CLEC1A	ENSG00000150048	0.59	4.7x10 ⁻⁰³
RAB13	ENSG00000143545	0.56	4.9x10 ⁻⁰³
LILRB3	ENSG00000204577	0.6	5.1x10 ⁻⁰³
GPR84	ENSG00000139572	0.94	5.3x10 ⁻⁰³
ASAP2	ENSG00000151693	0.5	5.4x10 ⁻⁰³
NFE2	ENSG00000123405	0.56	5.5x10 ⁻⁰³
SLC22A4	ENSG00000197208	0.56	5.8x10 ⁻⁰³
LILRA6	ENSG00000244482	0.61	5.9x10 ⁻⁰³
LILRA4	ENSG00000239961	-0.63	5.9x10 ⁻⁰³
OSCAR	ENSG00000170909	0.52	6.1x10 ⁻⁰³
CFAP45	ENSG00000213085	0.66	6.1x10 ⁻⁰³
GAS6	ENSG00000183087	0.56	6.3x10 ⁻⁰³
COCH	ENSG00000100473	-0.7	6.3x10 ⁻⁰³
PF4	ENSG00000163737	0.71	6.5x10 ⁻⁰³
ANXA3	ENSG00000138772	0.81	6.8x10 ⁻⁰³
NLRC4	ENSG00000091106	0.55	6.9x10 ⁻⁰³
PROS1	ENSG00000184500	0.8	7.1x10 ⁻⁰³
ADGRE2	ENSG00000127507	0.51	7.3x10 ⁻⁰³
ITGA2B	ENSG00000005961	0.8	7.5x10 ⁻⁰³
PDCD1	ENSG00000188389	-0.62	7.6x10 ⁻⁰³
CMTM5	ENSG00000166091	0.72	7.7x10 ⁻⁰³
NTNG2	ENSG00000196358	0.55	7.7x10 ⁻⁰³
HES1	ENSG00000114315	-0.56	7.7x10 ⁻⁰³
FCAR	ENSG00000186431	0.6	7.9x10 ⁻⁰³
H2BC3	ENSG00000276410	-0.52	8.0x10 ⁻⁰³
OLIG1	ENSG00000184221	-0.55	8.0x10 ⁻⁰³
CCR9	ENSG00000173585	-0.63	8.1x10 ⁻⁰³

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
ENSG00000260861	ENSG00000260861	0.53	8.2x10 ⁻⁰³
APOBR	ENSG00000184730	0.54	8.2x10 ⁻⁰³
GCA	ENSG00000115271	0.57	8.3x10 ⁻⁰³
PCSK6	ENSG00000140479	0.63	8.3x10 ⁻⁰³
MAP1A	ENSG00000166963	0.61	8.4x10 ⁻⁰³
CD14	ENSG00000170458	0.52	8.4x10 ⁻⁰³
ACSS3	ENSG00000111058	0.56	8.7x10 ⁻⁰³
TGM2	ENSG00000198959	0.77	8.8x10 ⁻⁰³
PLOD2	ENSG00000152952	0.68	8.9x10 ⁻⁰³
ARHGAP6	ENSG00000047648	0.59	8.9x10 ⁻⁰³
EGF	ENSG00000138798	0.68	9.0x10 ⁻⁰³
ALOX12	ENSG00000108839	0.67	9.9x10 ⁻⁰³
ABLM3	ENSG00000173210	0.71	1.0x10 ⁻⁰²
ESAM	ENSG00000149564	0.56	1.0x10 ⁻⁰²
BMX	ENSG00000102010	0.78	1.1x10 ⁻⁰²
PROK2	ENSG00000163421	0.67	1.1x10 ⁻⁰²
GAS2L1	ENSG00000185340	0.54	1.1x10 ⁻⁰²
NRCAM	ENSG00000091129	-0.76	1.1x10 ⁻⁰²
CXCL8	ENSG00000169429	0.86	1.1x10 ⁻⁰²
CYP4F3	ENSG00000186529	0.5	1.2x10 ⁻⁰²
QPCT	ENSG00000115828	0.55	1.2x10 ⁻⁰²
KCNMA1	ENSG00000156113	0.71	1.2x10 ⁻⁰²
MPIG6B	ENSG00000204420	0.78	1.2x10 ⁻⁰²
SIGLEC9	ENSG00000129450	0.51	1.3x10 ⁻⁰²
S100A9	ENSG00000163220	0.66	1.3x10 ⁻⁰²
ITGB3	ENSG00000259207	0.74	1.3x10 ⁻⁰²
TREM1	ENSG00000124731	0.51	1.3x10 ⁻⁰²
LILRA5	ENSG00000187116	0.57	1.3x10 ⁻⁰²
SPX	ENSG00000134548	0.65	1.4x10 ⁻⁰²
MMRN1	ENSG00000138722	0.69	1.4x10 ⁻⁰²
C3AR1	ENSG00000171860	0.51	1.4x10 ⁻⁰²
NRGN	ENSG00000154146	0.57	1.5x10 ⁻⁰²
NEK2	ENSG00000117650	-0.55	1.5x10 ⁻⁰²
KLRC1	ENSG00000134545	-0.5	1.5x10 ⁻⁰²
VNN2	ENSG00000112303	0.51	1.5x10 ⁻⁰²
BEX3	ENSG00000166681	0.54	1.6x10 ⁻⁰²
TRPM2	ENSG00000142185	0.55	1.6x10 ⁻⁰²
KAZN	ENSG00000189337	0.61	1.6x10 ⁻⁰²
AATK	ENSG00000181409	0.53	1.7x10 ⁻⁰²
IRAK3	ENSG00000090376	0.5	1.7x10 ⁻⁰²
IRAG1	ENSG00000072952	0.51	1.7x10 ⁻⁰²
CACNA2D3	ENSG00000157445	-0.74	1.8x10 ⁻⁰²
AQP10	ENSG00000143595	0.73	1.8x10 ⁻⁰²
CTDSPL	ENSG00000144677	0.56	1.9x10 ⁻⁰²
PLBD1	ENSG00000121316	0.53	2.0x10 ⁻⁰²
JUP	ENSG00000173801	-0.56	2.0x10 ⁻⁰²
CXCL1	ENSG00000163739	0.52	2.0x10 ⁻⁰²
CALD1	ENSG00000122786	0.64	2.0x10 ⁻⁰²
GP9	ENSG00000169704	0.59	2.1x10 ⁻⁰²
PYGL	ENSG00000100504	0.54	2.1x10 ⁻⁰²
GRB10	ENSG00000106070	0.52	2.2x10 ⁻⁰²
CAV2	ENSG00000105971	0.64	2.3x10 ⁻⁰²
OPLAH	ENSG00000178814	0.62	2.3x10 ⁻⁰²
SPARC	ENSG00000113140	0.62	2.4x10 ⁻⁰²
ZNF80	ENSG00000174255	-0.54	2.4x10 ⁻⁰²
CLEC4D	ENSG00000166527	0.58	2.4x10 ⁻⁰²
ITGA7	ENSG00000135424	0.87	2.4x10 ⁻⁰²
STXBP1	ENSG00000136854	0.53	2.4x10 ⁻⁰²
ITGB5	ENSG00000082781	0.56	2.4x10 ⁻⁰²
NKX3-1	ENSG00000167034	0.56	2.6x10 ⁻⁰²
MME	ENSG00000196549	0.53	2.7x10 ⁻⁰²
CEACAM4	ENSG00000105352	0.51	2.7x10 ⁻⁰²

Continues on next page

Gene Name	Ensembl ID	log ₂ FC	BH-adjusted <i>p</i> -value
PF4V1	ENSG00000109272	0.59	2.7x10 ⁻⁰²
C4BPA	ENSG00000123838	0.97	2.8x10 ⁻⁰²
LRRN3	ENSG00000173114	-0.71	2.9x10 ⁻⁰²
MANSC1	ENSG00000111261	0.54	2.9x10 ⁻⁰²
VNN1	ENSG00000112299	0.67	3.0x10 ⁻⁰²
MGAM2	ENSG00000257743	0.62	3.0x10 ⁻⁰²
SH3BGRL2	ENSG00000198478	0.58	3.1x10 ⁻⁰²
SYNM	ENSG00000182253	0.54	3.1x10 ⁻⁰²
PRKAR2B	ENSG00000005249	0.51	3.1x10 ⁻⁰²
VIL1	ENSG00000127831	0.62	3.2x10 ⁻⁰²
MARCO	ENSG00000019169	0.54	3.2x10 ⁻⁰²
MGAM	ENSG00000257335	0.55	3.5x10 ⁻⁰²
SMPD3	ENSG00000103056	-0.52	3.6x10 ⁻⁰²
RAB6B	ENSG00000154917	0.69	3.8x10 ⁻⁰²
PPBP	ENSG00000163736	0.59	4.0x10 ⁻⁰²
IL1R2	ENSG00000115590	0.68	4.1x10 ⁻⁰²
GP6	ENSG00000088053	0.52	4.2x10 ⁻⁰²
SEC14L5	ENSG00000103184	0.62	4.2x10 ⁻⁰²
F13A1	ENSG00000124491	0.54	4.3x10 ⁻⁰²
HLA-DQB1	ENSG00000179344	0.57	4.7x10 ⁻⁰²
NSUN7	ENSG00000179299	0.52	4.9x10 ⁻⁰²

APPENDIX G

BEST AVAILABLE TREATMENT STUDY FOR MIS-C (BATS)

The Multi-System Inflammatory Syndrome in children (MIS-C), also known as Paediatric Inflammatory Multisystem Syndrome Temporally Associated with SARS-CoV-2 (PIMS-TS), was originally recognised during the first wave of the COVID-19 pandemic in April 2020. The first reports from hospitals in the United Kingdom described this disease as a life-threatening illness characterised by high fever, rash and red eyes with symptoms suggesting the involvement of multiple organs including abdominal pain, vomiting and diarrhoea, confusion or lethargy, and in severe cases, heart failure, multi-organ failure, and even death. Guidelines for the diagnosis and treatment of this new disease were published by the Royal College of Paediatrics and Child Health in the United Kingdom, the Centers for Disease Control and Prevention in the United States of America, and then by the World Health Organization (WHO). Since these first descriptions, the disorder has emerged as an important childhood illness occurring worldwide approximately 4 weeks after SARS-CoV-2 infection, which may have had minimal symptoms in the child.

Faced with a new and life-threatening childhood illness, paediatricians worldwide attempted to treat the disorder with drugs that had been shown to work in other inflammatory paediatric disorders. Because MIS-C closely resembles Kawasaki disease in its clinical features, treatment with intravenous immunoglobulin (which contains antibodies from multiple donors) was the most widely used treatment, often combined with high doses of steroids or other immunomodulatory drugs. Although treatments with intravenous immunoglobulin, steroids and other immunomodulator drugs were widely adopted and included in national guidelines on the treatment of MIS-C, there was no clear evidence as to which of the treatments was best or more effective or harmful.

Evidence from randomised clinical trials is needed to determine which of several treatment options is best. However, the speed with which the new disease emerged made the conduct of randomised controlled trials difficult, because such trials may take many months to set up and to obtain the necessary funding, ethical approvals, and participant numbers. In the absence of randomised controlled trials, the Best Available Treatment Study for MIS-C (BATS) was established by a team of paediatricians at Imperial College London. Paediatricians worldwide were invited to join an online study, collecting data on patients with suspected MIS-C under their care. The BATS researchers reasoned that paediatricians in different hospitals and countries would treat patients according to their “best guess” as to what treatment might work. Although the treatments were not randomised, they reasoned that if a large enough number of patients were studied, it would be possible to establish which treatments worked and which did not by comparing the outcome of patients treated with different regimes.

A database was created in REDCap and the study was launched on 14th June 2020, just a few weeks after the first recognition of MIS-C. It was enthusiastically adopted by paediatricians in 39 different countries.

Over 2000 patients were included in the study, with 680 children receiving initial treatment with

intravenous immunoglobulin, 487 receiving steroids, and 698 receiving a combination of steroids and immunoglobulin. The study compared the rate of progression or recovery in children treated with each of these three regimes. A statistical method was used to account for differences in factors such as age and severity which may have affected the outcome and treatment assignment. No significant difference between the three treatments was found in the rate of progression of illness or recovery from organ failure. There was a suggestion that patients treated with combined immunoglobulin and steroids had a more rapid rate of improvement but this effect was mostly seen in the less severely ill patients. Of particular importance was that patients treated with each of the three regimes had similarly low rates of coronary artery abnormalities and similarly high rates of recovery from intensive care.

MIS-C is likely to remain an ongoing paediatric problem worldwide as long as SARS-CoV-2 infection continues to circulate and cause repeated waves of illness. Although the frequency of the disorder has decreased in many countries, some countries such as China are now experiencing high rates of SARS-CoV-2 infection and are likely to experience an increase in cases of MIS-C. Newly emerging variants and ongoing waves of infection are likely to continue to cause the disease in children lacking previous immunity. The finding from the BATS study, that treatment with steroids is as effective as treatment with immunoglobulin or combined immunoglobulin and steroids is of considerable importance, because immunoglobulin is in short supply globally and is expensive, and therefore unaffordable in many settings. The results of the study are likely to affect recommendations on the treatment of MIS-C by the WHO and other international bodies.

My role in this project was to create and manage the database.

Publications arising from the work described in Appendix G:

McArdle A.J.*, Vito O.*, Patel H., Seaby E.G., Shah P., Wilson C., Broderick C., Nijman R., Tremoulet A.H., Munblit D., Ulloa-Gutierrez R., Carter M.J., De T., Hoggart C., Whittaker E., Herberg J.A., Kaforou M., Cunningham A.J., Levin M. for the BATS Consortium. Treatment of multisystem inflammatory syndrome in children. *the New England Journal of Medicine* 2021;385:11-22

Channon-Wells S.*, Vito O.*, McArdle A.J.*, Seaby E.G., Patel H., Shah P., Pazukhina E., Wilson C., Broderick C., D'Souza G., Keren I., Nijman R., Tremoulet A.H., Munblit D., Ulloa-Gutierrez R., Carter M.J., Ramnarayan P., De T., Hoggart C., Whittaker E., Herberg J.A., Kaforou M., Cunningham A.J., Blyuss O., Levin M. on behalf of the BATS Consortium. Immunoglobulin, glucocorticoid, or combination therapy for multisystem inflammatory syndrome in children: a propensity-weighted cohort study. *The Lancet Rheumatology* 2023; 2665-9913

Ouldali N.*, Son M.B.F.*, McArdle A.J.*, Vito O., Vaugon E., Belot A. Leblanc C., Murray N.L., Patel M.M., Levin M., Randolph A.G., Angoulvant F for the BATS consortium, the Overcoming COVID-19 Investigators, and the French Covid-19 Pediatric Inflammation Consortium and Pandor study group. Immunomodulatory therapy for MIS-C: An Individual Patient Meta-Analysis. *Submitted to and accepted for publication by PEDIATRICS*

APPENDIX H

COPYRIGHT PERMISSION INFORMATION

Table H.1: Copyright permissions information

Name of work	Source of work	Copyright holder and contact	Permission requested on	Permission yes/no	Permission note
Figure 1.1: Estimated TB incidence in 2021 for countries with at least 100,000 incident cases	World Health Organization. Global tuberculosis report 2022. Report 2022 [2]	© World Health Organization 2022	NA	Yes	Permission not needed: Licensed with a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO licence (CC BY-NC-SA 3.0 IGO; https://creativecommons.org/licenses/by-nc-sa/3.0/igo)
Figure 1.3: Time from infection to T cell response: a comparison between other pathogens and <i>Mtb</i>	K. B. Urdahl, S. Shafiani, and J. D. Ernst. "Initiation and regulation of T-cell responses in tuberculosis". In: Mucosal Immunol 4.3 (2011), pp. 288–93. ISSN: 1935-3456. DOI: 10.1038/mi.2011.10. URL: https://www.ncbi.nlm.nih.gov/pubmed/21451503 [28]	© 2011 Society for Mucosal Immunology	26/04/2023	Yes	Permission obtained
Figure 1.4: The host immune response to <i>Mtb</i> in the lungs	A. O'Garra et al. "The immune response in tuberculosis". In: Annu Rev Immunol 31 (2013), pp. 475–527. ISSN: 1545-3278. DOI: 10.1146/annurev-immunol-032712-095939. URL: https://www.ncbi.nlm.nih.gov/pubmed/23516984 [32]	© 2013 by Annual Reviews	26/04/2023	Yes	Permission obtained

Continues on next page

Name of work	Source of work	Copyright holder and contact	Permission requested on	Permission yes/no	Permission note
Figure 1.5: The dynamic spectrum from <i>Mtb</i> exposure to tuberculosis disease in children	J. A. Seddon et al. "The evolving research agenda for paediatric tuberculosis infection". In: Lancet Infect Dis 19.9 (2019), e322–e329. ISSN: 1474-4457. URL: https://www.ncbi.nlm.nih.gov/pubmed/31221543 [83]	© 2019 Elsevier Ltd	26/04/2023	Yes	Permission obtained
Figure 1.10: Multiple omics data types and approaches for disease diagnostic and characterisation	Y. Hasin, M. Seldin, and A. Lusis. "Multi-omics approaches to disease". In: Genome Biol 18.1 (2017), p. 83. ISSN: 1474-760X. URL: https://www.ncbi.nlm.nih.gov/pubmed/28476144 [109]	© The Author(s)	NA	Yes	Permission not needed: Licensed with a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/)
Figure 1.12: Overview of the steps involved in DNA microarray experiments	P. A. Bryant et al. "Chips with everything: DNA microarrays in infectious diseases". In: Lancet Infect Dis 4.2 (2004), pp. 100–11. ISSN: 1473-3099. URL: https://www.ncbi.nlm.nih.gov/pubmed/14871635 [129]	© The Lancet	26/04/2023	Yes	Permission obtained
Figure 1.13: Overview of the steps involved in a RNA-Seq experiment	K. R. Kukurba and S. B. Montgomery. "RNA Sequencing and Analysis". In: Cold Spring Harb Protoc 2015.11 (2015), pp. 951–69. ISSN: 1559-6095. URL: https://www.ncbi.nlm.nih.gov/pubmed/25870306 [131]	© 2015 Cold Spring Harbor Laboratory Press	26/04/2023		Permission asked

Continues on next page

Name of work	Source of work	Copyright holder and contact	Permission requested on	Permission yes/no	Permission note
Figure 1.14: Overview of the Luminex® assay	T. Koudriavtseva et al. "Coagulation/ Complement Activation and Cerebral Hypoperfusion in Relapsing-Remitting Multiple Sclerosis". In: Front Immunol 11 (2020), p. 548604. ISSN: 1664-3224. URL: https://www.ncbi.nlm.nih.gov/pubmed/33193314 [170]	© 2020 Koudriavtseva, Stefanile, Fiorelli, Lapucci, Lorenzano, Zannino, Conti, D'Agosto, Pimpinelli, Di Domenico, Mandoj, Giannarelli, Donzelli, Blandino, Salvetti and Inglese	NA	Yes	Permission not needed: open-access article distributed under the terms of the Creative Commons Attribution License (CC BY)
Figure 1.15: Conceptual model of multi-omics and human disease	Y. V. Sun and Y. J. Hu. "Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases". In: Adv Genet 93 (2016), pp. 147–90. ISSN: 0065-2660. URL: https://www.ncbi.nlm.nih.gov/pubmed/26915271 [196]	© 2016 Elsevier Inc	26/04/2023	Yes	Permission obtained
Figure 1.17: Overview of multi-omics data integration tools	F. McGee et al. "The State of the Art in Multilayer Network Visualization". In: Computer Graphics Forum 38.6 (2019), pp. 125–149. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13610 [204]	©2019 The Authors	NA	Yes	Permission not needed: open-access article distributed under the terms of the Creative Commons Attribution License

Continues on next page

Name of work	Source of work	Copyright holder and contact	Permission requested on	Permission yes/no	Permission note
Figure 1.17: Overview of multi-omics data integration tools	I. Subramanian et al. "Multi-omics Data Integration, Interpretation, and Its Application". In: Bioinform Biol Insights 14 (2020), p. 1177932219899051. ISSN: 1177-9322. URL: https://www.ncbi.nlm.nih.gov/pubmed/32076369 [204]	©2020 The Authors	NA	Yes	Permission not needed: open-access article distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/)
Figure 2.6: Theoretical plot of a PCR output curve	Y. Cao et al. "Digital PCR as an Emerging Tool for Monitoring of Microbial Biodegradation". In: Molecules 25.3 (2020). ISSN: 1420-3049. URL: https://www.ncbi.nlm.nih.gov/pubmed/32041334 [218]	©2020 by the authors	NA	Yes	Permission not needed: open-access article distributed under the terms of the Creative Commons Attribution(CC BY) license (http://creativecommons.org/licenses/by/4.0/)