

**Applying multi-omics analyses to investigate
Mycobacterium tuberculosis infection: a dual perspective**

Ashleigh Cheyne

A thesis submitted for the degree of Doctor of Philosophy

April 2022

Department of Life Sciences

Imperial College London

Declaration of originality

I, Ashleigh Cheyne, declare that the work performed in this thesis is my own unless otherwise acknowledged.

Copyright Declaration

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.

Acknowledgements

Before I thank the many individuals who have shared in my PhD experience over the years, I want to highlight how challenging this experience has been. Several academic challenges and personal tragedies have altered my perspective of science, and life in general. In addition, the COVID-19 pandemic changed my PhD completely and has had a profound impact on my experiences during my studies. Without the support of the individuals described later, my experience would have been very different.

First and foremost I would like to thank my supervisors Dr Gerald-Larrouy-Maumus, Prof Michael Levin, Dr Myrsini Kaforou, and Dr Sandra Newton both for the unique opportunity of working together and for their continuous support throughout my PhD. They have each taught me different aspects of research, for which I am extremely grateful.

Further thanks go to my colleagues and friends in both my labs, the CMBI, and at St Mary's for their company, encouragement, and advice over the years. Thanks to the NIH project team for their amazing insight in running such a large and collaborative international study, with special thanks to Ivonne, Ortensia, and Ivana for the amazing times we shared in South Africa together. Additional thanks to the compassionate support from Michael and Sonia, for the scientific, emotional, and social support since my Masters. Additional thanks go to all my students for their hard work but in particular Petra, who has shown incredible endurance and continuous enthusiasm for her project.

Special thanks to Nora who has helped me since my Undergraduate and has shared in my vision of helping people through infectious disease research. A huge thanks to Millie who has always been there for me throughout my PhD, without your friendship I'm not sure I would have made it! Also, a big thanks to Signe who I've missed so much since your move to Oxford but who continued to support me from afar.

Massive thanks goes to the kempo team, Stuart, Aled, Matteo, Mappeo, Jackie, Sergey, Michal, and everyone else I trained with in the dojo. Your exercises, encouragement, and overall loveliness would have made my experience through my PhD significantly less enjoyable. Your comradery will never be forgotten. Extra special thanks to my grading partner and PhD buddy Matteo. You have supported me through the literal worst and I can never thank you enough.

Of course a huge thanks to my family, mum and Ryan, who have no idea what I'm doing but have supported me unconditionally nonetheless! You have encouraged me both before and during my PhD and I will always be grateful. Good luck reading my thesis guys!!

Finally, so many thanks go to my partner Timur. I don't have the words to express how you've helped me through my time at Imperial, just know that I couldn't have done it without you.

Abstract

Mycobacterium tuberculosis (Mtb) is the causative agent of tuberculosis (TB) and is the second highest cause of death from an infectious disease worldwide. This is despite years of research to understand and tackle TB, leading to a vaccine, diagnostic tests, and treatments. Despite these interventions, several aspects of the disease are poorly understood or neglected in this field of research. A greater understanding of the infection in these areas will enable the development of more effective prevention and treatment strategies. This thesis addressed these concerns through using a multi-targeted approach investigating several aspects of Mtb infection which are understudied.

First, the characterisation of paediatric TB was performed through differential gene expression analysis using a cohort of paediatric TB patients from South Africa. This cohort was then used to validate previously existing host transcriptomic biomarker signatures for TB to identify if the current adult or paediatric signatures meet the guidelines set out by the World Health Organisation (WHO) for development of new non-sputum based diagnostic tests for TB. None of these signatures met these guidelines using their original models. Optimisation of the signature models increased the performance of some of these signatures to reach the WHO guidelines, suggesting that there is potential for certain signatures to be developed into diagnostic tests in the future. Second, the differences between Mtb lineages was investigated through the integration of multi-omics datasets. Differences in non-synonymous single nucleotide polymorphisms (nsSNPs) and metabolites from 14 Mtb strains corresponding to 3 lineages were identified. The differences in metabolites was then correlated with nsSNP blocks (groups of correlated nsSNPs). This identified a novel negative correlation between ergothioneine, a metabolite involved in redox pathways, and a nsSNP block only present in Mtb strains from lineage 1. Finally, host-pathogen interactions were investigated through dual RNA-seq of whole blood infected with BCG *lux*, a surrogate strain for Mtb. Dual RNA-seq is a relatively new technique which has never been formed from whole blood samples infected with Mtb. A workflow to combine the previously defined whole blood assay with dual RNA-seq was developed and optimised. A pilot dual RNA-seq experiment was performed to identify the number of reads, and the number of genes, obtained from both the host and bacteria through this workflow. Further characterisation of the whole blood assay was also performed through the evaluation of live immune cell proportions across the time-course of the assay and through comparison of the transcriptome of Mtb infected whole blood with the transcriptomes of adult and paediatric TB patients. Overall, this thesis provides new insights into the complex biology of Tuberculosis.

Contents

Declaration of originality	2
Acknowledgements	3
Abstract	4
List of Tables	8
List of Figures	10
Abbreviations	13
Chapter 1: Introduction	15
1.1 <i>Mycobacterium tuberculosis</i> and tuberculosis infection	15
1.1.1 Mtb pathogenesis	15
1.1.2 Mtb evolution	17
1.1.3 Mtb diagnosis	20
1.1.4 Host infection and immune response	22
1.1.5 Modelling TB disease through host-pathogen interactions	30
1.2 Omics technologies and data analyses in tuberculosis research	39
1.2.1 Genomics	40
1.2.2 Transcriptomics	44
1.2.3 Host whole blood transcriptomic biomarker signatures for TB	53
1.2.4 Metabolomics	58
1.2.5 Bioinformatics methods	65
1.2.6 The development of multi-omics research	71
1.2.7 Multi-omics integration methods	71
1.2.8 Multi-omics integration in infectious diseases and TB	74
1.3 PhD Aims	78
1.3.1 Research Question	78
1.3.2 Aims	78
Chapter 2: Materials and Methods	79
2.1 Data availability across chapters	79
2.2 Methods used to investigate Mtb lineages	80
2.2.1 Mycobacterial strains	80
2.2.2 Mtb culture and growth	83
2.2.3 Metabolomics analysis of Mtb lineages	83
2.2.4 Genomics	85
2.2.5 Metabolomics of Mtb lineages and integration analysis	86
2.3 Methods used to discovery host biomarker signatures for TB	92

2.3.1 Study design and RNA-seq for biomarkers analysis	92
2.3.2 Transcriptomics for <i>in vivo</i> gene expression analysis and biomarker discovery.....	93
2.4 Methods used to investigate host-pathogen interactions in TB	100
2.4.1 Mycobacterial culture and growth.....	100
2.4.2 Human blood donors	100
2.4.3 Phytohemagglutinin (PHA) and Purified Protein Derivative (PPD) stimulation of whole blood stimulation.....	100
2.4.4 Infection of whole blood with BCG <i>lux</i>	101
2.4.5 Enzyme-linked immunosorbent assay for cytokine analysis of WBA or PPD/PHA assay supernatant	102
2.4.6 Flow cytometry analysis for WBA cell proportion analysis	104
2.4.7 RNA extraction optimisation and dual RNA-seq	109
2.4.8 Transcriptomics for dual RNA-seq analysis	111
2.4.9 Transcriptomics for comparison of <i>in vitro</i> and <i>in vivo</i> Mtb infection models.....	112
Chapter 3: The host: Host blood RNA signatures for paediatric TB	115
3.1 Background.....	115
3.2 Results	122
3.2.1 Cohort.....	122
3.2.2 Pre-processing and data exploration	124
3.2.3 Differential gene expression analysis.....	128
3.2.4 Pathway analysis of significantly differentially expressed genes	136
3.2.5 Performance of published signatures on the paediatric cohort.....	144
3.3 Discussion	161
3.4 Appendix	166
Chapter 4: The bacterium: Multi -omics integration of genomic and metabolomic level information in TB lineages.....	167
4.1 Background.....	167
4.2 Results	171
4.2.1 Creating a multi-omics integration approach for genomic and metabolomic data	171
4.2.2 SNP calling analysis and genomic-based clustering of Mtb strains	175
4.2.3 Metabolomics analysis and metabolomic-based clustering of Mtb strains	178
4.2.4 Integration of SNP and metabolite datasets	184
4.3 Discussion	204
Chapter 5: Analysing host-pathogen interactions in tuberculosis infection through applying dual RNA-seq to the whole blood infection assay	208
5.1 Background.....	208
5.2 Results	211

5.2.1 Optimisation of BCG <i>lux</i> growth	211
5.2.2 Evaluation of PPD/PHA response in donors	213
5.2.3 Validation of BCG <i>lux</i> growth in whole blood and cytokine analysis	214
5.2.4 Optimisation of flow cytometry analysis of whole blood infected with BCG <i>lux</i>	222
5.2.5 Optimisation of RNA extraction procedures from whole blood infected with BCG <i>lux</i>	229
5.2.6 Pilot dual RNA-seq analysis of whole blood infected with BCG <i>lux</i>	235
5.2.7 Main dual RNA-seq experiment.....	264
5.3 Discussion.....	268
5.4 Appendix 1: FASTQC report.....	272
5.4 Appendix 2: Comparing host-pathogen interactions in an <i>in vitro</i> whole blood infection assay with <i>in vivo</i> Tuberculosis infection	274
5.4.1 Background	274
5.4.2 Results – Comparison of cohort datasets	276
5.4.3 Results – Pre-processing of in vitro and in vivo datasets	278
5.4.4 Results – Differential gene expression analysis of the individual datasets	280
5.4.5 Results – Comparison of gene expression signatures between in vitro and in vivo datasets	283
5.4.6 Results – Concordance and discordance analysis of the in vitro and adult in vivo datasets	286
5.4.7 Results – Concordance and discordance analysis of the in vitro and paediatric in vivo datasets.....	288
5.4.8 Results – Concordance and discordance analysis of in silico cell populations from in vitro and in vivo datasets	290
5.4 Discussion	294
5.4 Appendix 3: GO term analysis	296
Chapter 6: General Discussion	302
6.1 TB research requires an interdisciplinary outlook.....	302
6.2 The shortcomings and future possibilities of the WBA.....	306
6.3 Conclusions.....	308
Chapter 7: Bibliography	309

List of Tables

Table 1.1: The advantages and disadvantages of different Mtb infection models.	33
Table 1.2: WHO guidelines for TPPs for triage and confirmatory non-sputum based diagnostic tests for TB	54
Table 1.3: Host blood transcriptomic biomarker signatures for TB.	56
Table 2.1: Mycobacterial strains used for Mtb lineage analyses.	81
Table 2.2: Modified BioCyc pathways.	89
Table 2.3: Previously published host transcriptomic biomarker signatures for TB.	98
Table 2.4: Mycobacterial strains used for host-pathogen interaction analyses.	100
Table 2.5: Laboratory-prepared ELISA reagents.	102
Table 2.6: Fluorochrome panel used for flow cytometry analysis.	106
Table 2.7: Serial dilutions used for antibody titrations	107
Table 3.1: Host blood transcriptomic signatures for paediatric TB.....	117
Table 3.2: Signature performances of paediatric biomarker signatures.	120
Table 3.3: Patient cohort characteristics.	123
Table 3.4: Top 50 differentially expressed genes which differentiate between TB vs OD patients. NA = not applicable.....	130
Table 3.5: Top 50 differentially expressed genes which differentiate between TB vs LTBI patients. NA = not applicable.....	134
Table 3.6: Top 50 significant pathways when comparing TB and OD groups from Reactome pathway analysis.	139
Table 3.7: Top 50 significant pathways when comparing TB and LTBI groups from Reactome pathway analysis.	142
Table 3.8: The performance of published biomarker signatures using their original models in discriminating between TB and LTBI patients.....	147
Table 3.9: The performance of published biomarker signatures using their original models in discriminating between TB and OD patients.	149
Table 3.10: The performance of published biomarker signatures using optimised classification models in discriminating between TB and LTBI patients.....	152
Table 3.11: The performance of published biomarker signatures using optimised classification models in discriminating between TB and OD patients.	157
Table 4.2: nsSNP blocks containing nsSNPs which affect more than 1 gene.....	185
Table 4.3: Representation score for genes with nsSNPs present in the rs1849 nsSNP block.....	188
Table 4.4: Representation score for the top 50 genes with nsSNPs present in the rs6112 nsSNP block.	191
Table 4.5: Representation score for the genes with nsSNPs present in the rs15117 nsSNP block....	193
Table 4.6: Representation score for genes with nsSNPs present in the rs25137 nsSNP block.....	195
Table 5.1: RLU/CFU ratios of BCG lux growth curve.....	211
Table 5.2: Separation indices for flow cytometry antibodies.	223
Table 5.3: An example of baseline RNA concentrations from 3 uninfected WBA samples at 0 hrs. .	229
Table 5.4: The effect of DNase and clean-up treatment steps on RNA quantity and quality in uninfected WBA samples.....	231
Table 5.5: RIN measurements of uninfected WBA samples before and after DNase treatment.	232
Table 5.6: The effect of bead-beating on the concentration and quality of RNA from infected WBA samples.....	233
Table 5.7: RNA quality and quantity of BCG lux WBA samples prior to sequencing.	236

Table 5.8: HBB and HBA counts and percentages in host samples.	241
Table 5.9: Percentage of mapped rRNA reads in pilot dual RNA-seq samples. NBB = non bead-beat samples; BB = bead-beat samples.	242
Table 5.10: Summary of mapped reads to host and bacteria.	243
Table 5.11: Summary of host differential gene expression results of the pilot dual RNA-seq samples.	248
Table 5.12: Top 50 differentially expressed host genes for 0 vs 11 hrs time-points in the pilot dual RNA-seq experiment.	250
Table 5.13: Top 50 differentially expressed host genes for 11 vs 24 hrs time-points in the pilot dual RNA-seq experiment.	252
Table 5.14: Top 50 differentially expressed host genes for 0 vs 24 hrs time-points in the pilot dual RNA-seq experiment.	254
Table 5.15: Summary of bacterial differential gene expression results of the pilot dual RNA-seq bead-beat samples.	256
Table 5.16: Top 50 up and down-regulated BCG <i>lux</i> genes at 0 vs 11 hrs according to log ₂ fold change.	258
Table 5.17: Top 50 up and down-regulated BCG <i>lux</i> genes at 11 vs 24 hrs according to log ₂ fold change.	260
Table 5.18: Top 50 up and down-regulated BCG <i>lux</i> genes at 0 vs 24 hrs according to log ₂ fold change.	262
Table 5.19: RLU:CFU ratio calculation to determine BCG <i>lux</i> MOI used for inoculation of the WBA.	264
Table S5.4.4: Significantly differentially expressed genes in the <i>in vitro</i> and <i>in vivo</i> datasets	280
Table S5.4.11: GO term analysis for upregulated terms in the <i>in vitro</i> WBA dataset.	296
Table S5.4.12: GO term analysis for downregulated terms in the <i>in vitro</i> WBA dataset	298
Table S5.4.13: GO term analysis for paediatric <i>in vivo</i> dataset	300
Table S5.4.14: GO term analysis for adult <i>in vivo</i> dataset	301

List of Figures

Figure 1.1: An overview of TB pathogenesis.....	16
Figure 1.2: Phylogenetic and geographical distribution of MTBC	17
Figure 1.3: Proposed evolutionary divergence of MTBC	19
Figure 1.4: Defence mechanisms of the host macrophage and avoidance and suppression by Mtb..	23
Figure 1.5: Heterogeneity in the host immune response to TB	31
Figure 1.6: The lineage and composition of immune cells in host whole blood.....	36
Figure 1.7: Major metabolic pathway schematic in Mtb.....	60
Figure 1.8: Hypothesised changes in metabolism of ICL-deficient Mtb strains	62
Figure 1.9: Regression models.	67
Figure 1.10: Classification methods.....	68
Figure 1.11: Similarity Network Fusion methodology	73
Figure 2.1: Bioinformatic workflow for variant calling analysis.	87
Figure 2.2: Bioinformatics pipeline for host transcriptomics gene expression and biomarker analyses.	94
Figure 2.3: Example WBA experiment design. Infected whole blood (with BCG lux) and uninfected whole blood are incubated over time, in this case for 0, 24, and 48 hrs after infection.	101
Figure 2.4: Example gating strategy of the CD3 antibody marker (PE-Dazzle594) at a 1 in 28 dilution in uninfected whole blood at 0 hours.	107
Figure 2.5: Example gating strategies for flow cytometry experiments with and without antibody staining.	109
Figure 2.6: bioinformatics workflow for dual RNA-seq analysis.	111
Figure 3.1: Visualisation of dataset before and after normalisation.....	125
Figure 3.2: Visualisation of the effect of batches on the sample cohort comparing TB and OD patients	126
Figure 3.3: Visualisation of the effect of batches on the sample cohort comparing TB and LTBI patients.....	127
Figure 3.4: Identification of differentially expressed genes in TB compared to OD patients shows an upregulation of genes.....	129
Figure 3.5: Identification of differentially expressed genes in TB compared to LTBI patients shows an upregulation of genes.....	133
Figure 3.6: Pathway analysis of TB vs OD patients using Reactome	138
Figure 3.7: Pathway analysis of TB vs latent TB patients using Reactome	141
Figure 3.8: Literature and bioinformatic workflow of the validation of published biomarker signatures for TB on an independent paediatric cohort.....	145
Figure 3.9: Performance of published host transcriptomic signatures using their original classification models to diagnose TB from LTB.....	146
Figure 3.10: Performance of published host transcriptomic signatures using their original classification models to diagnose TB from OD.....	148
Figure 3.11: Performance of published host transcriptomic signatures using optimised classification models to diagnose TB from LTB.....	151
Figure 3.12: Relationship between signature size and sensitivity and specificity after stratification towards the minimum or maximum WHO guidelines to diagnose TB from LTB.	154
Figure 3.13: Performance of published host transcriptomic signatures using optimised classification models to diagnose TB from OD.	156

Figure 3.14: Relationship between signature size and sensitivity and specificity after stratification towards the minimum or maximum WHO guidelines to diagnose TB from OD.	159
Figure 3.15: Relationship between cohort country of origin and signature performance.	160
Figure S3.4: Heatmap of common biomarker signature genes. Signatures filtered for those with genes found in at least one other signature. Ordered by genes found in most signatures. Black boxes indicate gene is present in that signature and white boxes indicate gene is absent in that signature.	166
Figure 4.1: Genomic and metabolomic data integration workflow.	174
Figure 4.2: nsSNPs cluster Mtb strains by lineage and functional pathway analysis suggests major metabolic remodelling.	177
Figure 4.3: All metabolites cluster Mtb strains by their lineage.	179
Figure 4.4: Differentially abundant metabolites cluster Mtb strains by their lineage.	180
Figure 4.5: Heatmap showing log2 fold changes for the differentially abundant targeted metabolites from at least one strain compared to the H37Rv reference strain.	182
Figure 4.6: DIABLO integration analysis identifies nsSNP blocks which correlate with significantly differentially abundant untargeted metabolites.	200
Figure 4.7: Correlation analysis identifies nsSNP block correlations with differentially abundant targeted metabolites.	202
Figure 4.8: Correlation analysis identifies nsSNP block correlations with ergothioneine. A).....	203
Figure 5.1: BCG lux growth in different growth media.....	212
Figure 5.2: IFN γ production measured in whole blood from BCG vaccinated and unvaccinated donors upon stimulation with PPD, PHA, or RPMI.....	213
Figure 5.3: WBA using BCG vaccinated and unvaccinated donors.....	215
Figure 5.4: The effect of different whole blood volumes and MOI on BCG lux growth in the WBA over 96 hours of infection.	217
Figure 5.6: The effect of different blood volume and MOI conditions on luminescence ratios from the BCG lux WBA over 96 hours.	218
Figure 5.7: The effect of different sample volumes on the cytokine profile of whole blood infected with BCG lux at a MOI of 1:1 over 96 hours.....	220
Figure 5.8: The effect of different MOIs on the cytokine profile of 2 ml of whole blood infected with BCG lux over 96 hours.	221
Figure 5.9: Example histograms of the antibody titration of the CD15 fluorophore in whole blood.	224
Figure 5.11: Lymphocyte cell population analysis of whole blood upon infection of BCG lux at different MOIs over 48 hours.....	226
Figure 5.10: Granulocyte cell population analysis of whole blood upon infection of BCG lux at different MOIs over 48 hours.....	226
Figure 5.12: Monocyte cell population analysis of whole blood upon infection of BCG lux at different MOIs over 48 hours.	228
Figure 5.13: Pilot dual RNA-seq experimental design.	236
Figure 5.14: Granulocyte cell population analysis upon infection of whole blood with BCG lux for the pilot dual RNA-seq experiment.	238
Figure 5.15: Lymphocyte cell population analysis upon infection of whole blood with BCG lux for the pilot dual RNA-seq experiment.	238
Figure 5.16: Monocyte cell population analysis upon infection of whole blood with BCG lux for the pilot dual RNA-seq experiment.	239
Figure 5.17: Percentage of HBB gene counts in pilot dual RNA-seq samples.....	241
Figure 5.18: Proportion of reads mapping to host transcripts in bead-beat and non-bead-beat samples.	244

Figure 5.19: Proportion of reads mapping to bacterial transcripts in bead-beat and non-bead-beat samples.	245
Figure 5.20: PCA plots of host and bacterial transcripts from the pilot dual RNA-seq experiment after r-log normalisation.	247
Figure 5.21: Innate immune cell population composition upon whole blood infection with BCG lux for the main dual RNA-seq experiment.	266
Figure 5.22: Adaptive immune cell population composition upon whole blood infection with BCG lux for the main dual RNA-seq experiment.	267
Figure S5.4.2: Concordance and discordance analysis study design.	277
Figure S5.4.3: PCA of in vitro and in vivo datasets after pre-processing.	279
Figure S5.4.4: Gene expression profiles of in vitro and in vivo datasets.	281
Figure S5.4.5: Disco plots of concordant and discordant genes across datasets.	284
Figure S5.4.6: Concordant and discordant gene module counts.	285
Figure S5.4.7: Dot plots of summarised gene modules from discordance and concordance analysis comparing in vitro with the adult in vivo dataset.	287
Figure S5.4.8: Dot plots of summarised gene modules from discordance and concordance analysis comparing in vitro with the paediatric in vivo dataset.	289
Figure S5.4.9: In silico flow cytometry analysis of innate immune cell populations in in vitro and in vivo datasets.	291
Figure S5.4.10: In silico flow cytometry analysis of adaptive immune cell populations in in vitro and in vivo datasets.	293

Abbreviations

ADC	Albumin Dextrose Catalase	GO	gene ontology
Ag	antigen	GSEA	gene set enrichment analysis
AM	alveolar macrophage	GWAS	genome-wide association studies
ANOVA	analysis of variance	HIV	Human immunodeficiency virus
ATP	adenosine 5-triphosphate	HPLC-MS	High-performance liquid chromatography mass-spectrometry
AUC	area under the ROC curve	ICL	Isocitrate lyase
BAL	Bronchoalveolar lavage	ICU	intensive care unit
BL	same as above so pick and delete one	IFNs	interferons
BMDM	bone marrow-derived macrophages	IM	interstitial macrophages
BCG	Bacillus Calmette-Guérin	iNOS	inducible nitric oxide synthase
BWA	Burrows-Wheeler Alignment	IRF	interferon regulatory factors
CCA	canonical correspondence analysis	LAM	lipoarabinomannan
cDNA	complementary DNA	LAMP	Loop-mediated amplification test
CFU	colony forming units	LASSO	least absolute shrinkage and selection operator
CI	confidence intervals	LC/MS	liquid chromatography/mass spectrometry
CLR	C-type lectin receptors	lnRNA	long noncoding RNA
DCs	Dendritic cells	LTBI	Latent TB infection
dc-RT MLPA	Dual-colour-Reverse-Transcriptase-Multiplex-Ligation-dependent-Probe-Amplification	LTB	Latent TB infection
DGNT		MAIT	mucosal-associated invariant T cells
DR	direct repeat	MDR	multi-drug resistance
ELISA	enzyme-linked immunosorbent assay	MHC	major histocompatibility complex
EPTB	extrapulmonary TB	miRNA	micro RNA
ER	endoplasmic reticulum	MOI	multiplicity of infection
FACS	fluorescence activated cell sorting	mRNA	messenger RNA
FADH2	flavin dinucleotide	MSH	mycothiol
FBA	hydroquinone form	MSMD	Mendelian susceptibility to mycobacterial disease
FcyR	flux balance analysis	Mtb	<i>Mycobacterium tuberculosis</i>
FMO	Fc receptor	MTBC	<i>Mycobacterium tuberculosis</i> complex
FS-PLS	fluorescence minus one		
GC/MS	forward-selection partial-least squares		
	gas chromatography/mass spectrometry		

NAATs	nucleic acid amplification tests	RF	random forest
NADH	nicotinamide adenine dinucleotide	RIN	RNA integrity
Ndk	nucleoside diphosphate kinase	RLU	relative light units
NMR	nuclear magnetic resonance spectroscopy	ROC	Receiver Operating Characteristic curve
NO	Nitric oxide	ROS	reactive oxygen species
NOX	NADPH oxidase	RPMI	Roswell Park Memorial Institute
nsSNPs	non-synonymous single nucleotide polymorphisms	rRNA	ribosomal RNA
NTM	nontuberculous mycobacteria	SDE	significantly differentially expressed genes
OD	other diseases	SI	separation index
OPD	o-phenylenediamine-di-hydrochloride	SNF	Similarity network fusion
OPLS	orthogonal partial least squares	SNPs	Single nucleotide polymorphisms
OPLS-DA	orthogonal partial least squares discriminant analysis	sRNA	small noncoding RNA
OXPHOS	oxidative phosphorylation	STAR	Spliced Transcript Alignment to a Reference
PAMPS	pattern-associated molecular patterns	SVM	support vector machine
PBMCs	peripheral blood mononuclear cells	TACO	tryptophan aspartate-containing coat protein
PCA	principal component analysis	TB	tuberculosis
PCR	Polymerase chain reaction	TCA	tricarboxylic acid cycle
PE		TfR	transferring receptor
PEP	phosphoenolpyruvate	TLR	Toll-like receptors
PHA	Phytohemagglutinin	TNFα	tumour necrosis factor alpha
PI	phosphatidylinositol	TPPs	target patient profiles
PIP3	phosphatidylinositol (3,4,5)-trisphosphate	WBA	whole blood assay
PLS-DA	Partial Least-Squares Discriminant Analysis	WGCNA	weighted-gene correlation analysis
PPD	Purified Protein Derivative	XDR	extremely drug resistant
PRRs	pattern recognition receptors	2-DG	2-deoxy-D-glucose
PtpA	serine/threonine-protein phosphatase 2A		
RD	regions of difference		

Chapter 1: Introduction

1.1 *Mycobacterium tuberculosis* and tuberculosis infection

Prior to the COVID-19 pandemic, Tuberculosis (TB) was the highest cause of death from an infectious disease worldwide (WHO, 2020). In 2019, 10 million people became infected with TB and 1.4 million people died from the disease. In addition, a quarter of the world's population is estimated to be infected with *Mycobacterium tuberculosis* (Mtb), the bacteria which causes TB, although the disease is thought to be in a latent state in the majority (90%) of individuals. TB has remained as the top causes of death from an infectious disease since 2015 and, although rates are lowering each year, the current progress in TB is not on track to meet the World Health Organisation (WHO) and United Nations' targets by 2035 as part of the End TB strategy (WHO, 2014). In addition, the COVID-19 pandemic is estimated to have inhibited progress in ending TB by reducing the access to health services and treatment (Cilloni et al., 2020). The COVID-19 pandemic was predicted to potentially cause an increase of 0.2 – 0.4 million TB deaths in 2020 if diagnosis and treatment for TB reduced by 25 – 50% over at least 3 months. As the effects of the pandemic are ongoing, the burden of TB is likely to worsen and so it is increasingly important to investigate TB.

1.1.1 Mtb pathogenesis

Mtb begins infections through inhalation of bacterial droplets from an infected adult individual (**Figure 1.1**) (Kanabalan et al., 2021). The bacteria then enter the host's lungs where they infect local innate immune cells including alveolar macrophages. These macrophages then either control or kill Mtb, leading to latent TB infection (LTBI) or removal of infection, respectively. The control of Mtb by macrophages depends on several factors including macrophage activation, subtype (e.g. M1 or M2 macrophages) and bacterial virulence factors which can inhibit macrophage control. Macrophages which cannot control Mtb result in active infection, which is characterised by non-specific signs and symptoms such as coughing, haemoptysis (coughing up blood), weight loss, and transmission of disease. In this case, Mtb is phagocytosed by the macrophage in a compartment called the phagosome which should fuse with the lysosome within the macrophage cell to carry out different killing mechanisms. The bacteria can prevent this fusion and/or prevent the defence mechanisms within these compartments to escape the intracellular killing. The bacteria can then replicate within the macrophage and avoid detection from the immune response, leading to active TB disease. This leads to granuloma formation (multicellular structures of host immune cells and Mtb cells to contain Mtb

replication). In some immunocompromised individuals, such as children or those with human immunodeficiency virus (HIV), further forms of TB can develop such as disseminated TB where the bacteria travel to other organs other than the lungs. Alternatively, if the bacteria are controlled by the host immune response but not eradicated, LTBI can develop. This occurs where *Mtb* growth is restricted but these bacilli remain in a dormant state undetected by the host immune response, sometimes for years. LTBI can resolve if the immune response eradicates infection over time. However, if the host immune system becomes immunocompromised, for example with old age, and the LTBI can reactivate into active TB.

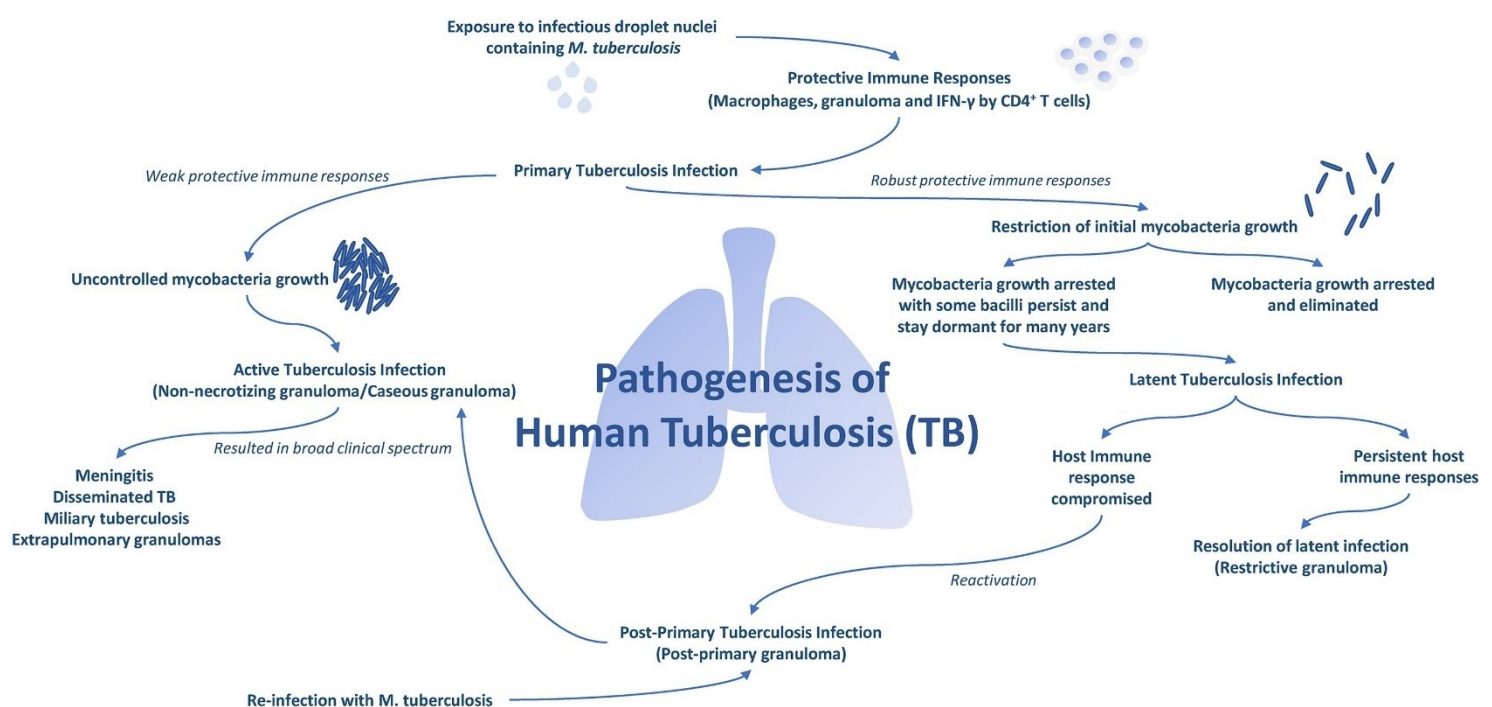
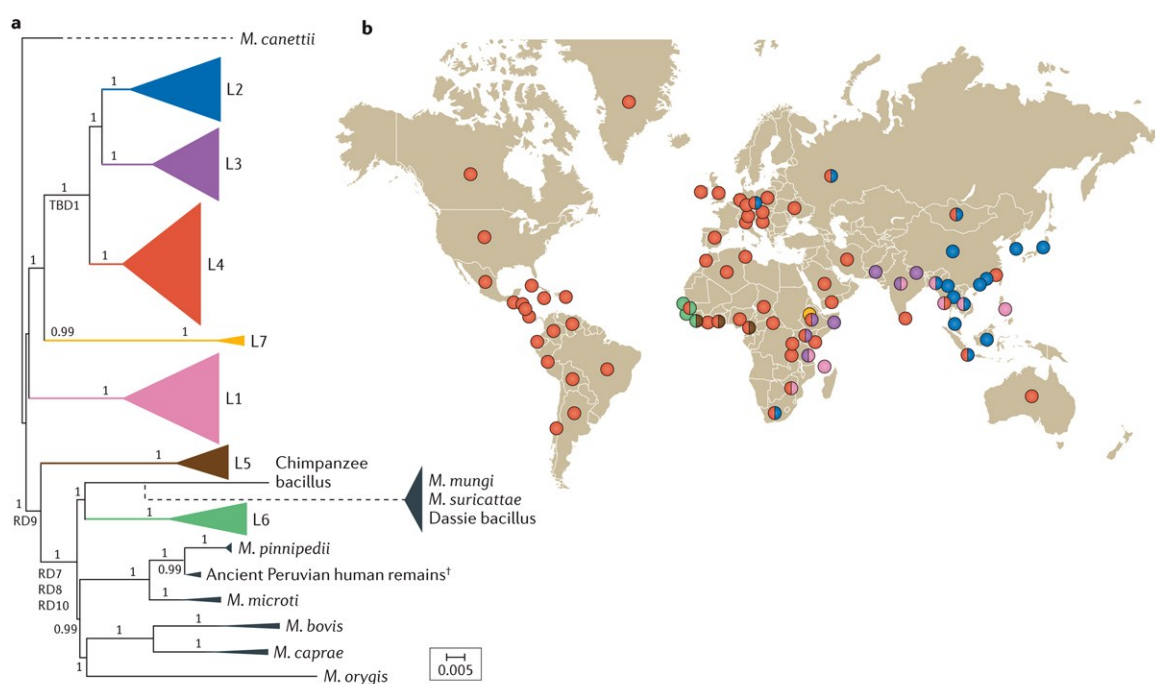


Figure 1.1: An overview of TB pathogenesis (Kanabalan et al., 2021). Reprinted from *Microbiological Research*, 246, Human tuberculosis and Mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery, 126674, Copyright (2021), with permission from Elsevier.

1.1.2 Mtb evolution

Mtb has been present for thousands of years and as a result has evolved alongside the human host (Comas et al., 2013). The bacterial species group called the *Mycobacterium tuberculosis* complex (MTBC) is the causative agent of human and animal TB (Gagneux, 2018). The MTBC includes *M. tuberculosis sensu stricto*, *M. africanum*, *M. orygis*, *M. bovis*, *M. microti*, *M. canettii*, *M. caprae*, *M. pinnipedii*, *M. suricattae*, and *M. mungi* (**Figure 1.2**). Within the MTBC, the species which are the most common causes of human TB disease are *M. tuberculosis sensu stricto* and *M. africanum*, which together make up the 7 phylogenetic Mtb lineages. Lineages 1, 2, 3, 4, and 7 consist of Mtb *sensu stricto* species whilst lineages 5 and 6 are made up of *M. africanum* species. There have also been recent publications suggesting there are also new lineages 8 and 9 (Coscolla et al., 2021; Ngabonziza et al., 2020). Hereafter I use the term Mtb as shorthand for all MTBC species which cause human disease, e.g. the Mtb lineages 1 – 9 and the reference Mtb strain H37Rv.



Nature Reviews | Microbiology

Figure 1.2: Phylogenetic and geographical distribution of MTBC (Gagneux, 2018). **A)** The phylogenetic tree is rooted with *Mycobacterium canettii* and is based on previous genomic data (Bos et al., 2014). The human lineages are shown in colour, with the size of each triangle representing the estimated size of these lineages in number of strains. The animal lineages and strains are indicated in grey. The scale indicates the number of nucleotide substitutions per site. TB1 = tuberculosis-specific deletion 1; RD7 – 10 = region of difference 7 – 10. **B)** Geographical distribution of human Mtb lineages. Reprinted by permission from Nature Reviews Microbiology: Springer Nature, Sebastien Gagneux, 2018.

Several studies indicate that MTBC originated from Africa, following human evolution from the same region and has diverged into lineages (Comas et al., 2013) which is also supported by the fact that Africa has the largest number of Mtb lineages (L1 – 9) (Gagneux, 2018). In contrast to the location of the original MTBC strain, identifying the original time of evolution of human TB has been challenging, with Comas, *et. al.*, estimating human TB developed least 70,000 years ago, while others have only suggested 6,000 years ago (Gagneux, 2018), depending on the methods used to estimate the age of the original MTBC strain. Comas, *et. al.*, also suggested a similar branching of human populations, as measured by mitochondrial macro-haplogroups which are commonly used to track human evolution, with Mtb lineages. While the Neolithic Demographic Transition of the human population was one of the major factors in the spread and evolution of separate Mtb lineages globally, other factors have also contributed to this, such a recent spread in agriculture in countries neighbouring China, around 3,000 – 5,000 years ago mirroring the expansion of the lineage 2 Beijing Mtb strains.

Similarly, the original ancestor of human TB is still unknown, as human TB strains have no known animal or environmental reservoir. However, through comparison studies of Mtb with closely related environmental nontuberculous mycobacteria (NTM) *M. kanasii* and *M. marinum*, it is thought that Mtb strains causing human disease has developed from an environmental pathogen and a combination of genetic deletion and acquisition events through horizontal gene transfer (Veyrier et al., 2009). The closest living relative of the most recent common ancestor of human TB strains, smooth tubercule bacilli, are mainly found in the Horn of Africa, also suggesting Africa as the origin of modern day Mtb (Cristina et al., 2005). Smooth tubercule bacilli are thought to be environmentally transmitted to humans and include the *M. canettii* strain commonly used as the root (i.e. surrogate for the most recent common ancestor) of Mtb phylodynamic studies. Their genomes are 10-115 kb larger than human Mtb strains and have a higher sequence variability (up to 65,000 single nucleotide polymorphisms, SNPs, difference) than Mtb strains (up to 2,000 SNPs difference) (Gagneux, 2018).

Different MTBC strains (both human and animal) show differences in specific genomic regions called regions of difference (RD) which vary from 2 to 12.7 kb in length (Brosch et al., 2002). In addition, strains within the MTBC can be classified as evolutionary ancient or modern based on the presence or absence of a specific deletion 1 (TbD1) which is present in certain strains in the MTBC but not others (Brites & Gagneux, 2015). Ancient strains contain the TbD1 site and include *M. africanum* and Mtb lineages 1, 5, 6, and 7 as well as *M. bovis*. In contrast, modern lineages including Mtb lineages 2, 3, and 4 do contain a deletion at the TbD1 site. These regions result from mobile genetic elements (RD3

and RD11), insertion sequences (RD5 and RD6), homologous recombination between insertion sequences leading to deletion (RvD2, RvD3, RvD4, and RvD5), and deletions in coding regions resulting in truncated genes through an unknown mechanism (RD1 – 2, RD4, RD7 – 10, RD12 – 14, and TbD1). These deletions indicate the evolution of different Mtb strains (**Figure 1.3**). In addition, the presence of these regions also suggests that the animal MTBC strains such as *M. bovis* derived from human Mtb strains as these strains do not contain any gene clusters not also found in *M. africanum* strains.

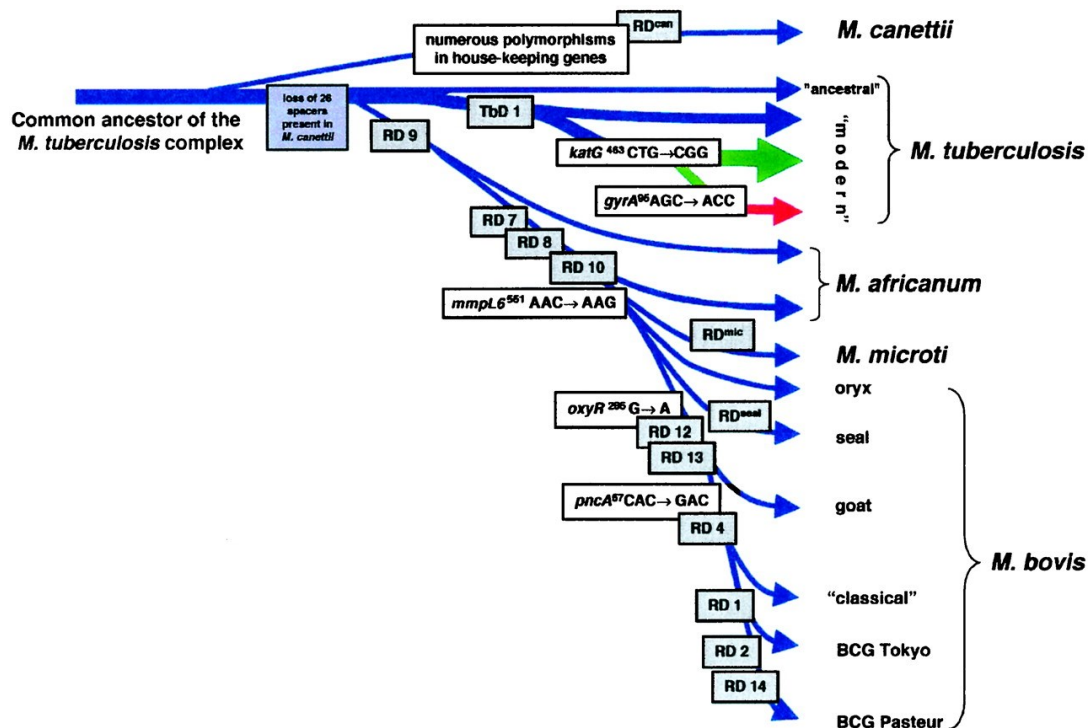


Figure 1.3: Proposed evolutionary divergence of MTBC (Brosch et al., 2002). The evolution is estimated from the presence or absence of conserved deleted regions and sequence polymorphisms. Grey boxes indicate loss of DNA. Blue arrows indicate strains which are characterised by the *katG* 463 sequence CTG and the *gyrA* 95 sequence ACC; green arrows indicate strains characterised by the *katG* 463 sequence CGG and the *gyrA* 95 sequence ACC; red arrows indicate strains characterised by the *katG* sequence CGG and the *gyrA* sequence AGC. Copyright (2002) National Academy of Sciences, U.S.A.

These lineages are groups of genetically similar strains which are classified according to specific molecular genotyping methodologies. The current hypothesis of low or no horizontal gene transfer in Mtb allows for these molecular typing methods to accurately classify Mtb strains by lineage. There are several different methods for molecular typing used, with spoligotyping, typing through polymerase chain reaction (PCR)-amplification of direct repeats, and large deletions of regions of difference being three of the most frequently used methods (Jagielski et al., 2014). Recently, however, a 90-SNPs

barcoding signature has been proposed to classify a greater number of Mtb strains into the highest number of lineages (9) and sublineages (90) to date, and identified 30 new sublineages (Napier et al., 2020). While these methods have been useful to categorise strains for comparative analyses, there are no standardised typing methods which has led to challenges in comparing data across studies. Each of these lineages can be found in different locations worldwide (**Figure 1.1**). Certain lineages, such as lineage 4, can be found in several regions (Gagneux, 2018). Other lineages, such as lineage 2, can be found in a restricted number of countries such as those in Eastern Asia, and other lineages, such as lineage 7, have only been identified in a single country, in this case Ethiopia. Lineages are thus often referred to by location-based naming conventions, such as lineage 2 being identified as the Beijing lineage.

1.1.3 Mtb diagnosis

As part of the global effort to fight TB, several tools have been developed over the years to tackle the disease. There are several diagnostic tests available with different accuracies, sensitivities, and specificities. The 'gold standard' diagnostic test for TB is culturing the bacterium from a sputum sample (Pai et al., 2016). Bacilli are cultured either on Lowenstein-Jensen agar plates or in liquid culture using the Mycobacterial Growth Indicator Tube or BacT/ALERT MB assays. Detection of the bacilli is performed by molecular or immunochromatographic lateral flow testing.

No other diagnostic test can match sputum culture in terms of accuracy, despite the test being reviewed by the WHO in 2007 (Pai et al., 2016). However, this does not come without limitation, including that it can take several weeks to receive a result. This time to diagnosis is too long as TB symptoms are similar to many non-bacterial diseases and clinicians must decide if antibiotic treatment is necessary, which could lead to antibiotic resistance if misdiagnosed. Another major drawback of the sputum culture test is the difficulty in producing a sputum sample, particularly in paediatric TB patients.

Another common diagnostic test is smear microscopy which is performed on sputum or other biological samples (Pai et al., 2016). This involves either acid-fast staining using stains which accumulate in the cell wall of Mtb and the lipids in the cell wall prevents decolorisation using acid-containing reagents. The different stains used for this method have different sensitivities, for example the Ziehl-Neelsen stain which is commonly used in low-income countries, but is cheaper and faster than other strains, has a 10% lower sensitivity than fluorescent microscopy stains such as Auramine.

This test also has a particularly low sensitivity in patients with paucibacillary TB (low numbers of bacilli) such as paediatric TB patients.

Recent developments in diagnostic testing have produced few new tools in terms of tests with higher accuracy for active TB. The Gene Xpert which was approved for use by the WHO in 2010 has a high specificity (99%) and can also detect multi-drug resistant Mtb (MDR-TB) (Dorman et al., 2018). This test involves PCR amplification of bacterial DNA from sputum samples (Pai et al., 2016). The test also identifies rifampicin resistance through amplifying an 81 bp region of the *rpoB* gene in Mtb bacilli, which is where greater than 95% of rifampicin resistance mutations are found. However, the Gene Xpert requires sputum samples and therefore has the same limitations relating to sample collection as the standard culture method. It also has varying sensitivity depending on the phenotype of TB disease, such as lower sensitivity for HIV infected individuals, and has limitations in where it can be used as it requires constant power (WHO, 2014).

Other nucleic acid amplification tests (NAATs) have recently been endorsed by the WHO (WHO, 2021). These tests are based on different types of PCRs including TB-Loop-Mediated Amplification Test (LAMP) (Pai et al., 2016). In addition, for patients with HIV infection and low CD4 counts, the urine lipoarabinomannan (LAM) rapid test can also be used. This test uses a lateral flow assay which detects LAM, a cell wall component of Mtb, in urine samples. Furthermore, patients predicted to be infected with latent TB can be diagnosed using tests which stimulate the host immune system, with those infected showing positive responses. These tests include the tuberculin skin test (TST) and IFN γ stimulation assays (IGRAs) such as QuantiFERON-TB and T-SPOT (Meier et al., 2019). TST uses a mixture of Mtb antigens called purified protein derivative (PPD) which is injected intradermally and stimulates an inflammatory response (Haas & Belknap, 2019). The amount of inflammation is measured, and a positive response is determined if the inflammation is above a threshold (which depend on the history of the patient). IGRAs are *in vitro* tests which use other Mtb antigens such as early secretory antigen target-6 (ESAT-6) to stimulate an IFN γ response. QuantiFERON-TB uses an enzyme-linked immunosorbent assay (ELISA) to measure IFN γ production from whole blood whereas the T-SPOT uses an enzyme-linked immune absorbent spot (ELISpot) assay to measure IFN γ production from human peripheral blood mononuclear cells (PBMCs).

However, there are still several other new technologies with the potential for development into diagnostic tests (WHO, 2021). One of these include host or bacterial biomarker signatures. Biomarkers use the presence of key genes, transcripts, proteins, lipids, or metabolites to indicate the presence of

the bacterium either by detecting these molecules from the bacterium itself, or from a characteristic immune response in the host which is distinct for TB.

1.1.4 Host infection and immune response

The innate immune response to Mtb in active infection

Around 5 – 10% of individuals who become infected with Mtb develop active TB (Sia & Rengarajan, 2019). Active TB is caused by the inhalation of aerosolised droplets from an infected adult with active TB disease. After inhalation, Mtb infects alveolar macrophages residing in the host lungs, before infecting other recruited cells including neutrophils and dendritic cells (DCs) (Eum et al., 2010; Sia & Rengarajan, 2019). Mtb pattern-associated molecular patterns (PAMPs) are recognised by these immune cells to stimulate an immune response such as phagocytosis of the bacterium (**Figure 1.4**). These PAMPs include several proteins and lipids such as the virulence factor mannosylated LAM and the Mtb membrane component peptidoglycan, as well as other molecules for instance mycobacterial DNA. They are recognised by a variety of pattern recognition receptors (PRRs) such as C-type lectins, complement receptors, collectins, scavenger receptors, Fc receptors, glycosphosphatidylinositol-anchored membrane receptors and Toll-like receptors (TLRs). For instance, LAM is recognised by TLR2 and stimulates pro-inflammatory mediators such as nitric oxides (NO) production in macrophage cells (Sia & Rengarajan, 2019).

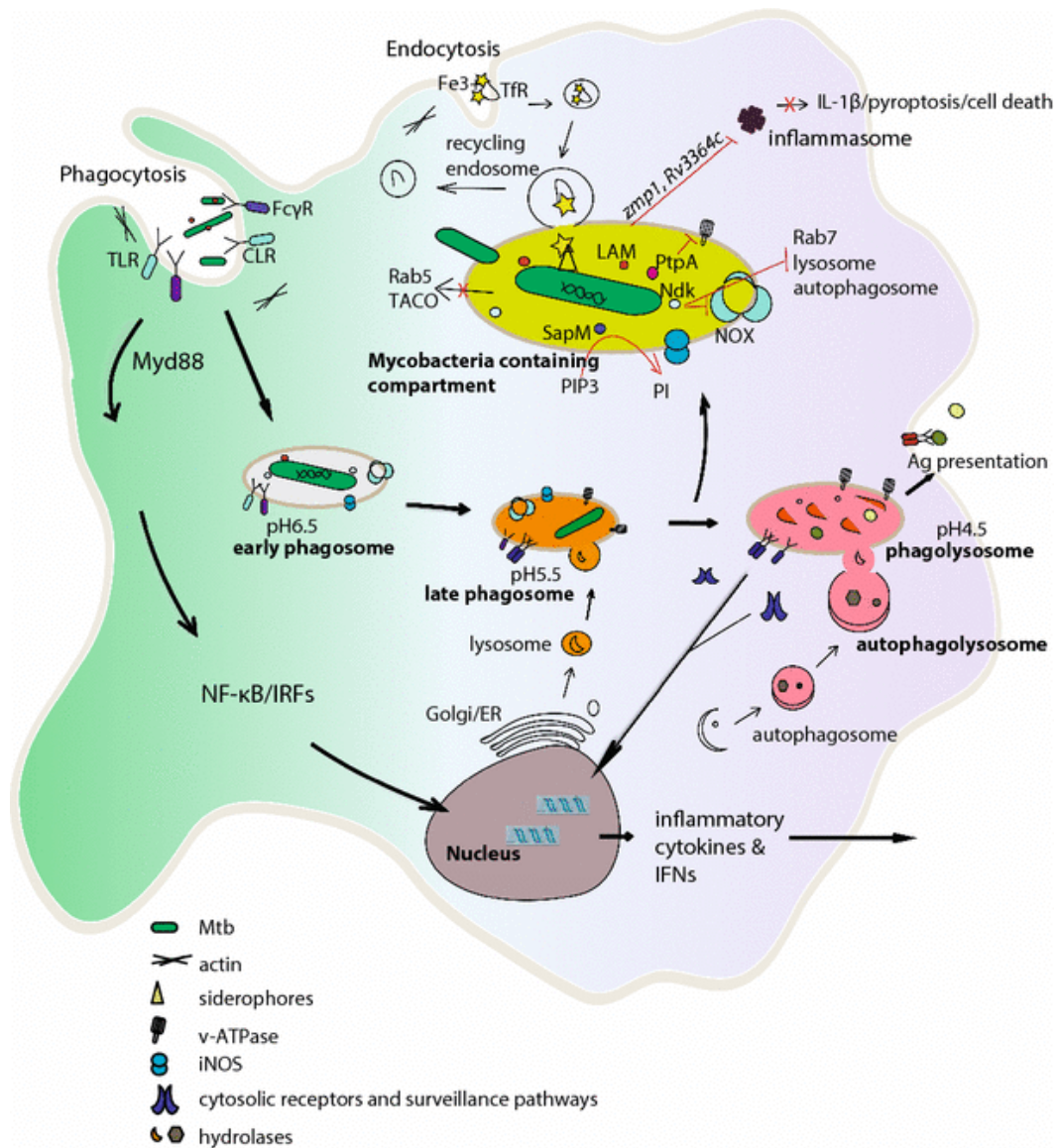


Figure 1.4: Defence mechanisms of the host macrophage and avoidance and suppression by Mtb (Awuh et al., 2017). Mtb are phagocytosed by host macrophage cells through various receptors such as TLRs. Within the phagosome there are multiple defence mechanisms including reactive oxygen species. Phagosomes then fuse with lysosomes, which is a process that can be inhibited by Mtb. Successful fusion can lead to the breaking down of Mtb and presentation of Mtb antigens to stimulate the adaptive immune system alongside cytokine production. TLR = toll like receptors; CLR = C-type lectin receptors; FcγR = Fc receptor; IRF = interferon regulatory factors; ER = endoplasmic reticulum; IFNs = interferons; TfR = transferring receptor; LAM = liparabinomannan; NOX = NADPH oxidase; PIP₃ = phosphatidylinositol (3,4,5)-trisphosphate; PI = phosphatidylinositol; PtpA = serine/threonine-protein phosphatase 2A; Ndk = nucleoside diphosphate kinase; TACO = tryptophan aspartate-containing coat protein; Ag = antigen. Reprinted by permission from Cellular and Molecular Life Science: Springer Nature, Jane Atesoh Awuh and Trude Helen Flo, 2017.

Within the lungs, the host immune response becomes activated as a result of these PAMP-PRR interactions and largely depends on the response stimulated by infected macrophages, one of the most important cells to control Mtb replication and proposed to be the predominantly infected cell type during infection. Alveolar macrophages phagocytose Mtb before transporting the bacterium to an intracellular compartment called the phagosome. The phagosome contains several defence and killing mechanisms such as the generation of reactive oxygen species (ROS) (Hmama et al., 2015). The phagosome then begins maturation which results in fusion of the phagosome with lysosomes or endosomes bringing into the compartment further defence molecules and mechanisms (**Figure 1.4**). This includes lowering the phagolysosomal pH through increasing the number of vacuolar H⁺-ATPase (V-ATPase) and lysosomal hydrolases. Other mechanisms include the production of harmful molecules such as free radicals from ROS and nitric oxide (NO) (Hmama et al., 2015), and the accumulation of toxic metal ions such as copper (Neyrolles et al., 2015).

In response to this, Mtb has developed subversion and suppression mechanisms to inhibit or avoid these host defence strategies. Mtb can inhibit intracellular trafficking, phagolysosome maturation, autophagy, host cell death, and also neutralize molecules used to kill the bacterium such as ROS and metals (de Martino et al., 2019; Sia & Rengarajan, 2019). For example, the ESX-1 secretion system is thought to increase the permeabilization of the phagosome, possibly through the ESAT-6 protein, thereby damaging its integrity. Mtb can also escape the phagosome and enter the cell cytosol using the ESX-1 (Chai et al., 2020). Furthermore, the Mtb protein PtpA can bind to the host macrophage V-ATPase inhibiting its function and preventing the acidification of the phagolysosome (Hmama et al., 2015). If Mtb is able to survive within the macrophage cell, it replicates there undetected from the immune response. The balance of killing and evasion strategies between the macrophage cell and Mtb bacilli determines the outcome of infection.

Macrophage cells and the phagosome have been extensively studied with respect to TB disease. However, recently it has been discovered that other innate immune cells can also become infected with Mtb and, moreover, play important roles in infection. One of these is the monocyte cell. Monocytes circulate in the bloodstream until they are recruited to the site of infection where they differentiate into macrophages thereby being directly relevant for TB disease. In one study CD11c^{Hi} monocyte-derived macrophages were suggested to be the predominantly infected cell in chronic TB in mice, implying there might be an important role for monocytes alongside alveolar macrophages (and other macrophage subsets) in TB (Lee et al., 2020). In addition, it has been shown that monocytes

which were transferred into Mtb-infected mice were the predominant producers of inducible nitric oxide synthase (iNOS) which is used to kill Mtb bacilli (Sia & Rengarajan, 2019).

Another immune cell which has been shown to play a role in TB is the neutrophil. Neutrophils, usually the first line of defence, are the first cells to be recruited in Mtb infection after stimulation by cell trafficking signals such as chemokines and chemoattractants (Eum et al., 2010; Sia & Rengarajan, 2019). Neutrophils were discovered to be the main infected cell in sputum, bronchoalveolar fluid, and in the cavities of 15 pulmonary TB patients (Eum et al., 2010). It is important to note that the discrepancies between which cells are the predominantly infected cell type during Mtb infection may result from different models used to study TB and differences in the markers used to distinguish cell populations. Regardless, neutrophils have been the topic of several TB studies. It was shown that a neutrophil-driven type I and type II IFN response was the main gene expression signature in active TB patients, although sample sizes were small in this study (Berry et al., 2010). Since this study, research has identified conflicting roles for neutrophils in TB with some studies identifying an immunomodulatory role through over-expression of programmed death-ligand 1 in the blood of TB patients (McNab et al., 2011) but the use of mice models has identified higher rates of neutrophil recruitment in strains susceptible to TB (DBA/2 mice) compared to those which are resistant to TB (C57BL/6 mice) (Sia & Rengarajan, 2019). Neutrophils have also been shown to inhibit Mtb replication in other studies but their excessive stimulation of the production of IL-17, a pro-inflammatory cytokine, may cause host tissue damage. Currently it is not clear if neutrophils are beneficial or detrimental to TB, however these studies indicate there may be certain environmental conditions which will influence the outcome of neutrophil involvement in infection.

In active TB, the host immune response is actively trying to fight the infection. Host inflammation can be protective against Mtb but can also be destructive when uncontrolled. Several studies have shown that too little, or excessive, inflammation, exacerbates disease and a balance is required for successful clearance of the bacteria (Sia & Rengarajan, 2019). For example, type I IFN signalling has been shown to increase host susceptibility to TB through dampening inflammation through inhibition of IL-1 β production which inhibits the inflammasome, stimulates the production of IL-10 which suppresses immune responses, and reduces IFN γ responsiveness in infected macrophages (McNab et al., 2014). In contrast, certain lipids which increase inflammation are required for successful eradication of Mtb as mice which cannot produce the proinflammatory prostaglandin E2 are more susceptible to TB than wild-type mice. These discrepancies may be due to the use of different infection models, or it may

reflect that inflammation must be used by the host at specific times during infection and in a controlled manner to be protective against TB.

DCs are another innate immune cell which can become infected in TB (Sia & Rengarajan, 2019). DCs are professional antigen presenting cells in that they function as the main source of communication between the innate and adaptive immune responses. DCs migrate from the site of infection to lung-draining lymph nodes and present bacterial antigens derived from other innate immune cells or from their own cytosol to activate naïve T cells. T cells must be stimulated by additional receptors such as costimulatory receptors and cytokines, as well as Mtb antigens, to be activated. Similarly, to macrophages, Mtb has been shown to infect DCs and use several mechanisms to prevent antigen presentation including inhibition of DC maturation in human and mice models, inhibition of the host endosomal sorting complex required for antigen presentation in mice (Portal-Celhay et al., 2016), and may inhibit costimulation and cytokine production of DCs through reduction of major histocompatibility complex (MHC) class II, CD40, CD86 protein expression (Bizzell et al., 2018). As the main cell type to activate the adaptive immune response, it is critical to further investigate how mycobacteria may subvert this response by the DCs.

The adaptive immune response to TB during active Mtb infection

Although the innate immune response is critical for successful eradication of Mtb, control of infection also depends on activation of the adaptive immunity.

B cells are part of the humoral adaptive immune response (Chan et al., 2014). This typically involves antigen presentation to naïve B cells which leads to the differentiation of B cells into plasma and memory B cells which target the present infection and future infections, respectively. In TB disease, however, their roles are not fully understood. The Bacillus Calmette-Guérin (BCG vaccine was developed in 1921 to stimulate a B cell response towards Mtb infection (Calmette & Guérin, 1920; Mahairas et al., 1996). After development of the BCG vaccine showed varying efficacy in different populations, it was suggested that B cell immunity was not required to control infection (Chan et al., 2014). Studies have shown that mice lacking B cells have increased lung damage during TB, possibly due to an increased recruitment of neutrophils (Sia & Rengarajan, 2019). In contrast, the production of type I IFN from B cells has been shown to stimulate macrophage differentiation towards an anti-inflammatory phenotype and therefore may limit their differentiation towards the pro-inflammatory

phenotype required to control infection. This suggests that there is some role for B cells during infection, however the importance of that role remains unclear.

In contrast, cellular immunity has been identified in several studies to provide protective immunity to TB (Chan et al., 2014). While there are still protective properties of humoral immunity, research has focussed predominantly on T cell responses in TB in the last few decades. CD4 T cells include T helper cells such as Th1, Th2, and Th17 cells, as well as regulatory FoxP3 T cells (Tregs). T helper cells produce cytokines according to pro-inflammatory (Th1 and Th17) or anti-inflammatory (Th2) profiles while Tregs regulate and suppress the immune response through production of inhibitory cytokines such as TGF β and IL-10 (Geginat et al., 2014). Studies using animal models have shown that CD4 T cells are essential for survival of Mtb infection as MHC II knock-out mice or CD4 depleted mice have high mortality rates to Mtb and CD4 depletion in cynomolgus macaques results in activation of TB disease from LTBI (Sia & Rengarajan, 2019). Th1 T cells are required for this successful control of infection through the production of IFN γ which activates macrophages to enable them to kill intracellular Mtb. IFN γ is produced by several immune cells including Th1 and Th17 cells and is one of the most important cytokines in TB. Studies have identified Mendelian susceptibility to mycobacterial disease (MSMD) in humans which is caused by genetic mutations in different genes involved in IFN γ signalling and support the notion that IFN γ signalling is necessary for control of Mtb infection.

Under specific conditions Th17 cells have shown to be either beneficial or detrimental to infection (Sia & Rengarajan, 2019). Th17 cells regulate inflammation through the production of inflammatory cytokines such as IL-17 and IL-22, as well as IFN γ (Tesmer et al., 2008). Th17 cells can be protective in some models of TB, as mouse mutants in the IL-17A signalling pathway have an increased number of bacteria (Sia & Rengarajan, 2019). Furthermore, IL-17-producing CD4 T cells were found in higher abundance in the peripheral blood mononuclear cells (PBMCs) and bronchoalveolar lavage (BAL) of BCG-vaccinated individuals compared with active TB patients. Despite the beneficial functions of Th17 cells, they also regulate neutrophils which, as described, can be harmful to the host during infection.

Other than CD4 T cells, CD8 T cells have also been implicated in TB (Sia & Rengarajan, 2019). CD8 T cells are cytotoxic T cells which identify and kill infected host cells to prevent the uncontrolled spread of infection, as well as secrete cytokines, including IFN γ and TNF α , which alter the immune response. These cells are activated through MHC class I stimulation with Mtb antigens from innate immune cells.

After stimulation, CD8 T cells produce perforins which break open the target cell membrane and allow the transport of granzymes into the target cell. Granzymes can then kill Mtb within infected cells. CD8 T cells have been shown to be important for protection from Mtb infection as mice, with defects in their CD8 T cell responses develop Mtb infection faster than wild-type mice. Furthermore, upon removal of CD8 T cells from rhesus macaques, these animals fail to develop protective immunity after BCG vaccination or through antibiotic treatments. This indicates a possible role for CD8 T cells in long-term Mtb infection and Mtb memory.

In addition to these T cell subsets, there are additional T cells which are implicated in TB including $\gamma\delta$ T cells and mucosal-associated invariant T cells (MAIT) cells found in the lung mucosal layer, as well as CD1-restricted T cells which specialise in the detection of lipid antigens which are found in abundance on the Mtb cell wall (Sia & Rengarajan, 2019).

Immune response differences between Mtb lineages

The majority of studies investigating the immune response to Mtb infection typically use the reference Mtb strain H37Rv. H37Rv is the most well-studied and most characterised of the Mtb strains after being derived from the H37 strain which was isolated in 1905 (O'Toole & Gautam, 2017). H37Rv is part of the Mtb lineage 4 which has implications for the understanding of the immune response as other lineages, and also sublineages, may produce different host immune responses to infection. CDC1551 is a clinical isolate. There are some studies which have investigated these differences in the host immune response, although our understanding is still limited.

Krishnan et al. investigated the innate immune response towards Mtb strains from lineages 1, 2, and 4 isolated from Vietnamese TB patients. This study demonstrated that strains from lineages 1 and 2 induced higher TNF α and IL-1 β from bone marrow-derived macrophages (BMDM) than strains from lineage 4, although strains in lineage 1 also stimulated higher levels of the regulatory cytokine IL-10 (Krishnan et al., 2011). They also investigated the response in mice and found that lineage 2 strains produced a higher colony forming unit (CFU) count within the first 2 weeks of infection in mice spleens and lungs than lineages 1 and 4 strains. This higher bacterial load for lineage 2 strains has been validated in another study which also performed a comparison between lineages 2 and 4 (Mourik et al., 2019). This study found that the lineage 2 strain was present in higher CFU in the lungs and spleens of mice at days 7 and 14 post infection, while the lineage 1 strain only showed higher CFU counts at

day 14 in the lungs and spleen of mice compared to H37Rv. In addition, when looking at cytokine profiles of the lungs of infected mice they found that lineage 2 stimulated higher levels of IL-4, IFN γ , IL-17a, and TNF α at day 14 compared to H37Rv. While IFN γ , IL-17a, and TNF α are pro-inflammatory cytokines and may be induced by Mtb to create excessive inflammation, the host may be producing IL-4, a cytokine associated with tissue repair, to counteract this increase in inflammation. Previous studies have found, however, that an increase in IL-4 is common to hypervirulent lineage 2 strains and negatively affects the host as it diverts macrophages from the pro-inflammatory phenotype, which can control intracellular Mtb, to an anti-inflammatory phenotype. In one study, lineages 2, 3, and 4 (the modern lineages), induced lower inflammatory responses including IL-6, IL-12p40, TNF α , IL-15, MIP-1 α , and CCL5 in human monocyte-derived macrophages at early (24-hour) timepoints than lineages 1, 5, or 6 (the ancient lineages) (Portevin et al., 2011).

Lineage 2 strains in particular have been under intense investigation. Specifically, strains or sublineages in lineage 2 which have been found to be more hypervirulent than other strains in lineage 2. Chen, et. al., focussed on ancient and modern Beijing strains, based on the presence or absence of RD181 respectively, from lineage 2 which identified that ancient Beijing strains stimulated higher levels of the pro-inflammatory cytokines IL-1 β and IFN γ than some modern Beijing strains (Chen et al., 2014). In contrast, strains in lineage 1 induced higher levels of IL-1 β , IL-12p70, IL-6, TNF α , and IFN γ than Beijing strains in a human monocyte-derived macrophage cell model (Chen et al., 2014). A more recent study showed that the BL spoligotype from lineage 2 has a higher bacterial load and stimulates greater TNF α , IFN γ , and IL-12p70 at different timepoints than another lineage 2 spoligotype, SIT1, and H37Rv strains in an alveolar macrophage cell model (Irene et al., 2021). Overall, while there is data suggesting a more virulent role for some lineage 2 strains, the different strains and classification techniques used across studies prevent many studies from being directly compared.

The studies described highlight new information on the differences in the host immune response to Mtb lineages. However, some of the results are conflicting on the level of virulence of lineage, or sublineages, of lineage 2 strains compared to other lineages. A possible explanation for this is that almost all of these studies include a subset of lineages, sublineages, or strains which make it challenging to compare across studies. In addition, there has been a focus on specific lineages (1, 2, and 4), or lineages found in individual countries, and a more global study would be beneficial to understand the virulence and host response across all Mtb strains. Despite this, the study of the

immune responses to these lineages is still essential for our understanding of TB as they reflect true infection more closely than the reference strain H37Rv.

1.1.5 Modelling TB disease through host-pathogen interactions

Host-pathogen interaction studies involve investigating how the pathogen and host interact with one another and respond to each other during infection. This type of analysis is particularly crucial for Mtb infection, as Mtb has evolved several immune evasion and subversion strategies as a response to attempts by the host to eliminate the bacteria. As mentioned, Mtb has developed these strategies to subvert the immune response in many different immune cell types and due to the complexity of these interactions there are several different presentations of TB disease in humans (**Figure 1.5**). These different presentations include pulmonary disease, extrapulmonary disease, LTBI, as well as different presentations within these categories. The spectrum of disease presentation is likely due to differences in the host-pathogen interactions between the immune response and specific Mtb strains. It is therefore important to understand these interactions between host and pathogen in different settings, such as acute or chronic disease, to reflect the different disease progressions observed during *in vivo* infection.

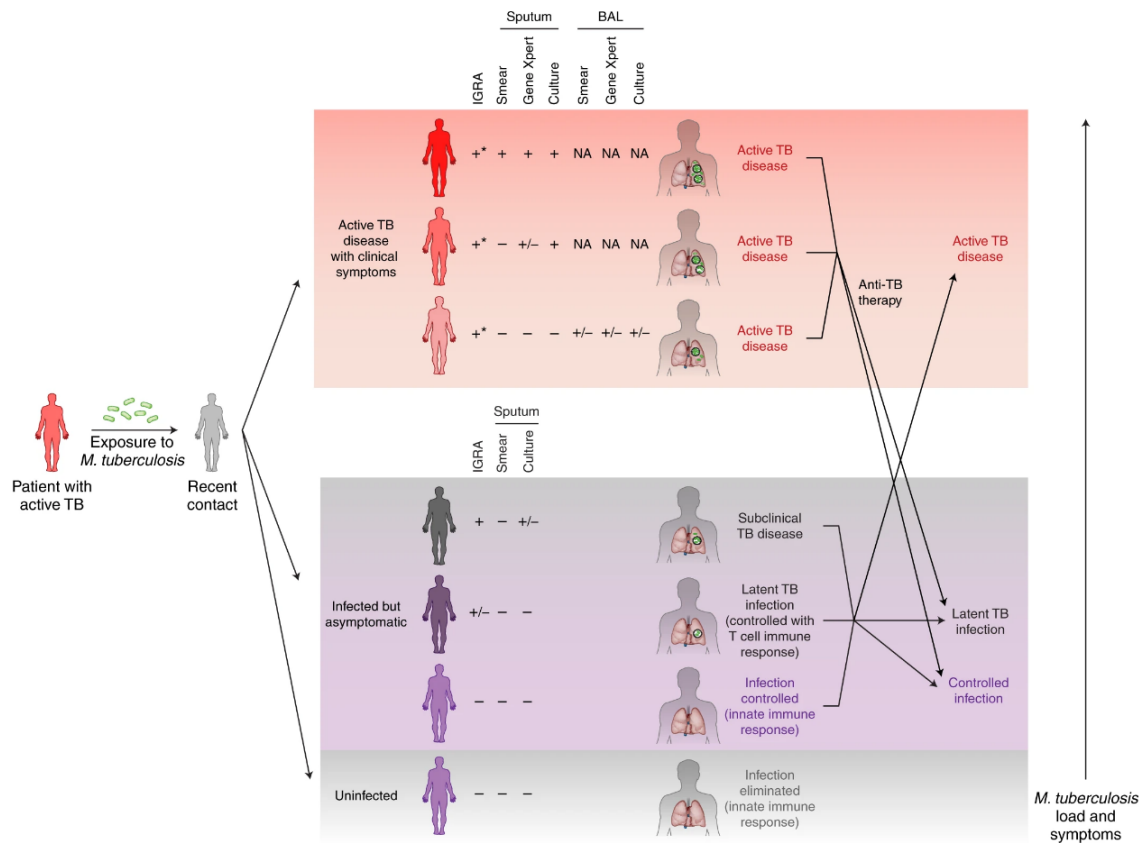


Figure 1.5: Heterogeneity in the host immune response to TB (Singhania et al., 2018). Exposure to an individual with active TB can result in a range of disease outcomes, shown in boxes. Most individuals remain asymptomatic (middle grey/purple box) and will either clear the infection and become uninfected, remain asymptomatic either due to control of the infection by their innate immune response, or their cellular immunity, or they will develop subclinical TB. Around 10% of exposed individuals will develop active TB (top red box). Active TB patients may also show a range of disease states, only some of which can be identified using certain diagnostic tests as described. * = usually positive; +/- = positive or negative; NA = not applicable. Reprinted by permission from *Nature Immunology*: Springer Nature, Singhania A, Wilkinson RJ, Rodrigue M, Haldar P, O'Garra A, 2018.

Animal models to study host-pathogen interactions

Deciding which infection model to use is an important step when investigating host-pathogen interactions in Mtb as many animal models used to study infectious disease models do not replicate true TB infection. It is challenging to model every aspect of TB disease in any one model of infection, and most studies have focussed on cell line models (particularly macrophage cell lines), mouse models, or samples from humans such as blood. However, there are limitations associated with each model (**Table 1.1**). Other than logistical differences such as cost and the availability of housing facilities, there are also large differences in the immune responses of the selected host.

Mice have been frequently used in many fields of biology to model disease with the most commonly used strains being C57BL/6 and BALB/c mice as they are the most studied and therefore the best characterised strains. Mouse models can be used effectively to model acute TB infection (the first 1 – 4 weeks of infection) until T cells become primed and migrate to the lungs (Flynn, 2006). However, when infected with TB most mouse strains are either resistant to infection or are susceptible and have early death (Chackerian & Behar, 2003). In addition, most mouse models do not replicate LTBI or granulomas accurately to what is seen in human disease as mouse granulomas are not structured in similar cellular layouts as seen in humans (Flynn, 2006). As a result of this, other animal models are often used instead of, or to supplement, mice studies including guinea pig, rabbit, and non-human primate models.

Table 1.1: The advantages and disadvantages of different Mtb infection models.

Model	Advantages	Disadvantages	Reference
Mice	- Most characterised <i>in vivo</i> system - Multiple mutants available - Many reagents available - Similar T cell responses as in humans - Similar chronic infection as in humans - Multiple immune cells interacting	- Relatively high cost - Requires ethics - Requires extensive training - Requires appropriate facilities - Does not model latency well (progressive disease in all mice) - Mouse granulomas do not reflect human granulomas	(Flynn, 2006)
Guinea pigs	- Multiple immune cells interacting - Similar granulomas as in humans	- High cost - Requires ethics - Requires extensive training - Requires appropriate facilities - Few reagents available - Most susceptible model	(Flynn, 2006)
Zebra fish	- Can measure mycobacterial virulence factors - Granuloma formation can be visually studied with transparent embryos - Cheaper to house	- Cannot be infected with H37Rv, only <i>M. marinum</i> - Does not represent other phenotypes of TB disease well	(Flynn, 2006; Zhan et al., 2017)
Moths	- Lower cost - Many reagents available - Requires less specialised training - Requires less specialised facilities - Multiple immune cells interacting - Can be used to study drug development for TB	- No adaptive immune system - Relatively new system and therefore uncharacterised	(Asai et al., 2019)
Non-human primates	- Pathology similar to human disease - Representative of latent TB - Can model coinfections such as HIV with TB	- Very high cost - Requires ethics - Small sample sizes used - Requires extensive training - Requires appropriate facilities	(Flynn, 2006)

Human challenge model	Represents mycobacterial disease accurately	BCG used for infections and not Mtb Limited sample types can be taken from donors High cost Requires ethics Requires donors	(Minassian et al., 2011)
Murine cell lines (e.g. macrophage cell lines)	Lower cost Requires less specialised training Requires less specialised facilities Can study bacilli invasion and intracellular replication	Only allows the study of a single cell type	(Keiser & Purdy, 2017)
Human cell lines (e.g. monocyte-derived macrophages, THP-1 cells)	Lower cost Requires less specialised training Requires less specialised facilities Can study bacilli invasion and intracellular replication	Only allows the study of a single cell type	(Fonseca et al., 2017)
Primary cell lines (e.g. PBMC-derived macrophages)	Lower cost Requires less specialised training Requires less specialised facilities Can study bacilli invasion and intracellular replication	Only allows the study of a single cell type Requires donors	(Fonseca et al., 2017)
Whole blood assays	Lower cost Requires less specialised training Requires less specialised facilities Multi-cellular Can model cell-cell interactions	Requires ethics Requires donors Cannot be used to model long-term immune responses	(von Both et al., 2018)

Cell-based models to study host-pathogen interactions

Macrophage cells are the most frequently utilised cell lines to study Mtb infection. These include mouse bone-marrow derived macrophages and human-derived lines such as THP-1 cells, U937 cells, and human monocyte-derived macrophages (Mendoza-Coronel & Castanon-Arreola, 2016; Pisu et al., 2020a). These different cell lines have different immune responses towards infection, such as the lower levels of phagocytosis observed in U937 cells than THP-1 and human monocyte-derived macrophages (Mendoza-Coronel & Castanon-Arreola, 2016). These *in vitro* cell lines are simpler and cheaper to use than *in vivo* models thereby making them a more accessible model than mouse models, however they have the limitation of only being able to model the interactions of one cell type with the pathogen without external environmental stimuli.

An alternative to using cell lines which, limit the number of cell types involved in the infection, or animal models, which do not always reflect true infection in humans, is the use of *in vitro* whole blood models to study TB. Whole blood models, including whole blood assays (WBAs), consist of several types of immune cells which can interact with each other allowing for analysis of complex responses between different cells. Specifically, whole blood contains several innate and adaptive host immune cells including neutrophils, eosinophils, basophils, monocytes, dendritic cells, B cells, and T cells (**Figure 1.6**), as well as red blood cells and plasma (Autissier et al., 2010). Therefore, the WBA is more representative of *in vivo* infection than *in vitro* cell line models as it can model the interplay between several immune cells and host-pathogen interactions with multiple cells involved in the infection. The assay can also be used to investigate a number of hypotheses which relate to Mtb infection and the host immune response which cannot be easily interrogated using single cell models. In addition, *in vitro* WBAs overcome the high cost and housing requirements of *in vivo* models while still consisting of a complex multi-cellular system. These assays utilise bioluminescent bacteria, either Mtb or the surrogate strain BCG which can be used in a lower safety level laboratory. These bioluminescent bacteria express the *luxAB* genes from *Vibrio harveyi* located in a pSMT1 vector under the heat shock protein 60 mycobacterial promoter (Snewin et al., 1999). The *luxAB* genes express a luciferase which uses reduced flavin mononucleotide as a cofactor to produce light. The bioluminescence properties of these strains allow the measurement of bacterial growth in the whole blood over time. This allows the added variable of time to be investigated using whole blood models as well as complex immune cell interactions.

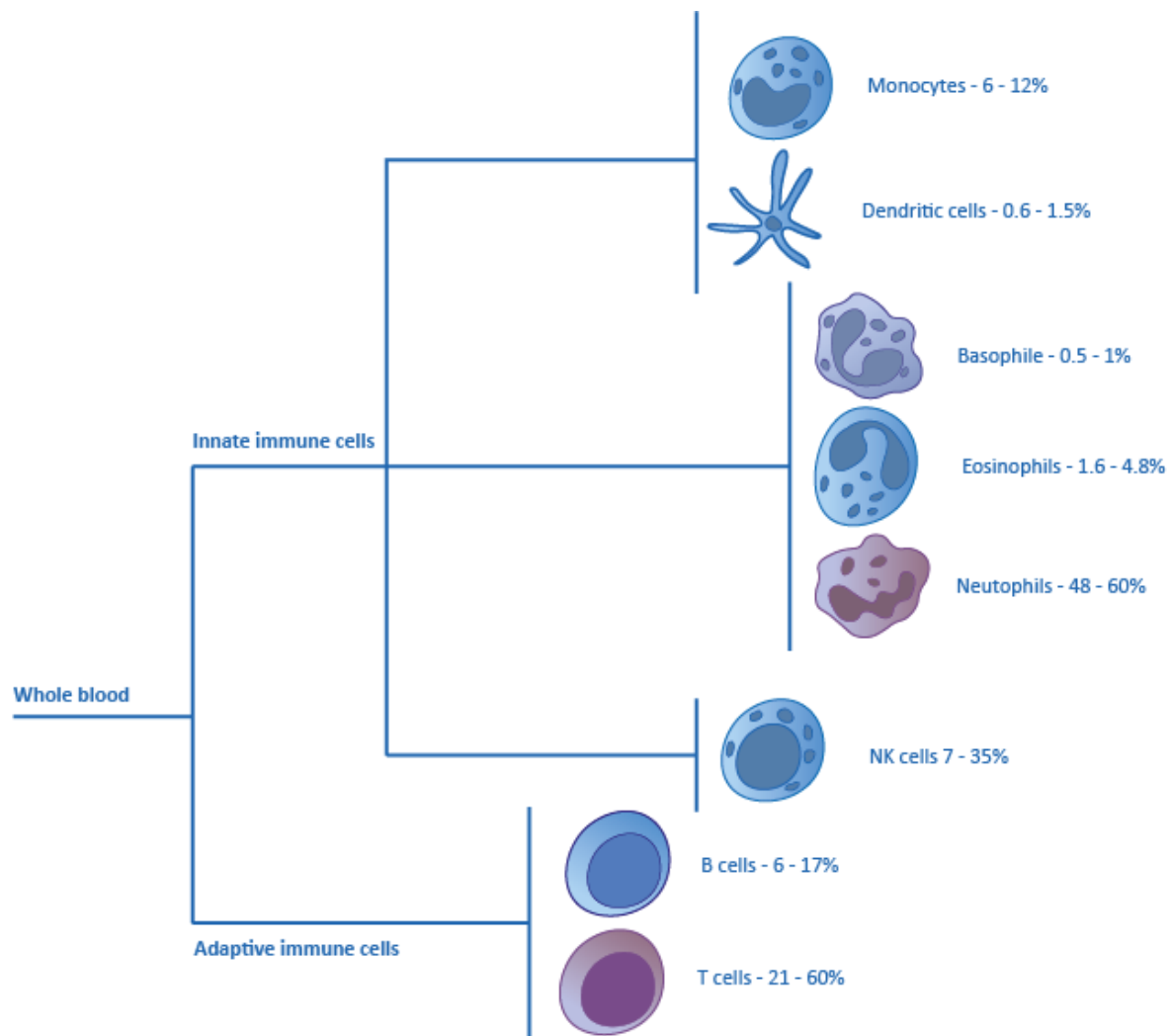


Figure 1.6: The lineage and composition of immune cells in host whole blood. Percentage of white blood cells in whole blood is indicated. Non-immune cells which are present in whole blood are excluded. Values from (Autissier et al., 2010).

This WBA has been used to study different host responses to Mtb infection, including innate and adaptive immune responses. These include investigation of the neutrophil levels in whole blood of pulmonary TB patients and the effect of neutrophil depletion on mycobacterial growth in whole blood (Martineau et al., 2007), the chemokine profiles of whole blood infected with BCG and the effect of infliximab treatment (Newton et al., 2008), the IFN γ response in HIV-infected children compared with healthy controls (Tena et al., 2003), and CD4 T cell responses in whole blood from tuberculin-positive and tuberculin-negative patients infected with BCG. While these studies have provided useful information on the host response to the bacterium in a model encompassing several types of immune

cells, they either briefly investigate the bacterial pathogenesis in terms of bacterial growth in whole blood, or do not investigate this aspect of the infection at all.

One of the first studies to use the WBA model to investigate the host-pathogen interactions investigated gene expression changes in Mtb infection (von Both et al., 2018). They infected peripheral whole blood from volunteers with Mtb *lux*, or the surrogate strain *M. bovis* BCG *lux*, which are bioluminescent strains of Mtb or BCG, respectively. Gene expression changes in Mtb-infected whole blood was measured over 96 hours using microarray. The authors discovered a predominantly down-regulated gene expression profile in infected whole blood samples with 75% of significantly differentially expressed probes having a $\log_2$ fold change ($\log_2FC$) of < -1 , compared to uninfected whole blood. Pathogen recognition and cytokine signalling pathways were increased in expression in infected whole blood at the early time points (up to and including 48 hours post-infection), with genes such as *IRAK1* (signalling molecule pathways), *NF- κ B*, *TNF α* , and *IL-6* (PRR pathways), and *TNF α* , *IL-6*, *IL-8*, *MIP-1 α* , and *MCP-1* (TREM1 signalling pathways) being increased. However, after the 48-hour time point the majority of host genes were suppressed including *TLR-1*, *TLR-4*, *TLR-6*, *DECTIN1*, (all of which are PRRs) *IRAK1*, *CD14*, *DAP12*, (all of which are involved in signalling pathways) and *CD11c* (a cell marker for a subset of macrophages). This change in expression included a broad range of pathways, including Mtb cell surface receptor recognition, phagocytosis, phagolysosomal degradation, chemokine and cytokine signalling, and antigen processing and presentation.

The authors suggest the downregulation of host immune responses may be due to the host immune system responding in down-regulation of inflammation upon infection of Mtb and other pathogens to prevent excessive inflammation and damage. The authors compared the transcriptional profiles of Mtb and BCG with a publicly available dataset of macrophage infection with *Leishmania*, as the mechanism of immunosuppression, to prevent host damage has been studied extensively in the field of parasitic infections. They found a correlation between transcripts from the host and the parasite. However, as these infections are in different systems (whole blood vs macrophage models) additional validation should be performed. This down-regulation of host gene expression during Mtb infection has only been observed in a handful of other studies which use macrophage cell line models at later time points (Brace et al., 2017; Muttucumaru et al., 2004) and a decrease in Th1 cell responses was also observed in a mouse model infected with a hypervirulent lineage 2 Mtb strain HN878 after 14 days post-infection (Ordway et al., 2007). Furthermore, at least some of the downregulation of gene expression observed in this study will be due to cell death as a result of infection or the time of whole

Despite the potential of the WBA to study host-pathogen interactions, there have been mixed results of its use in some studies. One study assessed a WBA to measure inhibition of Mtb replication following primary and secondary doses of the BCG vaccine in a low burden disease setting (Fletcher et al., 2013). They identified no growth inhibition measured through the WBA when compared to a PBMC-based assay which did detect some growth differences between naïve and BCG vaccinated individuals. Similarly, another study investigating growth inhibition of Mtb in whole blood from patients in a high burden setting found no differences between TB infected patients and uninfected patients, although their definitions of TB were based not on culture positivity but on the QuantiFERON-TB Gold In-Tube Assay (Baguma, et al., 2017). Another limitation of the assay is that, while it does contain different immune cell types, as found in whole blood, the similarities of the blood environment as a model of the lung environment has not been validated. Biomarker studies (discussed later) have shown that gene expression in whole blood does reflect host immune responses towards TB infection and can be used to identify TB disease and follow the immune response to treatment and overcoming the infection. These studies suggest that there may be some factors that whole blood can replicate well. Further studies should be performed to compare the immune responses in whole blood to the lung environment to identify which pathways can be accurately investigated using these models.

Overall, there are multiple models used to study TB. These have different advantages and disadvantages which result from logistical issues or are due to the complexity of TB. Depending on the type of analysis and biological question to be addressed, careful selection of an infection model must be identified.

1.2 Omics technologies and data analyses in tuberculosis research

With the advancement of next generation sequencing over the last decade there has been an increased number of studies using sequencing technology in biology and by advances in several other -omics fields including genomics, proteomics, transcriptomics, lipidomics, and metabolomics. These technologies have been incredibly helpful when studying infectious diseases and host-pathogen interactions which require analysis of several different systems. This advancement in technologies has also formed the development of new bioinformatic algorithms for data analysis and data management systems to make better use of these data.

Omic technologies enable researchers to investigate a broad range of biological processes in a relatively small-scale experiment. In this way, all genes, transcripts, proteins, lipids, small molecules, etc., in a system can be measured from one experiment. This allows the investigation of interacting pathways at a specific point in time, depending on the technology used. While genomics experiments can be easy, cheap, and provide insight into all the possible biological processes that a cell can produce, many of the processes identified through genomics may be suppressed or altered by downstream pathways such as during transcription. In contrast, proteomics, lipidomics, and metabolomics provide information on the current processes of a cell, but this type of experiment can be more difficult and expensive to perform.

Within the field of omics analyses there are multiple different types of omics technologies including genomics, epigenomics, transcriptomics, proteomics, lipidomics, and metabolomics. In this thesis, I will discuss genomics, transcriptomics, and metabolomics as these are the focus throughout my PhD. Further omics technologies not mentioned here are beyond the scope of this thesis. Specifically, I will expand on the topics of bacterial genomics and metabolomics to investigate Mtb lineages, host whole blood transcriptomic biomarkers in TB patients, and bacterial and host transcriptomics to understand host-pathogen interactions using dual RNA-seq.

1.2.1 Genomics

Genomics involves the study of the DNA of an organism for example through DNA sequencing. The completion of the first full human genome sequenced as part of the Human Genome Project in 2001 resulted in a significant decrease in cost of genome sequencing and was a pivotal step toward the generation of advanced sequencing technologies and the field of multi -omics (Levy & Myers, 2016). This has led to the development of next generation sequencing technologies which have allowed researchers to sequence the whole genome of samples as routine laboratory procedure. Many genomics experiments focus on genetic variation between samples, which include SNPs which are rare (<1%) differences in sequence bases, inversions, substitutions, deletions, insertions and copy number variations. As a result of this advancement, genomics has been used to study many different aspects of infectious diseases to answer different biological questions from sample variation to evolution.

Applications of genomics to understand Mtb and Mtb lineages

In 1998 the Mtb genome was fully sequenced which allowed the first look into Mtb on a global scale (Cole et al., 1998). The genome of Mtb was found to be 4.4 Mb in length. While genetic differences in Mtb strains were identified in earlier in studies comparing a small number of strains (di Pietrantonio et al., 2011), Mtb genomics was not studied more broadly until genomics became widely affordable.

The use of genomics to understand antibiotic resistance in Mtb, and also in Mtb lineages, is an area of intense research. There was an estimated 200,000 cases of MDR, pre-extremely-drug resistant (pre-XDR), and XDR-TB in 2019 which is an increase from the previous year (WHO, 2020). While this increase may be due to an increase in surveillance and testing, there remains a great need to understand current antibiotic resistance mechanisms to aid with treatment regimens, the development of new antibiotics, and prevent the spread of current drug resistant Mtb strains. Mtb can develop antibiotic resistance through phenotypic resistance, i.e. without genetic changes, and also through genomic variants. For instance, several SNPs have been associated with resistances in Mtb. Mtb resistance to the first and second-line antibiotics (first-line: rifampicin, isoniazid, pyrazinamide, and ethambutol; second-line: fluoroquinolones, injectable aminoglycosides, and thioamides) can be caused by SNPs in relevant genes such as *rpoB* which confers resistance to rifampicin. As a result, studies have continued to investigate SNPs which are possibly linked to resistance in different antibiotics in Mtb lineages. The first of these was published a decade ago and investigated the genomic differences between rifampicin-resistant and rifampicin-susceptible Mtb strains from TB patients

(Comas et al., 2011). The study also grew the rifampicin-susceptible strains for 45 weeks without rifampicin to simulate the evolution of resistance conferring SNPs. They then isolated non-synonymous SNPs (nsSNPs) found only in rifampicin-resistant strains from both datasets and identified 54 putative compensatory mutations in 38 genes and 10 intergenic regions. The genes *rpoA* and *rpoC*, coding for the alpha and beta prime subunits of the RNA polymerase, were found to constitute multiple mutations from the Mtb strains.

Further studies have investigated a greater number of strains although some are specific to certain countries, therefore limiting their wider application. For example, a study sequenced 1,000 genomes from Mtb isolates from Russia including strains from the Beijing, Euro-American, Central Asian, and East African-Indian lineages (Casali et al., 2014). They found that the diversity of the Euro-American lineage strains was higher than the Beijing strains. However, they found that the drug resistance gene locus with the most mutations was codon 315 of the *katG* gene which was most commonly substituted in Beijing isolates than Euro-American isolates. Similarly, SNPs which are linked to rifampicin resistance (present in the *rpoB* gene) were more commonly found in Beijing isolates than Euro-American isolates. A total of 14 nsSNPs were identified in the *rpoA* and *rpoC* genes as well. Another study investigated 161 isolates from China which were almost entirely identified as lineage 2 strains containing both drug-susceptible, MDR and XDR-TB strains (Zhang et al., 2013). They found 15 genes and 13 intergenic regions which showed strong associations with MDR and XDR isolates including *rpoB* and *rpoC* genes.

Additional studies have shown similar findings regarding *rpoA* and *rpoC* genes mutations as well as identifying associations between new genomic regions or variants with drug resistance using new approaches such as PhyC (Farhat et al., 2013) and genome-wide association studies (GWAS) analysis (Oppong et al., 2019). GWAS involve the screening of several thousand genomes to identify genetic variants associated with antibiotic resistance.

There are also several bioinformatic tools and packages which have been developed to predict the functional consequences of SNPs. The simplest type of analysis to investigate genomic findings is to visualise the effect of the SNP on protein structure by either using a known protein structure from the Protein Data Bank (Berman et al., 2000) or by predicting the structure using tools such as Phyre² (Kelley et al., 2015) or the recently published AlphaFold (Jumper et al., 2021). The difference between the

wild-type protein and mutated protein can then be visualised using modelling software such as PyMOL (<https://pymol.org/>). Additional analyses can be performed using tools ranging from homology-based scoring of deleterious effects of SNPs using SIFT (Sim et al., 2012) or PROVEAN scores (Choi & Chan, 2015), to structural impact prediction tools such as MISSENSE3D (Ittisoponpisan et al., 2019).

Due to the importance of understanding and preventing antibiotic resistance in TB treatment, the majority of studies investigating the implication of genomic alterations in Mtb lineages focus on their impact on drug resistance. Interestingly, one of the few studies to not predominantly focus on drug resistance conducted a systematic literature review of previously published articles which used whole genome approaches to investigate Mtb within-host microevolution (Ley et al., 2019). The studies used in this review were predominantly focussing on drug resistance but also included heterogeneity, diversity, transmission, outbreak, and relapse versus infection studies. Several of these studies investigated the functional consequence of the variants they identified. The majority of the functional analysis studies identified variants in genes predicted to be relevant for drug resistance and compensatory mechanisms. One common example were variants in the *fadD*, *fadE*, or *pks* gene families which are involved in cell wall biosynthesis pathways and were found in several studies to which the authors suggested may be linked to fitness compensation or drug resistance. An important finding was that only 11 out of the 38 studies used in the review discussed the potential functionality of the variants identified.

Applications of genomics in understanding the host response to TB disease

Host genetics also plays a significant role in TB disease, as exemplified by the differences in infection observed in different mouse strains (Bucsan et al., 2019). It is well-known in TB research that most mouse models, including the commonly used C57BL/6 strain, cannot replicate the necrotic granulomas observed in human disease. However, there are other mouse strains which have been developed to replicate certain conditions such as the granuloma, in strains such as the C3HeB/FeJ, B6.C3H^{sst1}, and CBA/J IL-10 KO mice. In the human host, the first known genetic variant associated with increased risk of mycobacterial infections was identified in 1996 (Newport et al., 1996). The variant was found in a region of the IFN γ receptor 1 (*IFNAR1*) gene. While sequencing technologies were not available then, the group used microsatellite analysis, northern blot analysis, and PCR amplification to identify a substitution which introduced a stop codon in the gene sequence, leading to loss of function of the *IFNAR1* gene. This led to these patients showing an increased susceptibility

to all mycobacterial infections and highlights the importance of the IFN α pathway in controlling Mtb infection.

Since then, larger genomic GWAS studies have identified further variants linked to increased susceptibility to TB in genes such as *WT1*, *SMOC1*, and *ASAP1* (Alessandra et al., 2018). In addition, associations of variants with increased protection from TB have also been found. For example, recently a variant in a regulatory region of the *IL12B* gene was significantly associated with resistance to TB in individuals living in Tanzania and Uganda which were hyperendemic for TB (p-value of 2.11×10^{-8}) (Sobota et al., 2016). *IL12B* encodes for the p40 subunit of IL-12 which is involved in the differentiation of Th17 cells and IFN γ production from these cells. IL-12 and its effects have been linked to Mtb disease, including use of IL-12 as a treatment in mice to protect against TB. This GWAS study suggests that the use of IL-12 or targeting cells involved in IL-12 stimulation may lead to new treatments for TB disease which can be used in humans.

As previously discussed, different Mtb lineages can produce different levels of disease or stimulate different immune responses in the host and several studies have hypothesised that differences in the genomes of Mtb lineages may cause these differences in immune response. One study identified two *TLR9* single nucleotide variants were in strong linkage disequilibrium with increased susceptibility to TB and TB meningitis (Graustein et al., 2015). This was associated with the Mtb variant rs352142 which was in turn associated with Mtb strains from lineages 1 and 4. Several other studies have linked Mtb variants associated with certain lineages which increase the risk of susceptibility to disease which are in linkage disequilibrium with different host genes (Correa-Macedo et al., 2019). In addition to these studies, protective variants have been identified, such as the rs9637876 variant found in the 5'UTR of the *IRGM* gene which is protective against lineage 4 Mtb strains. While these findings highlight important host genes related to specific Mtb lineages, these studies are often limited by the lack of biological investigation into why these associations have evolved and why they cause increases in susceptibility, or protection, towards specific lineages. Despite this, this is one of the few studies which attempt to understand the interactions between the host and Mtb. The majority of previous studies, as described below, focus on one side of the biology.

While the studies using GWAS have been extremely helpful in identifying key biological pathways for both susceptibility to and protection against TB, there are limitations to GWAS analysis. The main

limitation is the requirement of a large sample size to allow the identification of variants (Hong & Park, 2012). Another concern is the lack of identification of common variants (>1% of the population) linked to TB disease, which has been suggested to be due to genetic heterogeneity in host populations (Correa-Macedo et al., 2019b). Similarly, there are currently no associations between genetic variants and *Mtb* lineages, possibly due to low sample sizes. In addition, the associations of variants with TB must be functionally validated to allow for further biological insight into the implicated pathways involved in the disease.

1.2.2 Transcriptomics

Transcriptomics consists of the study of the transcriptome, which includes different RNA species such as messenger RNA (mRNA), ribosomal RNA (rRNA), small noncoding RNAs (sRNAs), long noncoding RNAs (lncRNAs), and micro RNAs (miRNAs). RNAs have different functions within the cell, with mRNAs encoding protein sequences, while rRNAs aid in translation of mRNAs into protein, and several RNAs are involved in regulation of transcription such as lncRNAs. As RNA has such a large array of functions, the study of transcriptomics has been utilised for a broad range of applications particularly with the increasing accessibility of microarray and RNA-seq and an increase in bioinformatic analysis tools and workflows available. The majority of current transcriptomics studies use PCR, microarray, or RNA-seq technologies to measure different aspects of the transcriptome, with the latter two being able to measure hundreds or thousands of transcripts at once.

Microarrays quantify gene expression through fluorescence intensity by measuring fluorescently-tagged complementary DNA (cDNA) molecules which attach to probes of known sequence on a chip. RNA-seq directly sequences the RNA or cDNA molecule itself and is therefore a more quantitative method. Unlike microarrays, RNA-seq can measure any transcript including those which have not yet been characterised, thereby being able to measure novel transcripts. While microarray chips targeting non-coding RNAs and miRNAs do exist, they are limited to those which are already characterised, whereas RNA-seq has enabled the investigation and discovery of novel non-coding and miRNAs (Han et al., 2015). In addition, transcriptomics can provide information on RNA splicing events. RNA splicing is the removal of intronic regions of a precursor mRNA molecules, to splice together the exons of the precursor mRNA, to form a mature mRNA (Wilkinson et al., 2020). RNA-seq can measure novel splicing events and can identify alternative splicing events through sequencing of transcription start sites and

can also identify new SNPs (Han et al., 2015). With developing bioinformatic tools, the applications of RNA-seq will continue to increase.

Applications of transcriptomics to understand Mtb

A common finding across Mtb transcriptomic studies is the gene expression changes in genes involved in Mtb metabolism, and particularly lipid metabolism. After sequencing the Mtb genome in 1998, it was revealed that a substantial portion of the Mtb genome encodes for molecules involved in lipid metabolism, particularly fatty acid metabolism with over 250 enzymes encoded in the Mtb genome compared with 50 in *Escherichia coli* (Cole et al., 1998). This lipid diversity was not explored fully until gene expression studies were performed.

Transcriptomics studies have built on the findings from genomics through identification of key Mtb specific enzymes such as isocitrate lyase, a virulence factor for Mtb and involved in Mtb metabolism of fatty acids (Muñoz-Elías & McKinney, 2005). Rachman, et. al., investigated the transcriptome of Mtb isolates from different tissue sites of TB patients, including lesions from granulomas, pericavity, and distant-lung and compared this to an *in vitro* culture of Mtb (Rachman et al., 2006). They identified that Mtb undergoes cell envelope remodelling in all datasets used but to a greater extent in the *in vivo* datasets, where twice as many genes involved in lipid biosynthesis were upregulated *in vivo* compared to *in vitro* including genes involved in fatty acid and mycolic acid modification genes. Metabolic changes have also been identified in *in vitro* models, such as one study which cultured Mtb in media containing cholesterol and fatty acids (which is similar to the macrophage phagosome environment) and compared this with Mtb grown in media containing dextrose (the standard carbon source for bacterial culture) (Aguilar-Ayala et al., 2017). Several lipid metabolism genes increased in expression including *ltp2*, *pks2*, and several *fad* genes.

In addition to metabolism, many transcriptional studies investigate the macrophage response to Mtb using mouse models, cell line models, or *in vitro* models (Fontán et al., 2008; Pisu et al., 2020b; von Both et al., 2018). One of the first studies to do this identified the large induction of genes (454 compared to 147 repressed genes) in murine BMDMs between 4 and 24- or 48-hours post-infection which was suggested to be associated with Mtb entry to the phagosome and the escape from the host defence mechanisms in this environment, including *icl* (Schnappinger et al., 2003). These include genes involved in transcriptional regulation, intermediary metabolism, lipid metabolism, and iron

acquisition. Additional *in vitro* models have shown an overall increase in gene expression in early (< 24 hours) infection (Rohde et al., 2007). Sharma, et. al., isolated Mtb from sputum samples of pulmonary TB patients as an *in vivo* model for investigating transcriptional differences compared to patients with lung cancer (Sharma et al., 2017). This analysis identified that the majority of genes were downregulated (394 compared to 163 genes) including genes involved in cell wall and cell processes and intermediary metabolism and respiration.

This contrast between cell models, which identified that the majority of genes are upregulated during infection, and human *in vivo* models, which have shown a downregulation of genes during infection could reflect the fact that these TB patients may have been living with TB for an extended period of time, whereas most *in vitro* models investigate shorter time points. This highlights that the choice of model, and the time-point used, may significantly impact the interpretation of the results. This also suggests that it is difficult to model true infection which may represent a more chronic infection phenotype using *in vitro* studies. To address this, there has also been extensive research on dormancy in Mtb infection and its relation to LTBI and the persistence of Mtb in different models.

Another consideration in using transcriptomic analysis to understand Mtb infection is the sample preparation. Unlike other bacteria, Mtb has a particularly lipid-rich cell wall which results in it being difficult to extract RNA from. Therefore, an important step in any Mtb RNA extraction protocol is to break open the Mtb bacilli. This is usually performed using a bead-beating step which mechanically breaks down the Mtb cells (Benjak et al., 2015). However, as RNA is sensitive to higher temperatures and degrades rapidly when handled, this extra bead-beating step adds difficulty in processing Mtb samples for RNA extraction. In addition, the bioinformatic tools available for pre-processing and analysis of Mtb transcriptomic data are limited compared to those available for human transcriptomic datasets. Many Mtb genes are uncharacterised resulting in difficulty in generating a biological interpretation of gene expression changes in many genes. Another limiting factor is the lack of a curated pathway analysis database for the Mtb transcriptome. While there are many tools and packages to perform pathway analysis using human transcriptomic datasets such as Ingenuity Pathway Analysis and gene set enrichment analysis (GSEA), there are fewer available tools for Mtb analyses (Krämer et al., 2014; A. Subramanian et al., 2005).

Transcriptomics has recently become accessible to many groups, leading to the greater number of studies of both host and Mtb. While these studies provide significant insight into disease, one major limitation is that almost all of the differentially expressed transcripts in studies focussing on Mtb involve metabolism, lipids, or proteins which are then often not validated through metabolomics or proteomic studies to show which of the changes in transcription result in changes in downstream molecules. It is therefore of great benefit to design and perform validation studies or multi omics analyses to address this issue.

The development of dual RNA-seq and its application to studying TB disease

Building on the depth of information which RNA-seq offers, in the last 10 years groups have also developed dual RNA-seq, a technique which allows the sequencing of more than one organism in a single sample. This technique builds on RNA-seq as it provides an insight into the communication and interaction of multiple cells in one system.

There are unique challenges associated with performing a dual RNA-seq experiment. One of the largest concerns when designing a dual RNA-seq experiment in infectious disease studies is the proportion of pathogen to host RNA available in each model. Some infectious diseases can produce high pathogen burden quickly within the host, such as *Salmonella enterica* infection, whereas others will have much lower pathogen burden. The latter is the case for Mtb which is a slow-growing bacteria and has a 24-hour doubling time. It may be beneficial to use a later time point of infection to increase the pathogen load, however depending on the model used this may also increase the amount of cell death which would dilute the gene signals measured and may not be representative of *in vivo* infection. For most infections, regardless of pathogen replication rate, either the pathogen must be enriched, or the read depth increased to obtain enough RNA for sufficient sequencing coverage. There are different methods to achieve pathogen RNA enrichment, including using laser capture microdissection, FACS through fluorescence labelling or cell wall binding dyes (Westermann et al., 2017), or differential lysis (Rienksma et al., 2015).

One of the first studies to perform a dual RNA-seq experiment utilised macrophage-like THP-1 cells infected with BCG to study transcriptional changes after 24 hours (Rienksma et al., 2015). They found an increased expression of cholesterol degradation and iron acquisition in BCG during infection and, as a potential response to this, an increase in *de novo* cholesterol synthesis expression in infected host

cells. As one of the first dual RNA-seq studies carried out, it revealed many of the design considerations which would be addressed in further studies. The study itself only used 8 biological replicates, with 2 being enriched for BCG through differential lysis of host and mycobacterial cells. Further limitations included the use of BCG and not validating in Mtb, use of one cell line, and the lack of validation of the gene expression findings by proteomics or metabolomics studies.

Westermann, et. al., 2016 develop dual RNA-seq further through using it to study *Salmonella enterica* serovar Typhimurium infection in HeLa cells (Westermann et al., 2016). The authors used GFP-expressing *Salmonella* to infect HeLa cells and isolated infected cells using FACS to enrich the pathogen RNA. They identified several significant differences in gene expression changes between infected and uninfected cells, specifically small regulatory RNAs (sRNAs) after they found that PhoP-induced sRNA in intracellular *Salmonella* (PinT) was increased by 100-fold during infection. PinT was positively correlated with the *Salmonella* pathogenicity island (SPI-2) and Janus kinase/signal transducers regulon. They then identified several alterations in the host immune response, such as an increase in early activation of the activators of transcription (JAK-STAT) signalling pathways such as SOCS3 in PinT mutants compared to the wild-type. Overall, the authors identified novel transcriptomic signals in host inflammatory pathways, identified the mode of action for these signals, and validated this through mutant models of infection. This initial dual RNA-seq study highlighted the power of such an experimental method and paved the way for others to use this technique to investigate additional infectious diseases.

Since then, groups have used the dual RNA-seq approach in different infection models, including to study Mtb infection. Another study integrated dual RNA-seq with metabolomics analysis in a human macrophage cell model (Zimmermann et al., 2017). The study infected THP-1 cells with Mtb and performed untargeted metabolomics analysis and dual RNA-seq analysis at different time points including 0-, 4-, 24-, and 48-hours post infection. For metabolomics experiments, to differentiate between host metabolism as a result of general phagocytosis rather than Mtb phagocytosis specifically, they used gamma-irradiated tubercule bacilli which do not have an active metabolism. They identified a large abundance of mycobacterial lipids such as monoacyl glycerols and phosphatidylinositol phosphates. Furthermore, a decrease in amino acids, nucleotides, and sugar phosphates was present which they suggest may be due to an uptake of these molecules by Mtb. Investigation of the host transcriptome identified an upregulation of infection associated pathways, including antigen processing and presentation, in the host. Similarly to previous studies, they

identified higher expression of catabolic pathway genes and cholesterol degradation intermediates specific to Mtb. As a possible response to this increase in cholesterol degradation by the bacterium, host steroid biosynthesis pathway genes were increased in expression. As mentioned, the cholesterol degradation product propanoyl-CoA is toxic to Mtb. Increased expression of the methylcitrate cycle was also present which degrades propanoyl-CoA.

One study used dual RNA-seq to investigate *Mycobacterium leprae* by analysing lesions from leprosy patients (Montoya et al., 2019). The authors used weighted-gene correlation analysis (WGCNA) to discover groups of correlated genes in bacterial and host transcriptomes in samples. A more recent study was the first to use an *in vivo* C57BL/6J mouse model to investigate Mtb infection in alveolar and monocyte-derived interstitial macrophages (Pisu et al., 2020b). This was the first instance where more than one cell type was used, although again the focus of this study was the macrophage cell. After 14 days of infection, fluorescently tagged alveolar and interstitial macrophages were FACS sorted from mice lungs, host and pathogen RNA was separated using differential lysis, RNA was extracted separately before an equal proportion of host and pathogen RNA was combined and dual RNA-seq was performed. Differential gene expression analysis was undertaken, comparing several tools, and also the *in vivo* results with differentially expressed genes from Mtb grown in broth media, Mtb infected BMDMs at 6 days post-infection, and other publicly available BMDM datasets. They identified 180 genes which were significantly differentially expressed in *in vivo* infection compared to the other datasets. Pathway analysis of these genes confirmed findings from earlier Mtb gene expression studies with an upregulation in cholesterol homeostasis, nitrogen assimilation, NO detoxification, and protein and lipid export machinery.

In addition, the authors identified differentially expressed genes between alveolar and interstitial macrophages (IMS) to understand their different immune responses and why Mtb preferentially infects alveolar macrophage (AMs). Using GSEA, they identified macrophage phenotype specific pathways which were regulated upon infection. The uninfected IMs showed a proinflammatory state, with an upregulation of pro-inflammatory pathways including NF- κ B and Th1 responses, and adhesion and chemotaxis pathways. Infected IMs increased further proinflammatory pathways such as ubiquitination, NO and ROS biosynthesis, and inflammatory responses were upregulated. These data suggested that IMs are already differentiated towards a pro-inflammatory phenotype. The AMs showed an upregulation of a larger number of fatty acid pathways upon infection than IMs including cholesterol and fatty acid biosynthesis, although these were downregulated compared to uninfected

AMs. The authors suggest that Mtb may predominantly infect AMs as a result of this increase in fatty acid metabolism compared to IMs.

Finally, a recent study used dual RNA-seq to investigate sputum from patients with and without active TB (Lai et al., 2021). Unlike previous studies, no enrichment of bacterial RNA was carried out. The authors identified 5,843 host genes which were differentially expressed between TB and non-TB patients. These genes corresponded to mostly up-regulated pathways in TB patients (9 out of the 11 pathways identified through GSEA). These up-regulated pathways included inflammatory responses stimulated by IFN γ , TNF α , and IFN α/β which confirms the findings of previous studies. There was also a significant difference in the metabolism response in the host in TB patients compared with non-TB patients. This was characterised by a decrease in expression of genes in the TCA cycle and the electron transport chain in TB patients. In the pathogen, a similar transcript profile was observed as has been found in cell culture models and human and animal infected tissues. The microbiome profiles in TB patients compared with non-TB patients was also different, with those without TB having a much higher number of microbiome species in their samples. However, patients with TB had very few Mtb reads available for analysis. The Mtb transcriptional profile was compared to Mtb culture transcriptomes, with several gene pathways being identified as similar to different Mtb culture stages such as stationary and exponential phases. Compared to exponential phase Mtb culture, sputum Mtb showed significantly higher expression of the KstR and KstR2 regulons which are involved in cholesterol metabolism.

The advent of dual RNA-seq is a new and exciting approach to study host-pathogen interactions in different models, and several groups have applied this towards Mtb infection. This approach has allowed for a more global understanding of disease as it can describe the responses of both host cells and Mtb towards infection. The current studies have used different experimental designs and sample preparation steps, to overcome specific challenges which dual RNA-seq brings. However, while different models have been utilised, all the current studies have again focussed on the interactions of Mtb with the macrophage cell. Further studies should aim to understand the roles of other cells, alongside macrophages, to fully understand host-pathogen interactions during infection.

Applications of transcriptomics to understand the host response to TB

Similarly to Mtb transcriptomics, host transcriptomics of TB has also utilised different infection models to answer distinct biological questions. Cell lines can again be used to model short-term cell-specific responses towards Mtb. As the most studied cell type of TB, several studies have investigated the transcriptional changes which occur during macrophage infection with Mtb. Of the pathways which are upregulated, Mtb infection in macrophages has been shown to increase the expression of genes involved in immunity and immune regulation, such as *IL-10*, *IL-8*, and inflammation such as *TNF α* , *IL-1 β* , and *IL-6* (Volpe et al., 2006). Furthermore, while some of these cytokines may be increased upon Mtb infection of human monocyte derived macrophages, the expression of certain cytokines might differ over longer infection times. They found that *IL-8* for instance, decreased after 3 days of infection, while *IL-1 β* remained highly expressed for at least 7 days. This has implications for further studies, as the immune response may vary between immunosuppressive responses to inflammatory at different points in time.

In addition, several studies have shown a decrease in certain host responses attributed to Mtb manipulation or suppression mechanisms. Schnappinger, et. al., found that Mtb infection stimulated lower levels of IL-12p40 expression, and protein abundance, than other bacteria (Schnappinger et al., 2006). In another study, heat-killed Mtb stimulated lower expression of IL-12p40 than live Mtb in primary mouse macrophages. IL-12 is involved in the Th1 response to Mtb and is critical for control of the infection. Together, this suggests that through transcriptomics analyses Mtb may suppress IL-12 production to prevent the activation of the adaptive immune response. Further studies have also identified that Mtb lipoproteins can inhibit this response, suggesting that even dead bacilli can prevent IFN γ signalling in the host.

In contrast, *in vivo* infections such as mouse models can be more reflective of true infection through allowing the communication of complex networks of cells and can also be used to model longer-term phenotypes. These studies, however, have again focussed on the macrophage cell, specifically through isolation and analysis of alveolar macrophages. Rothchild, et. al., found an upregulation of 131 genes and a downregulation of 65 genes in infected AMs compared to uninfected AMs isolated from mice BAL fluid at 24 hours post-infection (Rothchild et al., 2019). This confirms other studies in whole blood which also found an overall upregulation of genes early in infection (von Both et al., 2018). However, the authors used pathway analysis to identify a large increase in the NRF2-mediated oxidative stress response, but not other proinflammatory genes such as *TNF*, *IL1 β* , and *CXC9*, which was not previously shown and is in direct contrast with previous *in vitro* studies which show early proinflammatory

responses (Rothchild et al., 2019). They did show that proinflammatory pathways were upregulated by day 10 post-infection which was similar to gene expression changes in *in vitro* models using BMDMs. In contrast, the NRF2-associated genes were not increased in BMDMs, indicating that certain immune pathways, including this one, can only be modelled *in vivo* or through recreating an *in vivo* environment.

In addition to cell line and animal models is the use of patient whole blood. While TB is predominantly an infection of the lung, gene expression studies of whole blood have identified TB-specific ‘signatures’ of genes. Furthermore, these signatures partly reflect the immune response to infection as there are also several immune cells present in whole blood. As a result of this, and the ease of taking blood samples from patients, there has been a growing interest in host transcriptomics to develop a TB biomarker signature for the development of new diagnostic tests.

DiNadro, et. al., used a novel approach to investigate host endotypes during Mtb infection using previously published whole blood biomarker signatures (DiNardo et al., 2021). They used an unbiased clustering analysis of publicly available gene expression datasets to identify TB phenotype clusters which correspond to different disease endotypes. They defined two larger endotypes, endotype A and endotype B, which included almost all the patients, of which all HIV infected patients and paediatric TB patients were grouped into endotype A. Upon pathway analysis of the genes expressed within each endotype, endotype A showed an upregulation of type I and type II IFN, TNF, IL-2-STAT5 signalling, glycolysis, oxidative phosphorylation, and Myc targets. In contrast, endotype B showed decreased gene expression signalling of IFN γ signalling pathways and decreased expression of TNF and TNF via NF- κ B signalling pathways. The study also focussed on metabolic differences in the endotypes. They found that endotype A showed increased oxidative phosphorylation, glycolysis, the tricarboxylic acid (TCA) cycle activation, and glucose metabolism. In contrast, in endotype B all these metabolic pathways were downregulated. This is the first study to explore the possibility of signatures representing different disease subgroups in such an unbiased analysis and is one of few studies to investigate a broader transcriptional response to Mtb.

Overall, transcriptomics has been an incredibly useful tool for identifying and understanding important genes and pathways involved in the host response to infection, as well as the mycobacterial response to the host immune system. In addition, transcriptomics studies have built on previous

genomics findings and supplement well further omics technologies, such as the identification of the Mtb metabolism gene, *icl* (discussed below). Transcriptomics does come with limitations, however. While traditional sequencing experiments can be straightforward, some approaches are not yet standardised, such as dual RNA-seq, leading to difficulty in performing such experiments. Nevertheless, the applications of transcriptomics are vast and as the technology continues to become more affordable than ever, it is an extremely useful tool for investigating Mtb infection.

1.2.3 Host whole blood transcriptomic biomarker signatures for TB

Transcriptomics has also been used for biomarker analysis, usually with the aim of either creating diagnostic tests for specific diseases or generating markers for disease subsets or to follow disease progression (McNerney et al., 2012). These studies often focus on the host transcriptomic differences in infected versus uninfected patients. The first of such studies investigating TB disease identified a distinct expression profile of infected patients which included 393 gene signature distinguishing TB from healthy controls and a neutrophil induced IFN γ response in infected patients (Berry et al., 2010). Since then, many studies have identified smaller signatures including single gene signatures, with varying success in terms of accuracy (Turner et al., 2020), which can be converted to diagnostic tests more readily than larger signatures. Although these signatures are promising developments to modern diagnostic tests for TB, many were designed from single-country cohorts and show mixed effectiveness in diagnosing patients in other regions (Leong et al., 2020; Turner et al., 2020) and many signatures have not been validated in independent datasets (Jenum et al., 2016; Q. Li et al., 2015; Zhou et al., 2016).

In 2014 the WHO set out target product profiles (TPPs) which are a set of guidelines for different types of new TB diagnostic tests (**Table 1.2**) (WHO, 2014). These included a non-sputum based rapid test targeting new transcriptomic biomarker tests which were being published at the time. These guidelines emphasised the requirement for a novel test to improve on factors such as sensitivity, specificity, and cost. While a test with high sensitivity and specificity would be the ideal, there are compromises that can be made depending on the type of test required. If a test is being designed to rule in TB then high specificity is required, which is necessary to begin antibiotic treatment as TB symptoms overlap with many virus and parasitic diseases. In contrast, high sensitivity is required to rule out TB (when testing negative), which is useful in settings where further tests can be performed

for those testing positive. Currently, the Gene Xpert produces high specificity and can therefore be used for confirmatory testing depending on the setting.

Table 1.2: WHO guidelines for TPPs for triage and confirmatory non-sputum based diagnostic tests for TB (WHO, 2014). N/A = not applicable.

		Adult TB		Paediatric TB	
		Sensitivity	Specificity	Sensitivity	Specificity
Triage test	Minimum	90	70	N/A	N/A
	Optimum	95	80	N/A	N/A
Confirmatory test	Minimum	65*	98*	72	87
	Optimum	95	80	85	92

* For pulmonary TB in adults.

The adult biomarker signatures have recently been independently validated on an independent South African dataset (Turner et al., 2020). A main focus of this study was to identify whether any of the signatures meet the TPPs set out by the WHO for novel non-sputum based diagnostic tests. They evaluated 27 signatures derived from adult cohorts of different countries of origins for their performance on a South African cohort comparing TB patients (culture or gene Xpert positive) with patients with other diseases (culture or gene Xpert negative). However, they found that none of the signatures met the optimum WHO guidelines for a triage (95% sensitivity and 80% specificity) or confirmatory test based on sensitivity and specificity (65% sensitivity and 98% specificity), and therefore none have the potential for development into a diagnostic test used clinically.

In addition to the challenges faced in discovering a biomarker signature for adult TB, there have been fewer developments in diagnostic biomarker signatures for paediatric TB. The lack of an effective diagnostic test in paediatric TB has contributed to the 12% of deaths from TB in 2019 (144,000 children less than 15 years old) (WHO, 2020). Unlike adult TB, paediatric TB is paucibacillary in nature, meaning that there are fewer bacilli causing the disease compared to adult TB. This results in smear culture testing often being ineffective as often bacilli cannot be isolated from samples and acquiring smear samples in children is also challenging (Nicol & Zar, 2020).

Only 3 studies have been published which focus on identifying a host transcriptomic biomarker signature for development into a diagnostic test specifically designed for paediatric TB (**Table 1.3**) (Anderson, et al., 2014; Gjøen et al., 2017; Verhagen, et al., 2013). The first of these studies used a small sample group size (9 TB patients, 9 LTBI patients, and 9 healthy controls) of Warao children from

Venezuela and identified a 116 gene signature using a random forest algorithm which discriminated between the disease groups (Verhagen, et al., 2013). The authors then sought to validate their signature on other cohorts, but as this was the first paediatric cohort published, they used adult cohorts for validation. Their paediatric signature was found to perform well across all comparisons and in most cohorts (average prediction error rate of 19.2% for TB vs LTBI and 28.2% for TB vs LTBI vs healthy controls). In contrast, the signatures from the adult cohort studies performed poorly on their paediatric cohort when discriminating TB from LTBI and healthy controls (average prediction error rate ranging from 38 – 44.4%). They further identified a smaller 10 gene signature from their cohort and the adult cohorts used in their validation analysis using feature selection random forest methods which was validated on the same paediatric samples using qRT-PCR. This signature showed good predictive power with 5 of the genes classifying 78% of the TB patients correctly with 100% specificity. While this was a significant start in paediatric biomarker testing, the study used a small sample size of a niche cohort, combined their cohort with adult cohorts and therefore possibly diluting the paediatric signal when identifying a small gene signature, and did not validate this on an independent cohort or show the statistics for the full 10 gene signature.

Table 1.3: Host blood transcriptomic biomarker signatures for TB.

HIV = human immunodeficiency virus; TB = tuberculosis; OD = other diseases; DRS = disease risk score; LTBI = latent TB infection; HC = healthy controls; SVM = support vector machine model; RF = random forest model; UK = United Kingdom; USA = United States of America. * Unspecified HIV status

Signature name	Reference	Target group	Case	Control	Size	Platform	Signature use	Country
Anderson44	Anderson, et. al., 2014	Children (HIV+/-)	TB (n=87)	OD (n=43)	51 genes	Microarray	DRS	South Africa, Malawi
Anderson38	Anderson, et. al., 2014	Children (HIV+/-)	TB (n=87)	LTBI (n=134)	42 genes	Microarray	DRS	South Africa, Malawi
Araujo1	de Araujo, et. al., 2016	Adults*	TB (n=6)	HC and LTBI (n=28)	NPC2	RNA-seq	N/A	Brazil
Araujo3	de Araujo, et. al., 2019	Adults*	TB (n=8)	LTBI (n=21)	4 genes	RNA-seq	SVM	Brazil
Duffy10	Duffy, et. al., 2019	Adults (HIV+/-)	TB (n=93)	LTBI and OD (207)	10 genes	Microarray (from GSE37250 study)	RF	South Africa
Gjoen11	Gjoen, et. al., 2017	Children (HIV-)	TB (n=47)	Healthy controls and OD (n=36)	10 genes	dc-RT MLPA	LASSO regression	India
Gliddon3	Gliddon, et. al., 2019	Adults (HIV+/-)	TB (n=155)	OD (n=140)	3 genes	Microarray	DRS	South Africa, Malawi
Gliddon4	Gliddon, et. al., 2019	Adults (HIV+/-)	TB (n=155)	LTBI (n=167)	4 genes	Microarray	DRS	South Africa, Malawi
Huang11	Huang, et. al., 2015	Adults (HIV-)	TB (n=16)	Healthy controls and OD (n=79)	13 genes	Microarray	SVM	UK
Kaforou24	Kaforou, et. al., 2013	Adults (HIV+/-)	TB (n=151)	LTBI (n=77)	27 genes	Microarray	DRS	South Africa, Malawi
Kaforou33	Kaforou, et. al., 2013	Adults (HIV+/-)	TB (n=151)	OD (n=89)	44 genes	Microarray	DRS	South Africa, Malawi
Maertzdorf4	Maertzdorf, et. al., 2016	Adults (HIV-)	TB (n=60)	LTBI and HC (n=40)	4 genes	Microarray and RT-PCR	RF	India
Maertzdorf12	Maertzdorf, et. al., 2016	Adults (HIV-)	TB (n=60)	LTBI and HC (n=40)	15 gene	Microarray and RT-PCR	RF	India
Penn6	Penn-Nicholson, et. al., 2019	Adolescents (HIV-)	Progressors (n=46)	Non-progressors (n=107)	6 genes	RNA-seq	RISK6geo score	South Africa
Qian17	Qian, et. al., 2016	Adults (HIV-)	TB (n=16)	Healthy controls and OD (n=69)	17 genes	Microarray	Additive	UK
Roe1	Roe, et. al., 2016	Adults (HIV-)	TB (n=46)	Post recovery (n=31)	BATF2	Microarray	N/A	UK
Roe4	Roe, et. al., 2016	Adults (HIV-)	TB (n=23)	OD (n=35)	4 genes	Microarray	SVM	UK
Roe5	Roe, et. al., 2016	Adults (HIV-)	TB (n=23)	OD and post recovery samples (n=50)	5 genes	Microarray	SVM	UK
Roe2	Roe, et. al., 2020	Adults (HIV-)	TB (n=46)	Post recovery (n=31)	3 genes	Microarray	SVM	UK
Singhania20	Singhania, et. al., 2018	Adults (HIV-)	TB (n=NS)	LTBI and OD (n=NS)	20 genes	Microarray	Modified DRS	UK
Suliman4	Suliman, et. al. 2018	Adults (HIV-)	Progressors (n=45)	Healthy household contacts (n=141)	4 genes	RNA-seq	(GAS6+SEPT4) - (CD1C+BLK)	South Africa, Gambia
Suliman2	Suliman, et. al. 2018	Adults (HIV-)	Progressors (n=79)	Healthy household contacts (n=328)	2 genes	RNA-seq	ANKRD22/OSBPL10	South Africa, Gambia, Ethiopia
Sweeney3	Sweeney, et. al., 2016	Adults (HIV+/-)	TB (n=296)	LTBI and OD (n=727)	3 genes	Multicohort analysis	(GBP5+DUSP3)/2-KLF2	South Africa, UK, USA, Malawi, France
Verhagen8	Verhagen, et. al., 2013	Children (HIV-)	TB (n=9)	HC and LTBI (n=18)	10 genes	Microarray	RF	Venezuela
Walter45	Walter, et. al., 2016	Adults (HIV-)	TB (n=24)	LTBI (n=24)	51 genes	Microarray	SVM	USA
Walter35	Walter, et. al., 2016	Adults (HIV-)	TB (n=24)	OD (pneumoniae, n=24)	47 genes	Microarray	SVM	USA
Zak16	Zak, et. al., 2016	Adolescents (HIV-)	Progressors (n=37)	Healthy controls (n=77)	16 genes	RNA-seq	SVM	South Africa

The second study investigating paediatric TB signatures used a larger sample size (110 TB patients, 159 other disease patients, and 52 latent TB patients) of children from South Africa and Malawi (Anderson, et al., 2014). These were split into training and testing datasets where the training dataset was used to discover a 51 gene signature discriminating TB from other disease patients and a 42 gene signature discriminating TB from LTBI using the elastic net feature selection method and generated a Disease Risk Score for each patient. These performed well with an area under the receiver operator characteristic curve (AUC) of 86.2% when discriminating between TB and other disease patients in the testing dataset and a higher AUC of 89% in an independent validation cohort. It is also important to note that this study included HIV infected and uninfected individuals in their signature discovery, which is unique compared with most other studies, including those using adult cohorts. It is unclear if a TB signature for both HIV infected and uninfected individuals is necessary or if including this in the biomarker discovery model will dilute the signal and lower the accuracy of the signature derived from a mixed HIV cohort. However, this study used a larger sample size and independently validated their signatures. Despite this, their signatures are still large which would prove difficult to convert to an affordable biomarker test.

The final study using a paediatric cohort again selected a training and test dataset from a pool of 71 TB patients, 36 asymptomatic healthy control patients, and 26 symptomatic non-TB patients from India (GjØen et al., 2017). This cohort consisted of mainly HIV uninfected individuals (<1%). Least absolute shrinkage and selection operator (LASSO) regression analysis was used to identify a 10-transcript signature and a logistic regression, cross-validation, followed by age-adjusted LASSO regression analysis to identify a 7-transcript signature. These signatures performed extremely well on the test set with AUCs of 0.94 for both signatures when discriminating between TB and symptomatic non-TB controls. Although this is the highest AUC from the paediatric biomarker studies discussed, there was no independent validation of these signatures and so it is difficult to interpret the efficacy of these signatures.

While these studies were significant in the development of paediatric biomarkers, they are still far fewer in number compared to adult biomarker studies. They also contain different limitations which prevent their signatures from being developed into clinically relevant tests. Further studies should utilise these paediatric cohorts, alongside other larger cohorts, for independent evaluation of new paediatric biomarker signatures.

In summary, the field of host transcriptomics biomarker signatures is a fast evolving one, however emphasis is placed on signatures diagnosing adult TB and no guidelines have been set to determine if a multi-country signature is required, alongside a lack of specifics around disease phenotypes required for development of such tests, such as whether to include HIV infected patients. Furthermore, there has been discussion on whether a host transcriptomic signature could ever meet the WHO guidelines with the current studies and technologies available (Walzl et al., 2018). However, the development of a novel biomarker test, specifically for paediatric TB, is critically needed to overcome the limitations in current tests, and current studies are providing promising results by nearing the WHO guidelines.

1.2.4 Metabolomics

Metabolomics is the study of small molecules in a system to understand the synthesis and degradation of molecules towards different cellular processes. It is the direct output of the genomic and transcriptomic operations of a cell and provides a snapshot of the final molecules produced as a result of those processes. Metabolomics can be used to study a targeted selected of metabolites, called targeted metabolomics, or to measure all metabolites identified in a sample, called untargeted metabolomics.

Metabolomics can be studied using different technologies such as liquid chromatography mass spectrometry (LC/MS), gas chromatography mass spectrometry (GC/MS), or nuclear magnetic resonance (NMR) spectroscopy (Tounta et al., 2021). MS-based techniques separate molecules through their mass-to-charge ratio whereas NMR techniques use radio pulses in a magnetic field to affect the nuclei of molecules. These techniques have different advantages and disadvantages depending on the requirements of the study, such as LC/MS being highly sensitive and preferred for molecules of lower concentrations. While certain techniques may be more suitable for different sample types, different experimental set ups can also allow the measurement of different molecules. For example, within MS, LC/MS enables the largest coverage of the metabolome of samples, while GC/MS is preferred when measuring volatile compounds.

A powerful supplemental technique to standard metabolomics analysis is metabolic isotope tracing. This involves using an isotope labelled metabolite as a substrate in a chosen system, such as ^{13}C , ^{15}N

or ^2H labelled media as a growth substrate for a bacteria culture (Tounta et al., 2021). This enables the measurement of uptake of specific metabolites over a period of time and allows researchers to observe which metabolic processes utilise those metabolites. Furthermore, another technique called flux balance analysis (FBA) measures the metabolic flux (i.e. synthesis and breakdown) of molecules in a network (Orth et al., 2010). FBA uses reconstructions of genome-scale metabolic networks which contain all of the currently characterised reactions in a system and calculates the flow of metabolites through these networks. It can then be used to estimate metabolic flux in different systems, including growth rate, based on certain model constraints.

While metabolomics has been used extensively to study biology, it is a less widely used approach in the field of infectious diseases. This is partly due to the importance of proteins in the host immune response, such as antibodies, however both the host and pathogen metabolism has recently been implicated in several factors relevant to disease outcome such as virulence. With this recent appreciation of metabolomics for the study of infectious disease, and the development of bioinformatic tools to utilise this type of data, metabolomics in infectious disease is becoming a more widely used method to investigate TB.

Major metabolic pathways

Within host and pathogen metabolomics, the majority of studies focus on several major biochemical pathways which are common to both organisms, including central carbon metabolism, TCA cycle, and the pentose phosphate pathway (**Figure 1.7**). Glycolysis, included in the central carbon metabolism pathway, involves the conversion of glucose to pyruvate (Rhee et al., 2011). Pyruvate is then converted to adenosine 5-triphosphate (ATP) and also reduces nicotinamide adenine dinucleotide (NADH) to NAD^+ . As a result of this high ATP and NAD^+ production, this stage of glycolysis is frequently used by proliferating cells as it generates a large amount of energy. The pentose phosphate pathway is a branching point from the glycolysis pathway through the glucose-6-phosphate metabolites which is produced during an intermediate stage of glucose to pyruvate conversion. The activation of the pentose phosphate pathway is used when cells require either the synthesis of nucleotide and amino acid precursors for cell growth and proliferation, or the synthesis of reducing metabolites for redox metabolism and fatty acid synthesis.

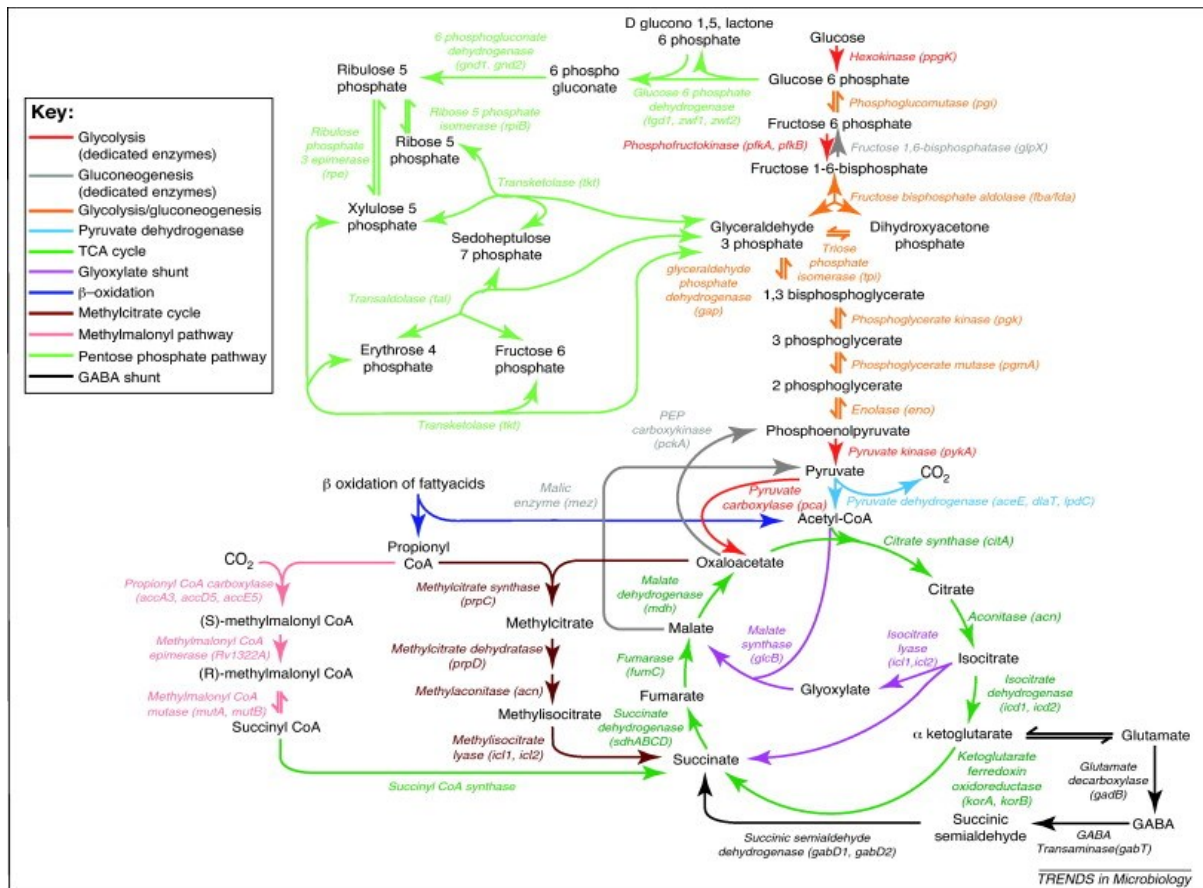


Figure 1.7: Major metabolic pathway schematic in *Mtb* (Rhee et al., 2011). CoA = coenzyme A; GABA = gamma-aminobutyrate. Reprinted from *Trends in Microbiology*, 19/7, Rhee, K.Y., De Carvalho, L.P.S., Bryk, R., Ehrt, S., Marrero, J., Park, S.W., Schnappinger, D., Venugopal, A. and Nathan, C, 307-314, Copyright (2011), with permission from Elsevier.

During the TCA cycle, pyruvate or fatty acids are used as substrates to produce multiple metabolites such as ATP, NADH, and flavin dinucleotide hydroquinone form (FADH₂). These can then be used to create different substrates with varying functions. There are several further pathways which branch from the TCA cycle using different intermediates. One of these pathways is oxidative phosphorylation (OXPHOS) which is used to generate energy using the intermediates of FADH₂ production. While OXPHOS produces energy, it is less efficient than other pathways and is often used by nonproliferating cells to maintain basal metabolism. In contrast, fatty acid oxidation is a pathway which produces a higher amount of energy (ATP in this case) than glycolysis. This pathway branches from the TCA cycle and involves the breakdown of fatty acids. Fatty acids can also be synthesised in the fatty acid synthesis pathway to produce lipids which are used to maintain cell structure and integrity of proliferating and growing cells.

Finally, amino acid metabolism is another important pathway as amino acids are used for various cellular functions including protein synthesis, fatty acid synthesis, and purine and pyrimidine synthesis. Amino acids are metabolised to and from different intermediates of the aforementioned pathways.

Applications of metabolomics to understand Mtb

Following the initial genomic and transcriptomic studies in Mtb, highlighting roles for Mtb specific enzymes such as ICL and fatty acid metabolic pathways for Mtb survival within the host macrophage, it was clear that metabolism played a central role in Mtb infection. Similarly, the field of metabolomics has focussed on understanding Mtb survival in the host macrophage cell and adaption to the phagolysosome (Park et al., 2021). Early studies identified that Mtb central carbon metabolism, which includes glycolysis, gluconeogenesis, the pentose phosphate shunt, and the TCA cycle, is an important pathway modulated throughout infection (**Figure 1.7**) (Rhee et al., 2011).

Several studies have investigated the preferred carbon source of Mtb in different settings. It has been shown that Mtb utilises preferentially cholesterol and fatty acids as the main source of carbon for growth within the host macrophage. The metabolism of cholesterol and fatty acids is through the TCA cycle, branching into the glyoxylate cycle, also known as the glyoxylate shunt, and methylcitrate cycle (**Figure 7**). The glyoxylate shunt is required to maintain the levels of TCA cycle intermediates during growth using fatty acids as a substrate including generating phosphoenolpyruvate (PEP) and pyruvate, and bypasses steps in the TCA cycle (Muñoz-Elías & McKinney, 2005). In Mtb, the glyoxylate shunt involves the action of 3 enzymes, *glcB* which encodes a malate synthase and converts acetyl-CoA or glyoxylate to malate, and two Icls encoded by *icl1* and *icl2* which convert isocitrate to glyoxylate or succinate. These Icls also act as methylisocitrate lyases, which are utilised in the methylcitrate cycle and convert methylisocitrate to succinate. These cycles have been shown to be essential for Mtb survival as Mtb mutants lacking both Icls were unable to survive when grown on fatty acid substrates and were cleared from mice lungs after 2 weeks of infection. In contrast, wild-type Mtb maintained a high bacterial load in mice lungs after at least 16 weeks of infection. However, unlike in other bacteria, Mtb has the ability to co-catabolize different carbon sources, allowing it to compartmentalise carbon sources for use in different downstream pathways (de Carvalho et al., 2010). This was shown in an *in vitro* study of Mtb grown on mixed carbon sources of dextrose and acetate, dextrose and glycerol, and acetate and glycerol. In fact, Mtb grew faster when replicating in mixed carbon cultures than in single carbon cultures. This co-catabolism of carbon sources has been observed in very few other bacteria

and the presence of this mechanism in Mtb is likely the result of evolving within the host macrophage cell where carbon sources are scarce, highlighting the importance of metabolism in Mtb infection.

A focus of many Mtb metabolomic studies is the role of metabolism in Mtb virulence, stress, and toxin-antitoxin pathways. Mtb susceptibility through the loss of Icls is partly due to the loss of propionate-CoA degradation in the methylcitrate cycle resulting in an unfavourable balance of NAD/NADH, membrane potential, and internal pH (**Figure 1.8**) (Eoh & Rhee, 2014). This implicated Icls as important for detoxification as well as virulence and metabolism. Deletions of the *prpD* gene, required for the methylcitrate dehydratase enzyme in the methylcitrate cycle, have shown a total growth defect when grown in liquid broth with propionate or odd-chain fatty acids as the primary carbon source and a partial growth defect in Mtb upon infection in murine BMDMs (Muñoz-Elías et al., 2006). However, the methylcitrate cycle was not shown to be essential in a mouse model, indicating that the methylcitrate cycle may only be required for virulence with specific carbon sources. Further analysis of the glyoxylate cycle also identified malate synthase as essential for Mtb growth *in vitro* when grown with acetate, propionate, and oleic acid (Puckett et al., 2017). Mtb lacking malate synthase was shown to be unable to replicate in mice. The susceptibility to oleic acid was found to be as a result of glyoxylate accumulation, and the exact mechanism of glyoxylate toxicity are still unknown but are suggested to be related to the build-up of by-products of acetate metabolism leading to ketoacidosis.

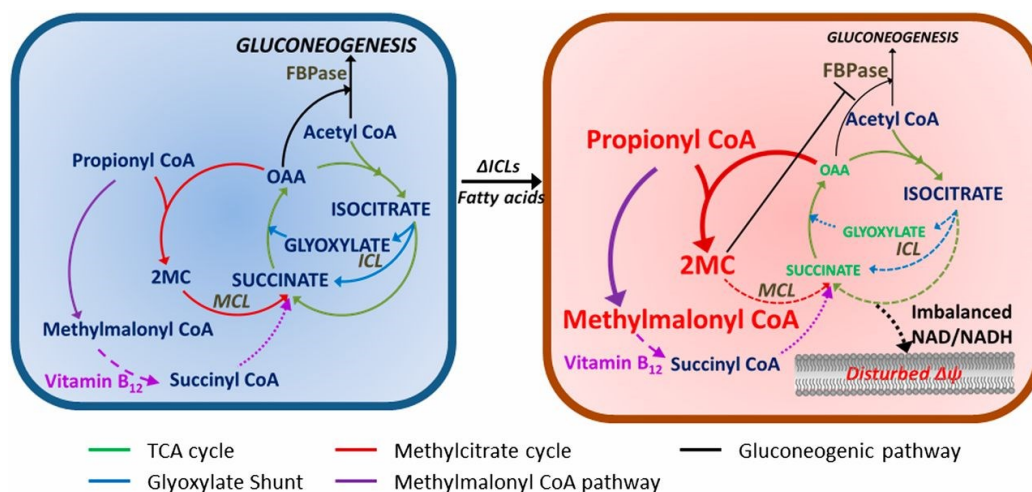


Figure 1.8: Hypothesised changes in metabolism of ICL-deficient Mtb strains (Eoh & Rhee, 2014). Central carbon metabolism pathway is shown in the blue scheme on the left, while the perturbations of this pathway caused by the loss of ICLs in Mtb mutants grown on fatty acid media is shown in the red scheme on the right.

Despite the known impact of Mtb metabolism on virulence and survival, there are few studies investigating the differences in metabolism between Mtb strains and lineages. One of these studies compared the metabolisms of hyper- and hypo-virulent strains of Mtb from the Beijing lineage from *in vitro* cultures (Meissner-Roloff et al., 2012). This study found 282 metabolites which separated the two phenotypes by principal component clustering (PCA) clustering and a moderate separation using partial least-squares discriminant analysis (PLS-DA). They further identified the top 40 metabolites from these 2 analyses based on effect sizes and through T testing. Some of the top metabolites included several amino acids which were at a lower abundance in the hyper-virulent Mtb strains including l-alanine, glycine, glutamine, phenylalanine, and aspartic acid. Similarly, glycolysis metabolites were decreased in abundance in the hyper-virulent strain, including glucose, galactose, mannose, myoinositol, glycerol, and ribitol. While this study was one of the first to compare Mtb strains, the lack of further bioinformatic analyses or validation studies make it challenging to interpret these results and compare them to *in vivo* infection.

While Mtb lineages have recently become an area of intense study in the fields of genomics and transcriptomics, this remains lacking in metabolomics. A possible reason for this is the lack of freely available developed bioinformatic tools to perform these analyses. Thousands of bioinformatic applications have been developed for the field of transcriptomics, although mainly targeted towards host transcriptomics. In contrast, the tools available for metabolomics analyses are lacking, with some not being openly accessible.

Applications of metabolomics to understand the host response to TB

While metabolism has been extensively studied in Mtb due to the large proportion of the Mtb genome and the influence it has on virulence depends on metabolic pathways and manipulation, there has also been a growing interest in host metabolism in response to infection and to Mtb specifically. Generally, metabolism is important for immunity to infection as immune cells themselves must carry out different metabolic functions such as synthesising or degrading cell surface receptors to detect invading pathogens. Each stage of the immune response, from pathogen detection, antigen presentation, and priming and activation of the adaptive immune response requires specific metabolic pathways. Likewise, different immune responses require distinct metabolic pathways to carry out these functions.

As with the previous omics technologies, the metabolism of the macrophage response to Mtb infection has been the main focus of host metabolomics studies. Mtb has been shown to manipulate host metabolism to aid in its survival during infection. For example, it was recently identified that live Mtb, but not dead Mtb, decreases both glycolysis and mitochondrial oxidative phosphorylation in infected THP-1 macrophages to reduce ATP production in these cells (Cumming, Addicott, et al., 2018). This would therefore prevent macrophages from producing the energy needed to carry out their functions in altering the adaptive immune response or killing the bacteria. In addition, macrophages infected with live Mtb increased fatty acid metabolism, specifically towards exogenous fatty acids during stressful conditions, which is suggested to be a reaction to Mtb preventing the use of endogenous fatty acids from within the cell. Further studies have identified that live, but not dead Mtb, also induce type I IFN metabolic changes in host macrophages, to decrease glycolysis and increase mitochondrial stress (Olson et al., 2021).

Comparative analysis on the metabolism of macrophages which can successfully control infection compared to those which cannot has also brought about insights into effective metabolic pathways for host control of infection. One study compared the transcriptome and metabolome of alveolar and interstitial macrophages, both of which are found in the lungs, to determine what type of metabolic phenotype they represent and how this might affect infection (Huang et al., 2018). The authors identified that alveolar and interstitial macrophages did not present with M1 or M2 phenotypes when comparing the transcriptomes but did present functionally distinct metabolisms. Interstitial macrophages produced more lactate and were inhibited by 2-deoxy-D-glucose (2-DG), a drug used to measure glycolytic metabolism, than alveolar macrophages, suggesting higher glycolytic activity in these cells. This suggested that alveolar macrophages presented with a lower glycolytic metabolism. This increased glycolytic activity of interstitial macrophages was also shown to be important for host control of Mtb infection, as inhibition using 2-DG increased bacterial replication in the mouse lungs. This is in contrast with fatty acid oxidation which was higher in alveolar macrophages and was associated with increased bacterial replication. There is therefore a consensus in both *in vitro* and *in vivo* studies that increased glycolytic activity is necessary for bacterial control during infection, and that Mtb shifts metabolism from this pathway towards increased fatty acid metabolism to increase survival.

In contrast to macrophage cells, there are far fewer studies investigating the metabolic response of adaptive immune cells to Mtb infection. It is likely that disruption of macrophage metabolism by Mtb

leads to limited, or less effective, activation of downstream adaptive immune responses. It has been shown that macrophages infected with the BCG vaccine strain of mycobacteria stimulate different metabolic responses in macrophages than those infected with live Mtb (Cumming, Addicott, et al., 2018), which may also contribute to the lack of effective Mtb immunity in BCG vaccinated adults. Moreover, there are studies investigating chronic disease which often focus on T cell tolerance and exhaustion during LTBI and within TB granulomas, which are outside the scope of this thesis.

1.2.5 Bioinformatics methods

Across the -omics technologies there are certain overarching statistical principals and analyses which have been developed alongside the sequencing technologies mentioned previously. Here, I will explain the background of these statistical principals which are used throughout this thesis.

Sequence alignment and mapping

In recent years, the use of microarray chips has become almost obsolete in the face of the decreasing costs of RNA-sequencing (RNA-seq). While microarrays produce fluorescence intensities, RNA-seq experiments produce reads. These reads comprise of short RNA sequences which must be aligned to a reference genome or transcriptome to be converted into gene counts and interpreted by researchers. There are hundreds of sequence alignment algorithms available, and many are designed for different types of data. In general, sequence aligners take RNA reads and align them to a reference sequence based on calculated homology estimates. These estimates can be calculated through different mathematical formulae and therefore, depending on the alignment algorithm chosen, will produce different results. A major challenge for aligners is aligning short-read sequences, sequences with high repetition, and mismatching, insertion, or deletion regions in sequence reads. Some of these challenges are overcome depending on the sequencing technology used, but often differences in accuracy of alignment result from these issues.

For the purposes of this thesis I have used two alignment algorithms which fall broadly into spliced-aware and spliced-unaware aligners. The difference between these broad categories is whether these algorithms consider sequences which are produced from spliced genomes/transcriptomes, such as human genomes/transcriptomes, or unspliced genomes/transcriptomes such as bacterial genomes/transcriptomes. In this study, the splice-aware alignment tool Spliced Transcript Alignment

to a Reference (STAR) was used to align human reads to a reference sequence (Dobin et al., 2013). For splice-unaware algorithms, Burrows-Wheeler Alignment (BWA) was used to align bacterial reads to reference sequences (Li, 2013).

BWA-MEM, the specific algorithm used in some analyses in this thesis, follows the seed-and-extend paradigm (Li, 2013). This describes finding a position with the longest exact match, known as the seed. Seeds which are colinear and are also close to each other are grouped into chains. Seeds are then ranked, based on the length of the chain it is found in and then by the length of the seed. In ranked order, seeds found in previous alignments are dropped, and the remaining seeds are extended to identify better alignment penalty scores, based on a threshold.

STAR aligns reads in 2 steps, 1) seed search and, 2) clustering, stitching, and scoring. In the seed search step, STAR searches for the longest sequence in the reference genome which matches a read, or a portion of the read. In the case of spliced transcriptomes, the read will only partially match the genome sequence. For any unmapped section of the read (e.g. due to splicing events), STAR searches the reference sequence for matches between the reference and the unmapped read sequence. Each portion of the read which matches a portion of the reference sequence is called a seed and so once all the seeds have been identified for one read, they are clustered and seeds which cover the reference sequence are stitched together based on the lowest mismatching rate.

Machine learning techniques for bioinformatics

Knowledge of various machine learning techniques commonly used in multi -omics data analysis is necessary to fully understand the advantages and limitations of the approaches used throughout this thesis. These techniques will therefore be briefly summarised here. For the purposes of this thesis, I will refer to sample features (such as genes, age, gender, strains, etc) as variables. Depending on the type of analysis, many of the bioinformatic methods described can be used to analyse samples and/or variables and therefore both terms can be used interchangeably throughout the following sections.

Clustering is a technique used to group similar samples, or variables, together based on similarity scores. Hierarchical clustering is an unsupervised approach, meaning classifying information, such as if a sample is case or control, is not used as input into the algorithm. In agglomerative clustering, each

sample is treated as an individual cluster, before being iteratively grouped with the other most similar clusters until all samples form one single cluster. This method is commonly used in understanding biological data and is what is used throughout this thesis in the form of dendrograms.

Regression analysis aims to identify structure or relationships within a set of samples. Linear regression identifies a linear structure in the data, i.e. it fits a straight line which minimises the sum of the squared differences of the variables and that line (**Figure 1.9A**). Linear regression is often used for differential expression analysis to identify variables which are over or under expressed compared to a control or other comparison (Love et al., 2014). Logistic regression fits a logistic function to model binary variables (**Figure 1.9B**). This is often used for classification problems (explained below) using variables such as SNP data.

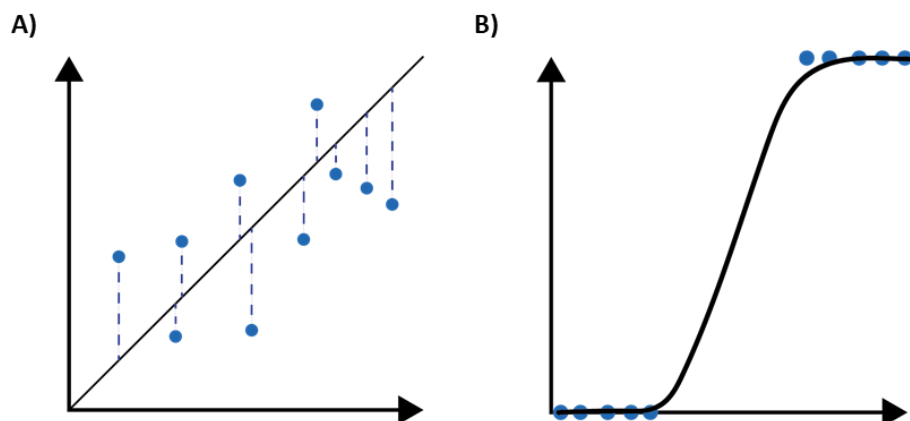


Figure 1.9: Regression models. **A)** Linear regression. Blue dots indicate variables. The black line indicates the relationship which produces the shortest distances between the black line and the variables. Blue dotted lines indicate distances from the line. **B)** Logarithmic regression. Blue dots indicate variables.

Classification analysis groups or categorises samples. In machine learning, classification algorithms use input (training) data to predict the classification of future (test) data. The training and testing datasets can either be selections from one dataset (often split 80:20 by training and testing, respectively, or a similar ratio) or can be made up of two independent datasets. There are several classification algorithms used to solve different problems, of which 3 are used in this thesis, including logistic regression. The second of these are random forest classifiers, which are based on decision trees.

Decision trees are similar to clustering models in that they initially use the entire dataset as 1 group before splitting this singular group into sequentially smaller groups based on similarity. While decision trees are a supervised method, clustering techniques are unsupervised. Random forest algorithms consist of multiple decision trees, each producing independent outputs from the same training data (**Figure 1.10A**). The overall prediction is created from the summary of these individual decisions.

Another classification technique is support vector machine (SVM) models. This technique is also supervised and aims to identify a hyperplane (or line) in space which separates or classifies the input variables (**Figure 1.10B**). In building the model, the variables which are closest to the hyperplane are defined as support vectors. The support vector model optimises the space between support vectors and the hyperplane to be the largest possible distance. The output of SVM models is again prediction of classification.

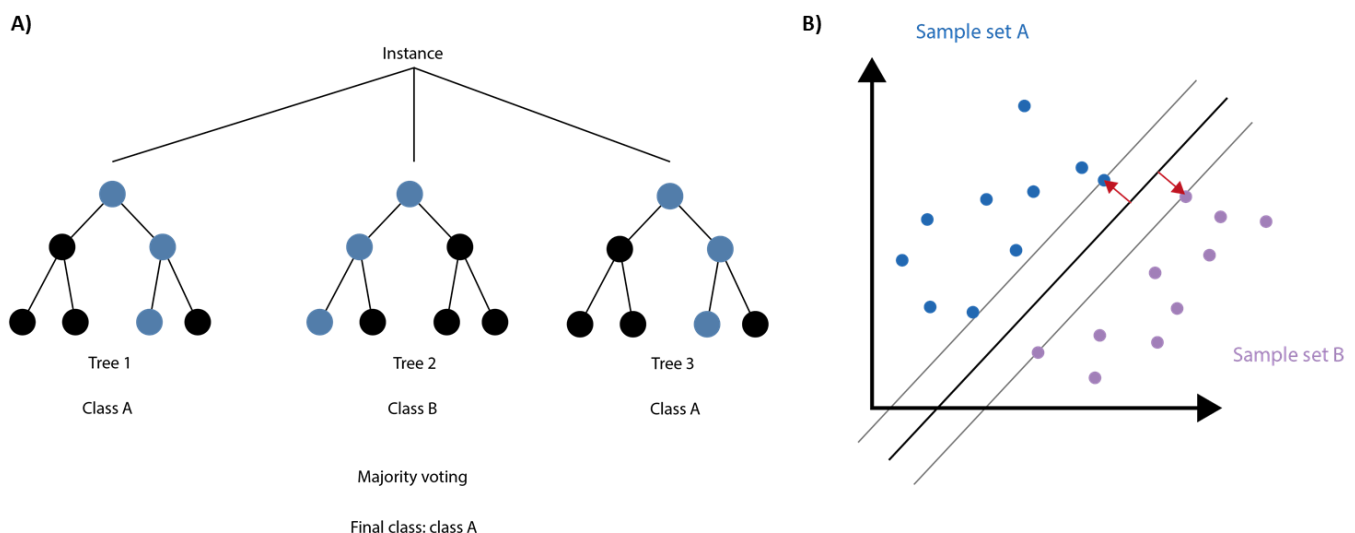


Figure 1.10: Classification methods. **A)** Random forest classification method. Dots indicate variables. Lines indicate decision making process. Final class is the final prediction based on the three trees in this instance. **B)** Support vector machine classification method. Dots indicate variables from two sample sets, A (blue) and B (pink).

The performance of the classification models described can then be evaluated using Receiver Operating Characteristics (ROC) curves. While there are different approaches to evaluate classifier performance, such as Precision-Recall curves, ROC curves are almost exclusively used in the field of host transcriptomic biomarkers for TB. A ROC curve measures the performance of a classification

model at different thresholds. It represents a probability curve which calculates different AUC statistics for each threshold, which is calculated from sensitivity (true positive) and false positive (or often defined as 1-specificity) rates. The higher the AUC, the more accurate the classifier is at differentiating between two binary variables (e.g. TB or not TB). An AUC of 0.5 (or 50%) describes the prediction accuracy as if classified by random. It is also possible to manually set different thresholds when performing ROC curve analysis to observe different sensitivities and specificities of your model. For example, to identify the maximum specificity for a model when the sensitivity is set to a cut-off value e.g. 70%.

Differential expression/abundance methods and pathway analysis

Differential expression analysis is commonly used to identify transcripts which are expressed statistically higher or lower in one condition or strain compared to another condition or strain, at a rate that is greater than expected due to random variation. This type of analysis produces log fold-change values representing the fold-change in expression between the case and control groups. Differential expression analysis is commonly applied in transcriptomics analyses, initially starting with microarray data before progressing to include RNA-seq data once this became affordable.

While there are several packages that can be used to perform this analysis, the majority of these packages were designed with microarray datasets in mind. Microarray analysis uses fluorophore-bound cDNA sequences from samples, with different coloured fluorophores corresponding to different sample types such as case and control. These sequences bind to probes on a chip, and the resulting fluorescence signal measured represents the level of expression of specific gene sequences. In contrast, RNA-seq data measure sequence reads, corresponding to specific transcripts. As these two technologies measure different outputs, the characteristics of their output data are different, resulting in differences in downstream analysis. Specifically, microarray data follow a Gaussian distribution (also known as normal distribution), where the data variance is determined by the mean. In contrast, RNA-seq count data is heteroskedastic, meaning that the variance of the data is highly variable. In count data, it is the genes with low mean expression values which are highly variable. To account for this, RNA-seq data is often modelled using a Negative Binomial distribution (Anders & Huber, 2010). As a result of the differences in data distributions, different models, and therefore bioinformatic packages, should be used for differential gene expression analysis. For example, for microarray analysis, limma (Ritchie et al., 2015) is commonly used which utilises a linear model, while

for RNA-seq analysis edgeR (Robinson et al., 2010) and DESeq2 (Love et al., 2014) are commonly used and utilise negative binomial models.

Similarly, differential abundance analysis is the identification of metabolites with statistically significantly higher or lower abundance in one condition or strain compared to another condition or strain, at a rate that is greater than expected due to random variation. This is one method which has been used to analyse metabolomics data. However, as metabolomics data differ in distribution to RNA-seq count data, they cannot be analysed using similar packages such as DESeq2. Furthermore, as the field of metabolomics analysis is still developing, there are fewer packages available in R which can be used for this analysis and often studies will use analysis of variance (ANOVA) tests instead to compare abundances of specific metabolites. In this thesis, the number of metabolites under investigation is large, and therefore could not be analysed using ANOVA test which are designed for a smaller number of comparisons and does not account for multiple testing. Instead, other regression analysis packages such as limma, which was originally designed for microarray data, can be used.

Dimensionality reduction

Dimensionality reduction techniques are used in bioinformatics to visualise complex datasets with a large number of variables or conditions such as many -omic datasets. There are two broad categories dimensionality reduction techniques fall into: feature extraction and feature selection. In feature extraction, new features are created which represent the original variance in the data, while feature selection selects original features which describe the majority of the variance in the data.

Within feature extraction techniques, PCA is often used in bioinformatics for visualisation of multivariate datasets. PCA is an unsupervised method which creates new features called principal components from highly correlated variables. Each principal component is orthogonal from the previous one and therefore each new component is linearly uncorrelated to the last and captures a different part of the dataset variance. Other methods, such as orthogonal partial least squares (OPLS) and orthogonal partial least squares discriminant analysis (OPLS-DA) are regression analyses often used in metabolomics analysis to identify discriminating variables between sample groups. These methods differ from PCA as they are supervised and therefore are typically used to identify discriminating features between variable groups.

In contrast, feature selection is a method where the number of features is reduced to a smaller number of the original features. There are several different methods used for this analysis but in this thesis the forward-selection partial-least squares (FS-PLS) is used. This is an in-house method which selects the variable which explains the highest proportion of variance in the data, then removes all the variables correlated with that initial variable. This step is then repeated iteratively until there are no more variables which explain variance in the data. Here, this is used to select genes which contribute to classifying different disease states to select a biomarker signature for a particular disease.

1.2.6 The development of multi-omics research

Omics research has significantly advanced research over the past few decades with respect to infectious disease. More recently, the use of multiple omics technologies together has opened further potential for uncovering new biological pathways and deepening the understanding of known pathways. The field of multi-omics started with interpreting multi-omics results individually to fully integrating multi-omics datasets using complex bioinformatic methods. Several integration methods have recently been developed, each with their own approaches to integrating different types of multi-omics datasets. Multi-omics techniques have been used in many fields, including infectious diseases, and can be particularly powerful when applied to complex diseases such as TB. Single omics techniques have advanced TB research, but by incorporating multi-omics approaches to studying TB researchers can investigate multiple levels of molecular and cellular interactions at once.

1.2.7 Multi-omics integration methods

Several integration methods have been developed using different bioinformatic approaches, with different applications, and are used through different platforms. Multi-omics analyses can be used for different types of analyses, including biomarker discovery and drivers of disease, discovering disease subgroups, and disease insights (Subramanian et al., 2020). Biomarker discovery identifies molecular features across different omics datasets which drive or represent the disease or group of investigation. Many complex diseases, like TB, cannot be classified into simple groups, such as active and latent TB patients, and instead multi-omics datasets from patients with these diseases can identify biological characteristics and pathways associated with specific phenotypes or subgroups. Multi-omics can also

be used to gain insights into diseases by discovering features or pathways which are contribute to the disease development or outcome.

The multi-omics approaches themselves have been previously categorised through different systems, including on the biological problem they were derived to solve or their statistical framework. Each method may use different underlying statistical methodologies, or even several methodologies in a single approach. These methods can be grouped differently depending on the study, but include similarity, correlation, fusion, network, Bayesian, and multivariate analyses to address biological hypotheses (Subramanian et al., 2020).

These techniques have been described in detail elsewhere (Subramanian et al., 2020), but will briefly be described here. Similarity approaches utilise similarities calculated between samples or variables, typically using distance metrics such as Euclidean distance. Network approaches use a graphical model to represent interactions between variables. Fusion methods identify patterns across datasets in a singular low-dimensional feature space. Similarity network fusion (SNF) is an example of a method which uses both similarity, network, and fusion approaches for integration (Wang et al., 2014). SNF creates similarity networks for each omics dataset using pairwise similarities which can also be represented as a network, and then fuses these networks into a single similarity network using a nonlinear method (**Figure 1.11**). Correlation approaches measure the association between variables. This association is often measured linearly, i.e. when one variable increases in measurement the other also increases, for example through using the Pearson correlation coefficient. Bayesian approaches is a very broad classification which can encompass several types of statistical analyses depending on the test used. It is largely based on Bayes' theorem which uses probabilities, evidence, and likelihood of an event based on a given priori (prior known information) and gives you a posteriori. Similarly, multivariate methods are a broad classification which refer to the analysis of more than one dataset. These types of analyses include PCA and canonical correspondence analysis (CCA), which are used in integration methods mixOmics (Rohart et al., 2017) and PMA (Witten et al., 2009) respectively.

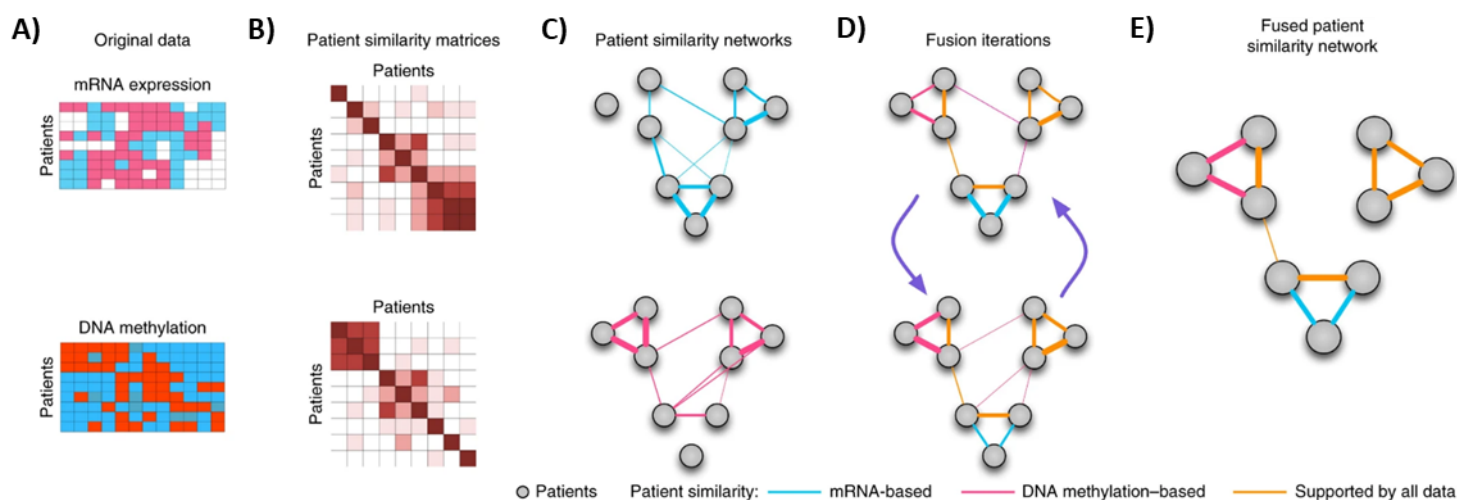


Figure 1.11: Similarity Network Fusion methodology (B.Wang et al., 2014). **A)** Two different - omics datasets originating from several patient samples are used as input. **B)** Similarity matrices are generated using these inputs to correlate patients. **C)** Similarity networks are generated using the same inputs, however here correlations are represented by networks with each patient represented by a node and pairwise similarities represented by edges. **D)** These networks are then fused and are iteratively updated with other networks to increase similarities. **E)** The result is a final fused network, with edge colours representing the original data type (e.g. mRNA or DNA methylation) which has contributed to those networks. Reprinted by permission from Nature Methods: Springer Nature, Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B. and Goldenberg, A., 2014.

There are certain factors which must be taken into account before choosing the appropriate multi-omics approach for a given study. Certain multi-omics approaches have been developed with specific types of omics datasets in mind and, as a result, some approaches cannot currently use specific types of omics data. For example, many methods require a minimum sample size to use them, for example, Multi-Omics Factor Analysis recommends a sample size greater than 15. While it can be beneficial to have larger sample numbers in multi-omics experiments as multiple testing requires more stringent p-value cut offs and lower sample sizes may not provide enough power to meet these cut offs, these large-scale experiments are often very expensive and time-consuming. By limiting the sample size of which these programmes can use it decreases the accessibility of these approaches to smaller groups. Furthermore, multi-omics datasets also require a large amount of data storage space and, alongside that, high computational power to perform certain integration analyses if using large datasets. Compounding this issue is the technical expertise available to perform multi-omics analysis, as all approaches require bioinformatic expertise and all of these methods are only available in one programming language, meaning it is more difficult to utilise approaches not developed in your chosen language.

Underlying these considerations is the issue that few of these methods and tools have been independently validated except in a handful of studies (Cantini et al., 2021; Chauvel et al., 2020; Rappoport & Shamir, 2018; Sathyanarayanan et al., 2020; Tini et al., 2019). These independent studies only use a small selection of multi-omics integration techniques, for example Sathyanarayanan, et. al., used 9 techniques whereas Rappoport, et. al., used 25 techniques (Rappoport & Shamir, 2018; Sathyanarayanan et al., 2020). However, these studies claim their chosen techniques represent the general underlying mathematical principals of all currently available methodologies and are therefore relevant despite not exhaustively comparing all techniques (Cantini et al., 2021). The studies also evaluate the methods based on different outcomes, through measuring different types of scores to measure the accuracy of the tests, such as F-score, or similarities between tests, such as Jaccard indices. They also measure different criteria, such as signal to noise or biological pathways identified, although number of clusters identified is a common approach in several of these studies. Finally, the studies identify a few key integration techniques which perform better, on average, than others, although these techniques are not common across studies. The lack of standardisation of methods and evaluation criteria used leads to it being extremely difficult to identify which technique is most appropriate for certain types of analyses. Another issue to consider is that all these studies use simulated data and cancer data but have yet to evaluate these techniques on other datasets from other diseases. Overall, there is still no gold standard identified in which mathematic approaches, and therefore integration techniques, are appropriate to use for analysis of real-world biological data.

1.2.8 Multi-omics integration in infectious diseases and TB

Despite almost all the integration methods currently available having been developed using different cancer datasets (Subramanian et al., 2020), recently they have been applied to the field of infectious diseases. Parasitic, viral, and bacterial infections have been studied using multi-omics approaches, including *Plasmodium* species, SARS-CoV-2, and *Mtb* (Montaldo et al., 2021; van Crevel et al., 2021; Zhou et al., 2021).

In the field of Malaria research, multi-omics have been widely used for the identification of *Plasmodium* genes, transcripts, proteins, and metabolites which contribute to a variety of pathways involved in different stages of the parasite life cycle, symptomatic and asymptomatic malaria, and pathways involved in the dissemination of the parasite from both human and mosquito hosts (Zhou

et al., 2021). Multi-omics approaches have also frequently been used in COVID-19 studies to characterise disease pathogenesis. For example, one study combined single-cell and multi-omics analyses to identify different immune cell populations as well as immune cell profiles in COVID-19 patients with different disease severities (Single-cell multi-omics analysis of the immune response in COVID-19), while another used different combinations of interactomes, proteomes, transcriptomes, and bibliomes (genes identified from the DisGeNET data collection (Barh et al., 2020)) from several published studies to identify commonly upregulated pathways in both acute and long-term COVID-19 patients (Barh et al., 2021). Multi-omics approaches have also been used to identify drug targets for diseases such as COVID-19 (Barh et al., 2020).

While there are many single -omics studies of TB, there are only a small number of integrated multi-omics studies of this disease. In addition, two of these studies focus on drug resistance with one using genomics, proteomics, and metabolomics to identify the mechanism of capreomycin resistance (Zhao et al., 2019) and another study using the same three technologies to identify the mechanism of *p*-aminosalicylic acid resistance (Wei et al., 2019), both being *in vitro* studies. While important for designing therapeutics to treat TB, drug resistance is only one aspect of TB research.

Integrated multi-omics analysis has also been used to investigate Mtb lineages. Specifically, one study used genomics, transcriptomics, and proteomics to identify differences between the B0/W148 cluster of Beijing Mtb strains from the H37Rv reference strain (Bespyatykh et al., 2019). This included higher expression of the *ethA* gene, implicated in drug resistance, in strains in this cluster as well as decreased expression of genes within the DosR regulon confirmed with an observed decrease abundance of proteins within this regulon. This Beijing cluster was also investigated using genomics, transcriptomics, proteomics, and metabolomics by another group (Bespyatykh et al., 2020). Here, the authors performed multi-omics analyses on isolates of these strains from a patient undergoing anti-TB therapy to understand the changes in the bacterium during this process. Similarly, another study investigated the virulence of the avirulent H37Ra Mtb strain, derived from the H37Rv reference strain, using an integrated genomics, transcriptomics, and proteomics approach (Pinto et al., 2018). They identified genomic and proteomic evidence for the presence of several virulence pathways thought to be absent in H37Ra, such as the *esxO* and *esxK* genes and proteins. They also redefined the H37Ra genome through identification of 82 novel peptides linked to 63 novel protein-coding genes. Finally, one group generated a novel multi-omics approach to integrate SNPs with metabolomic data to investigate the genomic drivers of metabolomic differences in Mtb lineages (Øyås et al., 2020). A difference in the

TCA cycle, pyruvate, glycolysis, and glycine, serine, and threonine metabolic pathways was identified between lineages. Their approach allowed the identification of SNPs which resulted in downstream metabolic changes, providing targets for further experimental validation.

Multi-omics analyses can therefore be used to identify drug targets, further characterise Mtb strains, and link causal SNPs with downstream changes in Mtb lineages. Additional applications include the supplementation of biomarker studies with multi-omics to identify biomarkers with higher accuracy. One such study has been performed in TB, using miRNA, metabolomic, and cytokine/chemokine datasets from adult TB patients from a multi-country cohort which were HIV positive (Krishnan et al., 2021). While only finding candidate biomarkers in some of these datasets, when combining all 3 their decision-tree model identified a mixed signature comprising of a single miRNA and a single metabolite as the optimal signature with high accuracy (an AUC of 0.965). Although this study was performed on HIV positive individuals only, and was not validated in an independent cohort, it does provide the first example of attempting to improve current biomarker signature performances by supplementing single-omics datasets with new data.

There are a few examples of other novel applications of multi-omics analyses. One COVID-19 study used a unique approach to multi-omics analysis, through the identification of disease 'domains' which can be defined as conceptual repositories, to summarise previously published omics studies to identify findings across the omics technologies (Montaldo et al., 2021). Multiple omics studies confirmed findings in other studies, such as that platelet degranulation and coagulation cascades are involved in disease severity as well as changes observed in TCA metabolites which was suggested to be linked to exacerbated lung functions and lower blood oxygen content. Although confirmatory and not 'integration' in the traditional sense, this approach allows for the confirmation of single- or multi-omics findings across studies which may be particularly valuable in identifying key drivers of disease. Another application could be the combination of 'systems biology' with 'systems epidemiology', using multi-omics data to aid epidemiology findings for complex diseases such as TB (Ahmed & Hasnain, 2011). Finally, studies in other infectious diseases have suggested the integration of more than just -omics datasets, through including the integration of -omics data with other variables such as clinical information of cell populations to understand the linking of molecular results with clinical results and different disease outcomes (van Crevel et al., 2021).

The integration of multi-omics datasets is a new area of biology, particularly in infectious disease research and TB. The COVID-19 pandemic has seen more integrated multi-omics analyses used than ever before, which may help encourage its use for improving our understanding of other infectious diseases. Within TB research, the applications of integrated multi-omics approaches are few but varied. There are clearly benefits to using these approaches, either confirmation of known pathways relevant to disease, discovery of new pathways relevant to disease, or supplementation of other approaches. As the experimental set-up becomes more standardised and straightforward to implement, and as new bioinformatic packages and algorithms are developed to analyse and integrate these datasets, integrated multi-omics analyses will become ever more present in infectious disease research.

1.3 PhD Aims

1.3.1 Research Question

- How does the host and bacterium contribute to tuberculosis infection?
 - How does the paediatric host contribute to infection and can these responses be used to develop an accurate diagnostic test for paediatric TB?
 - How does the bacterium contribute to infection?
 - How do both host and bacterial interactions contribute to infection?

1.3.2 Aims

- 1) What immune pathways in the paediatric host are altered in response to Mtb infection?**
 - a. Investigate transcriptomic changes in paediatric TB patients from South Africa and identify genes/pathways which are affected in TB vs other disease (OD) groups and TB vs latent TB infection groups.
 - b. Evaluate previously published biomarker signatures to identify whether if any published TB-specific signatures can be used as a diagnostic test for paediatric TB in our South African paediatric cohort.
- 2) How has Mtb evolved to successfully infect the host?**
 - a. Investigate the genomic and metabolomic differences between Mtb strains and lineages.
- 3) How successful are the host and bacterial mechanisms at dealing with infection?**
 - a. How successful are the host mechanisms at responding to infection?
 - b. How successful are bacterial mechanisms at responding to the host immune response?
 - c. Are host or bacterial pathways identified in the *in vivo* datasets in Aims 1 and 2, overlapping with those in an *in vitro* infection model?

Chapter 2: Materials and Methods

Several authors contributed to this thesis and are described and acknowledged in the relevant Chapters. A summary of their contribution can be found here.

- 2.2.1 Mtb strains were kindly gifted by Dr Brian Robertson and his group at Imperial College London
- 2.2.2 Mtb culture and growth was performed by Dr Gerald Larrouy-Maumus and Agnieszka Broda at Imperial College London
- 2.2.3 Mtb metabolite extraction and HPLC-MS was performed by Dr Gerald Larrouy-Maumus and Agnieszka Broda at Imperial College London
- 2.3.1 Patient cohort was recruited by Prof Heather Zar and her group at Red Cross Children's Hospital and the University of Cape Town
- 2.3.1 Patient cohort RNA extraction was performed by Dr Victoria Wright, Giselle D'Souza, and Dr Samuel Nichols from the Levin group at Imperial College London
- 2.3.2 Patient RNA-seq pre-processing in Bash was performed by Dr Dominic Habgood-Coote in the Levin group at Imperial College London
- 2.4.9 The analysis in this section was performed by my student, Petra Bachanová, at Imperial College London and was supervised by me throughout

Reagents are sourced from Sigma-Aldrich, United States (US), unless otherwise stated.

2.1 Data availability across chapters

All sequencing data and code will become openly available upon publication of results. The genomics data used in this study is available from the European Nucleotide Archive (<https://www.ebi.ac.uk/ena/browser/home>, ENA accession number: ERP001731). The transcriptomics data will be available from the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>). The code used to produce the analyses in this Thesis will be available at the GitHub repository <https://github.com/ashleigh-cheyne>.

2.2 Methods used to investigate Mtb lineages

2.2.1 Mycobacterial strains

Mycobacterium tuberculosis (Mtb) strains used throughout this study, their origin, and lineage can be seen in **Table 2.1**.

Table 2.1: Mycobacterial strains used for Mtb lineage analyses.

Strain	Spoligotype*	Lineage**	Origin of acquisition	Origin of patient sample
Mtb H37Rv	N/A	Lineage 4 (EA)	Laboratory collection	N/A
Mtb 232	EA14_VNM	Lineage 1 (IO)	Dr Robertson's group	Vietnam
Mtb 281	ZERO	Lineage 1 (IO)	Dr Robertson's group	Vietnam
Mtb 346	EA14_VNM	Lineage 1 (IO)	Dr Robertson's group	Vietnam
Mtb 372	EA14_VNM	Lineage 1 (IO)	Dr Robertson's group	Vietnam
Mtb 119	BEIJING	Lineage 2 (BJ)	Dr Robertson's group	Vietnam
Mtb 212	BEIJING	Lineage 2 (BJ)	Dr Robertson's group	Vietnam
Mtb 345	BEIJING-LIKE	Lineage 2 (BJ)	Dr Robertson's group	Vietnam
Mtb 374	BEIJING	Lineage 2 (BJ)	Dr Robertson's group	Vietnam
Mtb 649	BEIJING-LIKE	Lineage 2 (BJ)	Dr Robertson's group	Vietnam
Mtb 173	H1	Lineage 4 (EA)	Dr Robertson's group	Vietnam
Mtb 293	T1	Lineage 4 (EA)	Dr Robertson's group	Vietnam
Mtb 318	T1	Lineage 4 (EA)	Dr Robertson's group	Vietnam
Mtb 440	H3	Lineage 4 (EA)	Dr Robertson's group	Vietnam
Mtb 639	H3	Lineage 4 (EA)	Dr Robertson's group	Vietnam
MTB_DY_28	N/A	Lineage 1 (IO)	(Comas et al., 2013)	Ghana
MTB_N0014	N/A	Lineage 1 (IO)	(Comas et al., 2013)	Thailand
MTB_N0043	N/A	Lineage 1 (IO)	(Comas et al., 2013)	Burma
MTB_N0132	N/A	Lineage 1 (IO)	(Comas et al., 2013)	The Philippines
MTB_N70_read	N/A	Lineage 1 (IO)	(Comas et al., 2013)	India
MTB_N72_read	N/A	Lineage 1 (IO)	(Comas et al., 2013)	India
MTB_N73	N/A	Lineage 1 (IO)	(Comas et al., 2013)	India
MTB_T83_12	N/A	Lineage 1 (IO)	(Comas et al., 2013)	Vietnam
NG-5604_BTBS-280	N/A	Lineage 1 (IO)	(Comas et al., 2013)	Ethiopia
1-6	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
1027	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
1336	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
1864	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
19-5	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
19-6	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
19-7	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
58	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
72	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
759	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
861	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ-1165	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ-1168	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ-1343	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ-1438	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ-1972	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ-50	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ10-3	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ1331	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ1335	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ1439	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ1580	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ1597	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ1605	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ1645	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China

GQ1885	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ1973	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ229	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ254	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ366	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ657	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ73	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ74-3	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ762	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ856	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ859	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
GQ904	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
MTB_N0010	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	Indonesia
MTB_N0020	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	Mongolia
MTB_N0039	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	Laos
N0002b4	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	Cambodia
N0003b4	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	Vietnam
N0005	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	Indonesia
N0034b4	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	China
N0041b4	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	Vietnam
N0044b4	N/A	Lineage 2 (BJ)	(Comas et al., 2013)	Singapore
1180Cb4	N/A	Lineage 3	(Comas et al., 2013)	Nepal
751Bb4	N/A	Lineage 3	(Comas et al., 2013)	Nepal
855Bb4	N/A	Lineage 3	(Comas et al., 2013)	Nepal
MTB_N0056	N/A	Lineage 3	(Comas et al., 2013)	Vietnam
MTB_N04_read	N/A	Lineage 3	(Comas et al., 2013)	India
MTB_N24_read	N/A	Lineage 3	(Comas et al., 2013)	India
MTB_N37	N/A	Lineage 3	(Comas et al., 2013)	Pakistan
MTB_SG1_12	N/A	Lineage 3	(Comas et al., 2013)	India
NG-5604_BT BH-273	N/A	Lineage 3	(Comas et al., 2013)	Ethiopia
MTB_DY_195	N/A	Lineage 4 (EA)	(Comas et al., 2013)	Ghana
MTB_DY_22	N/A	Lineage 4 (EA)	(Comas et al., 2013)	Ghana
MTB_DY_8	N/A	Lineage 4 (EA)	(Comas et al., 2013)	Ghana
MTB_N0101	N/A	Lineage 4 (EA)	(Comas et al., 2013)	Nicaragua
MTB_N0120	N/A	Lineage 4 (EA)	(Comas et al., 2013)	Puerto Rico
MTB_N0126	N/A	Lineage 4 (EA)	(Comas et al., 2013)	Colombia
MTB_N0131	N/A	Lineage 4 (EA)	(Comas et al., 2013)	USA
MTB_N0135	N/A	Lineage 4 (EA)	(Comas et al., 2013)	USA
MTB_N0137	N/A	Lineage 4 (EA)	(Comas et al., 2013)	Mexico
MTB_N0142	N/A	Lineage 4 (EA)	(Comas et al., 2013)	USA
MTB_N0143	N/A	Lineage 4 (EA)	(Comas et al., 2013)	USA
MTB_N0144	N/A	Lineage 4 (EA)	(Comas et al., 2013)	China
MTB_N0149	N/A	Lineage 4 (EA)	(Comas et al., 2013)	USA
MTB_N0163	N/A	Lineage 4 (EA)	(Comas et al., 2013)	Iran
NG-5604_BT BH-587	N/A	Lineage 4 (EA)	(Comas et al., 2013)	Ethiopia
1001003	N/A	Lineage 5	(Comas et al., 2013)	Sierra Leone
DY21	N/A	Lineage 5	(Comas et al., 2013)	Ghana
MAF_1047301	N/A	Lineage 5	(Comas et al., 2013)	Ghana
MAF_1048001	N/A	Lineage 5	(Comas et al., 2013)	Ghana
MAF_256902	N/A	Lineage 5	(Comas et al., 2013)	Ghana
MAF_348203	N/A	Lineage 5	(Comas et al., 2013)	Ghana

MAF_349404	N/A	Lineage 5	(Comas et al., 2013)	Sierra Leone
MAF_533304	N/A	Lineage 5	(Comas et al., 2013)	Ghana
MAF_553604	N/A	Lineage 5	(Comas et al., 2013)	Ghana
MTB_DY_20	N/A	Lineage 5	(Comas et al., 2013)	Ghana
MTB_DY_26	N/A	Lineage 5	(Comas et al., 2013)	Ghana
541504	N/A	Lineage 6	(Comas et al., 2013)	Ghana
MAF_533604	N/A	Lineage 6	(Comas et al., 2013)	Ghana
MAF_823602	N/A	Lineage 6	(Comas et al., 2013)	Ghana
NG-5604_BT BH-1012	N/A	Lineage 7	(Comas et al., 2013)	Ethiopia
NG-5604_BT BH-935	N/A	Lineage 7	(Comas et al., 2013)	Ethiopia

*Classified according to the international spoligotyping database (Brudey et al., 2006).

**IO = Indo/Oceanic; BJ = East Asian/Beijing; EA = Euro/American.

2.2.2 Mtb culture and growth

Mtb lineage strains culture

Mtb culture and growth was performed by Agnieszka Broda and Dr Gerald Larrouy-Maumus (Larrouy-Maumus group). Mtb strains were cultured in 7H9 Middlebrook broth supplemented with 0.2% glycerol, 0.2% dextrose, 10 mM NaCl, 0.5 g/l Fraction V bovine serum albumin, and 0.05% tyloxapol to mid-log phase. Broth cultures were incubated in a shaking incubator (125 rpm) at 37 °C until reaching mid-log phase (OD₆₀₀ ~0.6) prior to metabolomics studies.

Doubling times were determined according to the formula

$$\text{Doubling time} = \frac{\log 2}{\log OD_{n+1} - \log OD_n} (t_{n+1} - t_n),$$

where $t_{n+1} - t_n$ denotes the difference between two subsequent timepoints, OD_n and OD_{n+1} denote the concentrations at t_n and t_{n+1} , respectively.

For agar media, Mtb was diluted in phosphate buffer saline (PBS) to the required inoculum concentration and spread on 7H11 agar plates supplemented with oleic acid, albumin, dextrose, and catalase (OADC) media and 0.5% glycerol. To measure mycobacterial growth, Mtb was plated on agar plates and incubated for 14 days before counting mycobacterial colonies to determine the bacterial concentration.

2.2.3 Metabolomics analysis of Mtb lineages

Metabolite extraction of Mtb lineage strains

Metabolite extraction was performed by Agnieszka Broda and Dr Gerald Larrouy-Maumus (Larrouy-Maumus group). Mtb strains were grown in supplemented 7H9 Middlebrook broth until mid-log phase as previously described. Mtb was inoculated onto 0.22 μ m nitrocellulose filters under vacuum filtration and placed onto supplemented 7H10 Middlebrook agar media as described above. The bacteria were incubated at 37 °C for 5 doubling times to generate sufficient growth for metabolomics studies. The bacteria were metabolically quenched by plunging the filters into acetonitrile/methanol/H₂O (2:2:1) solution at 4 °C. Mechanical lysis was carried out using 0.1 mm acid-washed zirconia beads for 1 minute using a FastPrep-24 Ribolyser at 6 m/s. The lysate was filtered twice through 0.22 μ m Spin-X column filters (CoStar, UK). A 100 μ l aliquot of the metabolite solution was mixed with 100 μ l of acetonitrile with 0.2% acetic acid at -20 °C before centrifuging at 17,000 x g at 4 °C for 10 minutes. The supernatant was transferred into LC/MS V-shaped vials (Agilent) and 4 μ l of supernatant was injected into the LC/MS instrument for metabolomic analysis.

High-performance liquid chromatography mass-spectrometry

High-performance liquid chromatography mass-spectrometry (HPLC-MS) was performed by Dr Gerald Larrouy-Maumus. Aqueous normal-phase LC was performed using an Agilent 1290 Infinity II LC system equipped with a binary pump, temperature-controlled autosampler set at 4 °C, temperature-controlled column compartment (150 mm x 2.1 mm; dead volume of 315 μ l), and a flow rate of 0.4 mL/min. Elution of polar metabolites was performed using solvent A (consisting of deionized water, resistivity $\sim$ 18 M Ω cm, and 0.2% acetic acid) and solvent B (consisting of 0.2% acetic acid in acetonitrile). The following gradient was used at a flow rate of 0.4 ml/min: 0-minute, 85% B; 0-2 minutes, 85% B; 3-5 minutes, 80% B; 6-7 minutes, 75% B; 8-9 minutes, 70% B; 10-11 minutes, 50% B; 11.1-14 minutes, 20% B; 14.1-25 minutes, 20% B; and 5-minute re-equilibration period at 85% B. Accurate mass spectrometry was performed using an Agilent Accurate Mass 6545 QTOF (Agilent Technologies, US). Chromatography of a reference mass solution using an isocratic pump connected to an ESI ionization source operated in the positive-ion mode was performed before continuous fusion to assure mass-axis calibration. The nozzle voltage was set to 2,000 V and the fragmentor voltage was set to 100 V; the nebulizer pressure was set to 50 psig; the nitrogen drying gas flow rate was set to 5 l/minute; the drying gas temperature was maintained at 300°C; and the MS acquisition rate was 1.5 spectra/sec. The m/z data from 50 to 1,200 were collected. The Agilent Accurate Mass 6546 QTOF has a mass spectrum error of < 5 parts per million, a mass range of 10,000 to 45,000 over the m/z range of 121-955 atomic mass units, and a dynamic range of 100,000-fold with picomolar sensitivity. Data was collected in the centroid 4-GHz (extended dynamic range) mode. Detected m/z data were identified using unique accurate mass-retention time and MS/MS fragmentation identifiers according

to distributions of accompanying isotopomers. Typical variation in metabolite abundances were in the 5 - 10% range.

Bicinchoninic acid assay

The protein concentration of samples used for LC-MS analysis was measured using the Pierce bicinchoninic acid assay Protein assay kit (Thermo Fisher, US) according to manufacturer's instructions. Protein concentration was used to normalise metabolite abundances according to total protein concentration in the original samples prior to the data analysis steps.

Data Analysis

For targeted metabolomic analysis, a list of metabolites of interest was created using Agilent MassHunter Pathways to personal compound database (PCDL) B.08.00 (Agilent technologies, US). For both untargeted and targeted metabolite analyses, molecular features were extracted and/or filtered using Agilent MassHunter Profinder B.08.00 sp3 in accordance with having a mass ranging from m/z 70 – 700 and an absolute minimum abundance of > 5,000. Additionally, Mass Profiler Professional B.15.1 (Agilent technologies, US) was used to perform untargeted analysis.

2.2.4 Genomics

Variant calling

DNA extraction and sequencing was performed as described previously (Comas et al., 2013). A workflow outlining the following steps can be seen in **Figure 2.1**. Quality control checks using FastQC (Andrews, 2010) was performed on FASTQ files. Sequences that passed quality control checks were aligned to the reference genome *M. tuberculosis* H37Rv (GCF_000195955.2_ASM19595v2 assembly, NCBI; (Agarwala et al., 2016)) using the BWA aligner (v0.7.17-r1188; (Li & Durbin, 2010)). Samples were annotated using the *M. tuberculosis* H37Rv annotation file (GCF_000195955.2_ASM19595v2 assembly, NCBI; (Agarwala et al., 2016)). Variant calling was performed using the mpileup function from the Samtools package (H. Li et al., 2009). Variants with a read depth greater than 7 reads and a mapping quality per variant greater than 50 were filtered for using the vcfutils.pl varFilter function (Danecek et al., 2011). The remaining variants were then annotated using SNPeff (Cingolani et al., 2012). The FASTQ files of the Mtb strains in the global dataset were downloaded from the ENA

database under project number PRJEB3223 from the original publication (Comas et al., 2013). These files were treated identically to the files described above, except they were not annotated.

2.2.5 Metabolomics of Mtb lineages and integration analysis

Metabolite analysis

Differential abundance analysis was performed on metabolites using the `lmFit` function in base R with a false discovery rate cut-off of <0.05 comparing each Mtb strain with the H37Rv reference strain. Metabolites which were identified as significantly differentially abundant in individual strains compared to the H37Rv reference strain were taken forward for analysis. Significantly differentially abundant metabolites were clustered and visualised using PCA analysis.

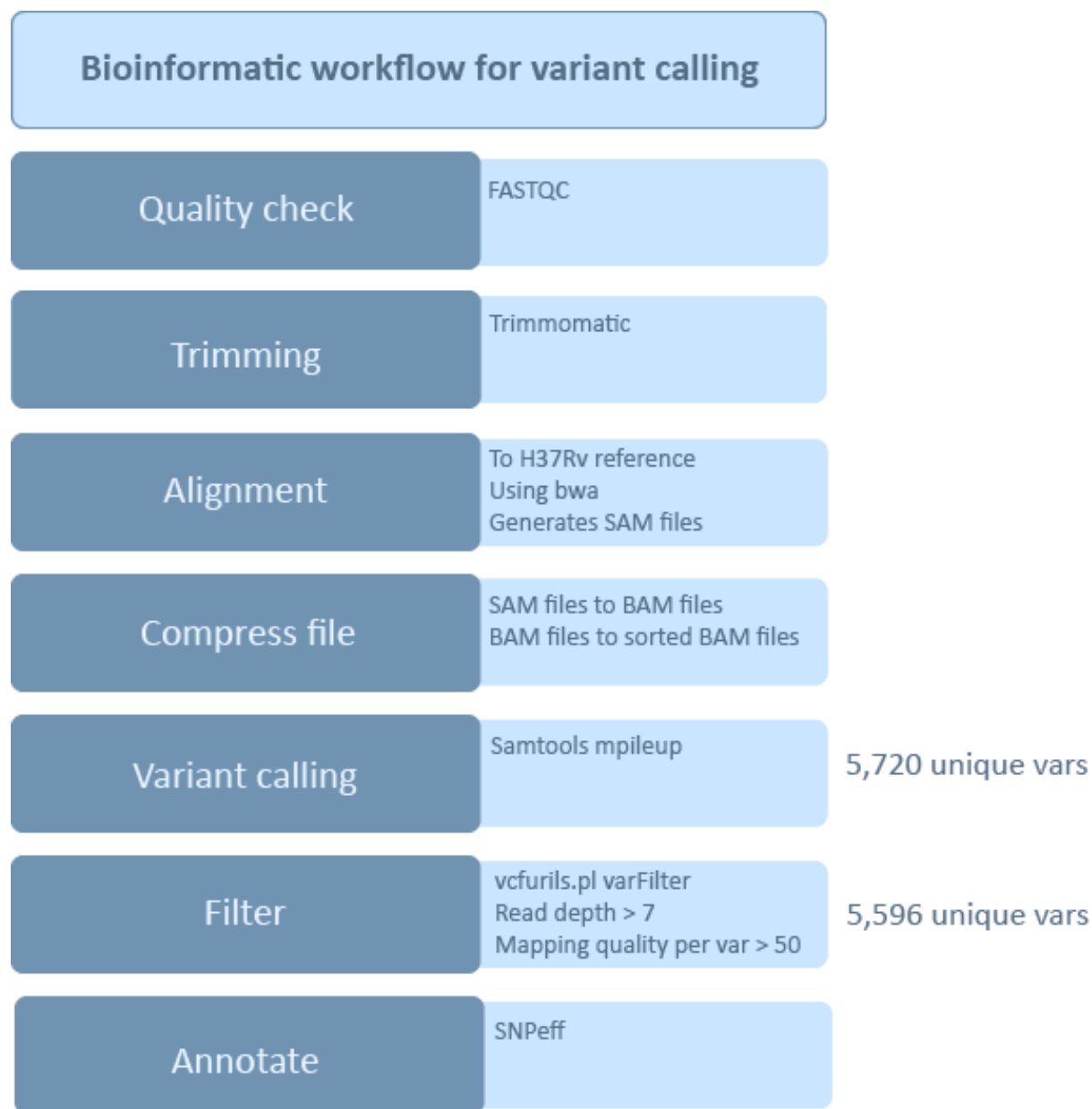


Figure 2.1: Bioinformatic workflow for variant calling analysis. Var(s) = variants.

SNP clustering and functional pathway analysis

SNPs were selected from variants for downstream analysis. Mtb strains were clustered using all SNPs, synonymous SNPs (sSNPs), and nsSNPs individually using a Euclidean distance-based dendrogram.

Functional categories for nsSNPs were obtained from BioCyc (Karp et al., 2019) and were assigned to the genes affected by the nsSNPs. Due to the availability of categories on BioCyc, functional pathways relating to metabolism are overrepresented. For ease of visualisation, specified BioCyc pathways were grouped into 'superfamilies' as shown in **Table 2.2**. These functional categories were filtered to remove unassigned categories and categories with less than 2 genes. nsSNPs were correlated (Pearson's correlation cut-off of absolute 0.99) to identify nsSNPs which were perfectly correlated. Correlated nsSNPs were grouped into 'SNP blocks' for downstream analysis.

Table 2.2: Modified BioCyc pathways. Functional categories were grouped into ‘superfamilies’ to aid in visualisation and interpretation.

Superfamily	Functional categories
Adenine and adenosine salvage	Adenine and adenosine salvage I Adenine and adenosine salvage II Adenine and adenosine salvage VI Adenine and adenosine salvage III
Cholesterol degradation	Super pathway of cholesterol degradation II (cholesterol dehydrogenase) Cholesterol degradation to androstenedione I (cholesterol oxidase) Cholesterol degradation to androstenedione II (cholesterol dehydrogenase)
Aminoimidazole ribonucleotide biosynthesis	5-aminoimidazole ribonucleotide biosynthesis I 5-aminoimidazole ribonucleotide biosynthesis II Super pathway of 5-aminoimidazole ribonucleotide biosynthesis
Adenosine deoxyribonucleotides de novo biosynthesis	Adenosine deoxyribonucleotides de novo biosynthesis Adenosine deoxyribonucleotides de novo biosynthesis II
CDP-diacylglycerol biosynthesis	CDP-diacylglycerol biosynthesis I CDP-diacylglycerol biosynthesis II
L-arginine biosynthesis	L-arginine biosynthesis I (via L-ornithine) L-arginine biosynthesis II (acetyl cycle)
Isoleucine biosynthesis	L-isoleucine biosynthesis I (from threonine) Super pathway of L-isoleucine biosynthesis I
L-proline biosynthesis	L-proline biosynthesis I L-proline biosynthesis III
NADH to cytochrome bd oxidase electron transfer	NADH to cytochrome bd oxidase electron transfer I NADH to cytochrome bd oxidase electron transfer II
Pyrimidine deoxyribonucleotides de novo biosynthesis	Super pathway of pyrimidine deoxyribonucleotides de novo biosynthesis Pyrimidine deoxyribonucleotides de novo biosynthesis I Pyrimidine deoxyribonucleotides de novo biosynthesis III Super pathway of pyrimidine deoxyribonucleotides de novo biosynthesis (<i>E.coli</i>)
Trehalose biosynthesis	Trehalose biosynthesis I Trehalose biosynthesis II Trehalose biosynthesis III Polyacyl trehalose biosynthesis

Mycolate biosynthesis	Super pathway of mycolate biosynthesis Mycolate biosynthesis
Super pathway of guanosine nucleotides de novo biosynthesis	Super pathway of guanosine nucleotides de novo biosynthesis I Super pathway of guanosine nucleotides de novo biosynthesis II Guanosine deoxyribonucleotides de novo biosynthesis I Guanosine deoxyribonucleotides de novo biosynthesis II
Super pathway of adenosine nucleotides de novo biosynthesis	Super pathway of adenosine nucleotides de novo biosynthesis I Super pathway of adenosine nucleotides de novo biosynthesis II
Phosphatidyl glycerol biosynthesis	Phosphatidyl glycerol biosynthesis I (plastidic) Phosphatidyl glycerol biosynthesis II (non-plastidic)
Pentose phosphate pathway	Pentose phosphate pathway Pentose phosphate pathway (non-oxidative branch) Pentose phosphate pathway (oxidative branch) I
Guanine and guanosine salvage	Guanine and guanosine salvage Guanine and guanosine salvage II
L-methionine biosynthesis	L-methionine biosynthesis III Super pathway of L-methionine biosynthesis (by sulfhydrylation)
L-tryptophan biosynthesis	Super pathway of L-tryptophan biosynthesis L-tryptophan biosynthesis
Trehalose degradation	Trehalose degradation II (cytosolic) Trehalose degradation VI (periplasmic)

Representation factor analysis

To identify over-represented genes, each SNP block was analysed at the gene-level. To do this, the number of blocks in which any one gene was included was divided by the theoretical number of blocks which could have included it as assumed from the number of strains where the SNP is present. This representation factor, R , is shown

$$R = \frac{N_{st}/N_{si}}{N_{sg}/N_{gi}},$$

where N_{st} is the total number of blocks theoretically possible in this dataset (14 strains resulting in 16,384 possible blocks) and N_{si} is the total number of blocks in the list of interest (in this case 116 blocks). N_{sg} represents the number of SNP blocks that a given gene could be present in, assuming the number of strains in the dataset with that SNP, and N_{gi} is the number of SNP blocks that contain a given gene.

Integration of nsSNP blocks and metabolites

nsSNP blocks and statistically differentially abundant untargeted metabolites were integrated using the DIABLO suite of functions from the mixOmics Bioconductor package (Singh et al., 2019). Correlation cut-offs were applied as indicated. nsSNP blocks and significantly differentially abundant targeted metabolites were correlated. Correlation p-values were calculated using the Hmisc package in R (Harrell Jr, 2021).

2.3 Methods used to discovery host biomarker signatures for TB

2.3.1 Study design and RNA-seq for biomarkers analysis

Study design and patient cohort definitions

Patients (<16 years old) presenting with suspected pulmonary TB between 2009 and 2013 were recruited to the study at Red Cross Children's Hospital, Cape Town, South Africa by the clinical team from Prof. Heather Zar's group at Red Cross Children's Hospital and the University of Cape Town. Patients were recruited if they presented with cough of any duration as well as either a household TB contact within the previous 6 months, presenting with weight loss (or failed to gain weight within the previous 3 months); a positive tuberculin skin test (TST) to purified protein derivative (PPD: PPD 2TU, PPD RT23, Statens Serum Institute); or a chest radiograph suggesting pulmonary TB. Exclusions included children who had previously received TB treatment or prophylaxis for more than 72 hours, children who were not non-Cape Town residents, children who could not provide informed consent, and in cases where an induced sputum and nasopharyngeal aspirates were not able to be obtained. This resulted in a total of 127 patients recruited who met the criteria described.

Patients were classified into confirmed TB (n = 70) and unlikely TB (other diseases, n = 47) using the NIH consensus definition (Graham et al., 2012). Those classified as confirmed TB were confirmed through culture of *Mtb*, and those classified as unlikely TB were confirmed through a negative *Mtb* culture, not being clinically diagnosed with TB, no prescription of TB treatment, and a documented improvement at a 3-month follow-up visit. Latent TB patients (n = 10) were classified according to no evidence of TB on chest radiograph but positive for reactive Mantoux tests with/without positive Quantiferon results. Latent TB patients were followed up at 3 months. All patients were additionally tested for HIV infection based on the WHO clinical staging involving measurement of CD4 count and HIV viral load (Abbot RealTime HIV-1 assay, Abbott, US). For patients who were diagnosed with confirmed TB, treatment was initiated under the recommendation of the treating doctor and follow-up visits were at 1, 3, and 6 months after for patients diagnosed with confirmed TB but not treated. Treatment response was assessed at follow-up visits through documentation of signs and symptoms and chest radiographs were given after treatment completion.

Blood samples were collected as close to hospital presentation as possible, irrespective of antibiotic use at the time of presentation. Patient classification, to confirmed TB or unlikely TB groups, was carried out by two experienced clinicians using previously defined criteria and was independently made to those managing the patients.

Written informed consent was provided from parents or legal guardians, and assent was given by patients older than 7 years. The study was approved by the Faculty of Health Sciences, Human Research Ethics Committee (HREC, number 389/2017) in association with the parent study (HREC number 045/2008).

RNA extraction and sequencing

RNA extraction was performed by Dr Victoria Wright, Giselle D'Souza, and Dr Samuel Nichols (Levin group). Patient whole blood was collected in PAXgene blood RNA tubes (PreAnalytiX, Germany), stored at -80 °C until further use. RNA was extracted manually or using the QIAcube with the PAXgene Blood RNA Kit (Qiagen, Germany) and DNase treatment was performed using the Turbo DNA-free Kit (Invitrogen, US), both by The Wellcome Centre for Human Genetics (Oxford, UK, <https://www.well.ox.ac.uk/ogc/>). RNA libraries were constructed using the Illumina TruSeq RNA Sample Preparation Kit (Illumina, US). Ribosomal and globin RNA depletion was performed using the ILLUMINA Ribo-Zero Gold kit (Illumina, US). Sequencing of 30 million paired end reads (75 or 100 bp in length) was performed on the Illumina HiSeq 4000 (Illumina, US) by The Wellcome Centre for Human Genetics (UK).

2.3.2 Transcriptomics for *in vivo* gene expression analysis and biomarker discovery

The workflow for the bioinformatics analysis for host transcriptomic differential gene expression analysis and biomarker discovery is shown in **Figure 2.2**.

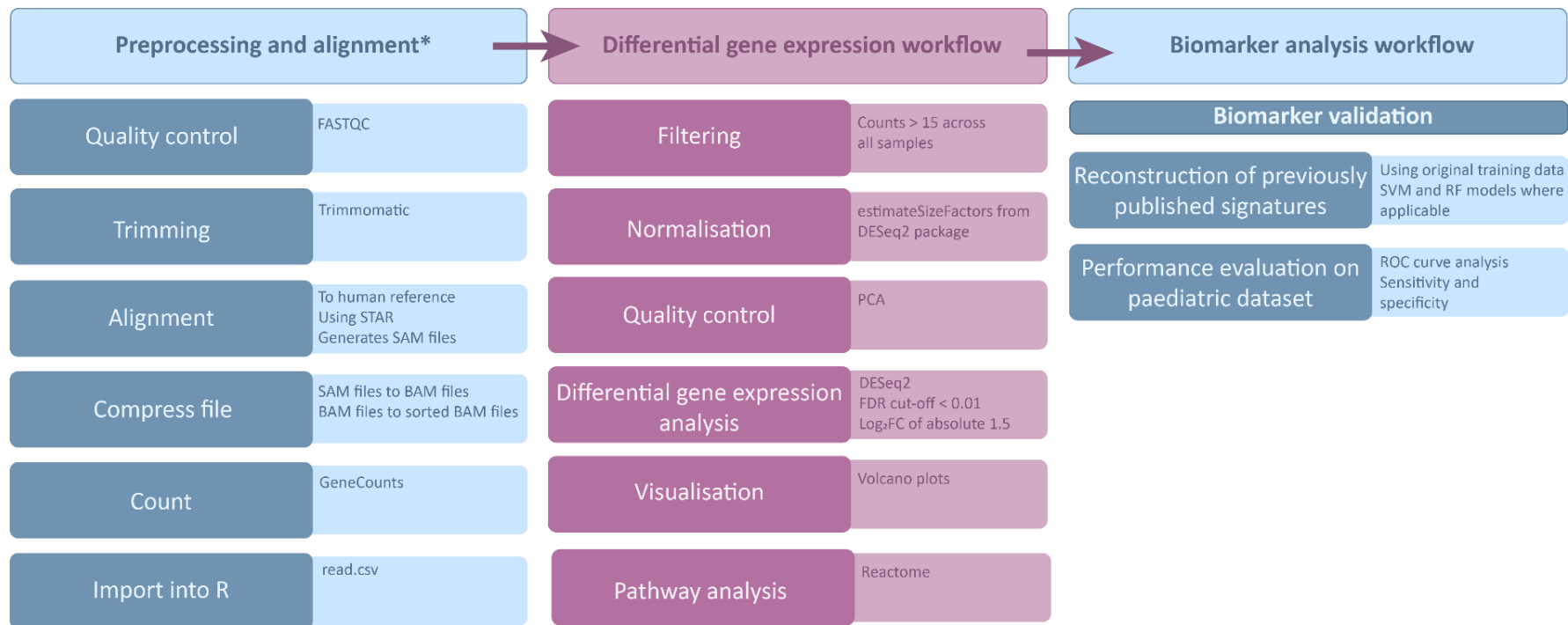


Figure 2.2: Bioinformatics pipeline for host transcriptomics gene expression and biomarker analyses. * = work performed by Dominic Habgood-Coote (Levin group, ICL).

Pre-processing

Data pre-processing was performed in Bash by Dominic Habgood-Coote (Levin group, ICL). Quality control checks of paired end reads was performed using FastQC (Andrews, 2010) and MultiQC (Ewels et al., 2016). Annotations were modified using BEDTools (Quinlan & Hall, 2010). Reads were aligned to the GCh38 reference genome version 89 ensembl (Zerbino et al., 2018) using the STAR aligner (Dobin et al., 2013) and annotation and read counting was performed using SAMtools (Quinlan & Hall, 2010) and FeatureCounts (Liao et al., 2014), respectively.

Quality control, differential gene expression analysis, and pathway analysis

All further analysis was performed using R (R Core Team, 2021). Normalisation was performed using the DESeq2 method (using the estimateSizeFactors function in DESeq2 (Love et al., 2014), before $\log_2$ transformation for visualisation using PCA analysis. Transcripts were $\log_2$ -transformed and normalised using the DESeq2 package. Differential gene expression analysis was performed using DESeq2 and significant transcripts were filtered for using a false discovery rate cut-off of <0.01 and a $\log_2$ -fold change ($\log_2FC$) of absolute 1.5 when comparing active TB with other diseases (OD) and active TB with LTBI. These cut-offs were used to allow for visualisation of significantly differentially expressed genes in volcano plots.

Pathway analysis was performed using Reactome (Gillespie et al., 2022). Gene expression analysis through Reactome was performed using the Camera parameter and variance stabilizing normalisation (vst) RNA-seq counts were used as input.

Feature selection

Feature selection was performed on the significantly differentially expressed transcripts from the comparisons of active TB with OD and active TB with LTBI. Feature selection was performed using an in-house algorithm called feature selection-partial least squared (FS-PLS, (Herberg et al., 2016)). FS-PLS selects the transcript with the highest discrimination between the 2 groups compared, removes all transcripts highly correlated with the chosen transcript, then repeats the process of selecting further transcripts with high discriminatory power until there are no transcripts which add to the discrimination. Cross-validation was applied to the FS-PLS algorithm to select a 3-gene signature over 100 attempts. The feature selection results were visualised using Upset plots.

Published signature evaluation on a paediatric dataset

Host transcriptomic biomarker signatures for tuberculosis (TB) from previous publications which met the following criteria were selected from the literature, 1.) were derived from whole blood, 2.) were designed as a biomarker signature for TB, and 3.) the original publication's training data was available to download to allow for reproduction of the original model. The selected signatures can be seen in **Table 2.3**.

Classification models for each signature were constructed using the method from their original publication. Certain signatures could not be fully replicated as they were generated from microarray datasets, datasets with outdated annotation, or they used genes which were not present in our dataset. For signatures which used multiple probes for individual genes, the probe with the highest expression values were selected. Signatures derived from support vector machine (SVM) models were reconstructed using the kernlab package in R (Karatzoglou et al., 2004) with the training data from the original publication. Signatures derived from random forest models were reconstructed using the randomForest package in R (Liaw & Wiener, 2002) with the training data from the original publication.

The accuracy of each signature at discriminating between active TB and other diseases (OD), and between active TB and LTBI, in the paediatric South African dataset was tested. Batch correction was performed using ComBat_seq function in R (Zhang et al., 2020) to correct for sequencing the cohort on different RNA-seq plates. Data were log₂-transformed and normalised using the estimateSizeFactors function from the DESeq2 package in R (Love et al., 2014). Transcripts with low counts across all samples (less than 15) were excluded from analysis as transcripts with lower number of reads across all samples were unlikely to be identified as differentially expressed and false positive signals due to low read counts could skew the results. ROC curves were created using the pROC package in R (Robin et al., 2011) to measure the accuracy of the signatures. The predict function in base R was used to measure the accuracy of the signatures derived from SVM or random forest models. Sensitivities and specificities of each signature were calculated using the ci.se and ci.spe functions, respectively. When measuring the maximum sensitivity, the specificity was set to the WHO minimal (>82%) and optimal (>92%) guidelines for a rapid non-sputum based diagnostic test for TB (WHO, 2014). Similarly, when measuring the maximum specificity, the sensitivity was set to the WHO minimal (>72%) and optimal (>85%) guidelines. The binom_test function in R was used to calculate the confidence intervals for the area under the ROC (AUROC) curves for each signature.

To create optimised models for each signature, batch correction, $\log_2$ -transformation, DESeq2 normalisation, and filtering to include transcripts with more than 15 counts across all samples was performed as transcripts with lower number of reads across all samples were unlikely to be identified as differentially expressed and false positive signals due to low read counts could skew the results. To identify the signature models with the best fit, i.e. the highest AUC, sensitivity, and specificity, for the paediatric dataset, logistic regression models were fitted to each signature using the `lm` function in base R. The performance statistics for AUC, confidence intervals, sensitivities, and specificities were measured as previously described for each signature comparing active TB from OD and active TB from LTBI.

Table 2.3: Previously published host transcriptomic biomarker signatures for TB. TB = tuberculosis; OD = other diseases; LTBI = latent tuberculosis infection; HC = healthy controls; DRS = disease risk score; dc-RT MLPA = dual-colour-Reverse-Transcriptase-Multiplex-Ligation-dependent-Probe-Amplification; SVM = support vector machine; RF = random forest; SA = South Africa; ML = Malawi; BZ = Brazil; IN = India; UK = United Kingdom; GA = The Gambia; ET = Ethiopia; USA = United States of America; FR = France; VZ = Venezuela.

Signature name	Reference	Target group	Case	Control	Size (genes)	Platform	Signature use	Country
Anderson44	Anderson, et. al., 2014	Children (HIV+/-)	TB (n=87)	OD (n=43)	51	Microarray	DRS	SA, ML
Anderson38	Anderson, et. al., 2014	Children (HIV+/-)	TB (n=87)	LTBI (n=134)	42	Microarray	DRS	SA, ML
Araujo1	de Araujo, et. al., 2016	Adults (unspecified HIV status)	TB (n=6)	HC and LTBI (n=28)	NPC2	RNA-seq	N/A	BZ
Araujo3	de Araujo, et. al., 2019	Adults (unspecified HIV status)	TB (n=8)	LTBI (n=21)	4	RNA-seq	SVM	BZ
Duffy10	Duffy, et. al., 2019	Adults (HIV+/-)	TB (n=93)	LTBI and OD (207)	10	Microarray (from GSE37250 study)	RF	SA
Gjoen11	Gjoen, et. al., 2017	Children (HIV-)	TB (n=47)	HC and OD (n=36)	10	dc-RT MLPA	LASSO regression	IN
Gliddon3	Gliddon, et. al., 2019	Adults (HIV+/-)	TB (n=155)	OD (n=140)	3	Microarray	DRS	SA, ML
Gliddon4	Gliddon, et. al., 2019	Adults (HIV+/-)	TB (n=155)	LTBI (n=167)	4	Microarray	DRS	SA, ML
Huang11	Huang, et. al., 2015	Adults (HIV-)	TB (n=16)	HC and OD (n=79)	13	Microarray	SVM	UK
Kaforou24	Kaforou, et. al., 2013	Adults (HIV+/-)	TB (n=151)	LTBI (n=77)	27	Microarray	DRS	SA, ML
Kaforou33	Kaforou, et. al., 2013	Adults (HIV+/-)	TB (n=151)	OD (n=89)	44	Microarray	DRS	SA, ML
Maerzdorf4	Maerzdorf, et. al., 2016	Adults (HIV-)	TB (n=60)	LTBI and HC (n=40)	4	Microarray and RT-PCR	RF	IN
Maerzdorf12	Maerzdorf, et. al., 2016	Adults (HIV-)	TB (n=60)	LTBI and HC (n=40)	15	Microarray and RT-PCR	RF	IN
Penn6	Penn-Nicholson, et. al., 2019	Adolescents (HIV-)	Progressors (n=46)	Non-progressors (n=107)	6	RNA-seq	RISK6geo score	SA

Qian17	Qian, et. al., 2016	Adults (HIV-)	TB (n=16)	HC and OD (n=69)	17	Microarray	Additive	UK
Roe1	Roe, et. al., 2016	Adults (HIV-)	TB (n=46)	Post recovery (n=31)	BATF2	Microarray	N/A	UK
Roe4	Roe, et. al., 2016	Adults (HIV-)	TB (n=23)	OD (n=35)	4	Microarray	SVM	UK
Roe5	Roe, et. al., 2016	Adults (HIV-)	TB (n=23)	OD and post recovery samples (n=50)	5	Microarray	SVM	UK
Roe2	Roe, et. al., 2020	Adults (HIV-)	TB (n=46)	Post recovery (n=31)	3	Microarray	SVM	UK
Singhania20	Singhania, et. al., 2018	Adults (HIV-)	TB (n=NS)	LTBI and OD (n=NS)	20	Microarray	Modified DRS (GAS6+SEPT 4)	UK
Suliman4	Suliman, et. al. 2018	Adults (HIV-)	Progressors (n=45)	Healthy household contacts (n=141)	4	RNA-seq	-	SA, GA
Suliman2	Suliman, et. al. 2018	Adults (HIV-)	Progressors (n=79)	Healthy household contacts (n=328)	2	RNA-seq	(CD1C+BLK) ANKRD22/O SBPL10	SA, GA, ET
Sweeney3	Sweeney, et. al., 2016	Adults (HIV+/-)	TB (n=296)	LTBI and OD (n=727)	3	Multicohort analysis	(GBP5+DUS P3) /2-KLF2	SA, UK, USA, ML, FR
Verhagen8	Verhagen, et. al., 2013	Children (HIV-)	TB (n=9)	HC and LTBI (n=18)	10	Microarray	RF	VZ
Walter45	Walter, et. al., 2016	Adults (HIV-)	TB (n=24)	LTBI (n=24)	51	Microarray	SVM	USA
Walter35	Walter, et. al., 2016	Adults (HIV-)	TB (n=24)	OD (pneumoniae, n=24)	47	Microarray	SVM	USA
Zak16	Zak, et. al., 2016	Adolescents (HIV-)	Progressors (n=37)	HC (n=77)	16	RNA-seq	SVM	SA

2.4 Methods used to investigate host-pathogen interactions in TB

2.4.1 Mycobacterial culture and growth

Mycobacterium bovis BCG *lux* (BCG *lux*) and *Mycobacterium tuberculosis* (Mtb) strains used in this study can be seen in **Table 2.4**. BCG *lux* was used as a surrogate strain for Mtb, where containment level 2 labs were required for development of experimental workflows and as a result of time constraints.

Table 2.4: Mycobacterial strains used for host-pathogen interaction analyses.

Strain	Origin of acquisition	Resistance
BCG <i>lux</i>	Prof Young's group	Hygromycin 50 µg/ml
Mtb H37Rv	Laboratory collection	None

2.4.2 Human blood donors

Ethical permission was obtained from the Imperial College Healthcare Tissue Bank Review Committee (reference number R18001). Healthy adult donors were recruited from Imperial College London and provided written informed consent. For the dual RNA-seq experiment, all donors identified as the same sex.

2.4.3 Phytohemagglutinin (PHA) and Purified Protein Derivative (PPD) stimulation of whole blood stimulation

Whole blood was diluted at a ratio of 1:10 in Roswell Park Memorial Institute 1640 (RPMI) media. 20 µl of 50 µg/ml phytohaemagglutinin (PHA), 20 µl of 100 µg/ml purified protein derivative (PPD, gifted by the Lalvani lab, ICL), and 20 µl of RPMI control was added to wells of a 96 well plate in quadruplet before adding 180 µl of diluted whole blood to wells. Plates were incubated in a 37 °C incubator with 5% CO₂ humidified atmosphere. Following incubation for 3 days (PHA, RPMI control) or 6 days (PPD, RPMI control), 150 µl of supernatant from corresponding wells was removed and stored at -80 °C for further use.

2.4.4 Infection of whole blood with BCG *lux*

I adapted the WBA which has been described previously (von Both et al., 2018). Blood samples were donated from healthy adults and diluted in equal volumes with RPMI supplemented with glutamine (**Figure 2.3**). Diluted whole blood was aliquoted into 1 ml and 2 ml volumes in bijou tubes and infected with BCG *lux* diluted in PBS to a multiplicity of infection (MOI) of 1:1 and 10:1 (assuming an average of 2×10^5 monocytes/ml in diluted whole blood). Uninfected samples, using the corresponding volume of PBS to mimic BCG *lux* infection volumes, were used as a control. For specific experiments, a bacterial growth control was included, which consisted of 1 ml and 2 ml of 7H9 Middlebrook broth inoculated with an MOI of 1:1 and 10:1 of BCG *lux*. Samples were incubated in a rocking incubator (Stuart equipment, UK) at 37 °C and 20 revolution/minute to continuously mix the blood. The inoculum of an MOI of 1:1 and 10:1 was plated on 7H11 agar plates as described earlier.

At specified time-points (including 0-, 6-, 12-, 24-, 48-, 72-, and 96-hours post-infection), samples were centrifuged (2,000 x g, 10 minutes) and 400 µl of supernatant was transferred into new tubes, double filtered using Spin-X-filters (0.22 µm, Costar, US), and frozen at -80 °C for cytokine analysis. 400 µl of PBS was added to the remaining blood samples to replace the removed supernatant.

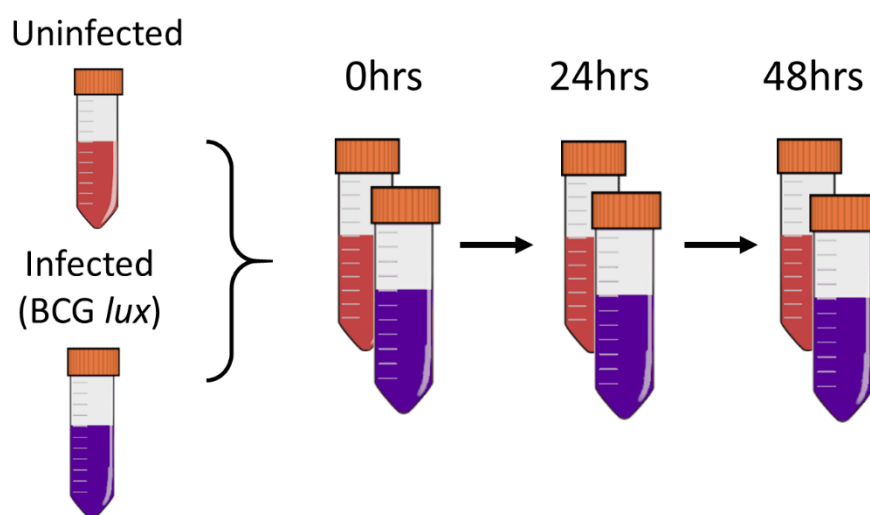


Figure 2.3: Example WBA experiment design. Infected whole blood (with BCG *lux*) and uninfected whole blood are incubated over time, in this case for 0, 24, and 48 hrs after infection.

The samples were aspirated into a Universal tube and 10 ml of sterile deionised H₂O was added. Samples were incubated at room temperature for 10 minutes to allow for red blood cell lysis. Samples were centrifuged (2000 x g, 10 minutes) and the supernatant discarded. 1 ml of PBS and 3 – 5 glass beads were added to each sample and the samples were vortexed for 15 seconds. Samples were diluted to a 1:10 concentration with PBS for luminescence measurements using a luminometer (Berthold, Germany) with 1% n-decyl aldehyde as a substrate. Luminescence measurements were measured by relative light units (RLU) and luminescence ratios were calculated by

$$Luminescence = \frac{RLU \text{ of BCG lux at 96 hrs}}{RLU \text{ of BCG lux at 0 hrs}}$$

2.4.5 Enzyme-linked immunosorbent assay for cytokine analysis of WBA or PPD/PHA assay supernatant

IFN γ enzyme-linked immunosorbent assay (ELISA)

IFN γ and ELISAs were performed on supernatant from the WBA or PPD/PHA assay. **Table 2.5** describes the list of laboratory-made reagents used throughout.

Table 2.5: Laboratory-prepared ELISA reagents.

Solution name	Components	Concentration/quantity	pH
PBS/tween	PBS	1x	N/A
	Tween 20	0.05%	N/A
PBS/tween/FCS	PBS	1x	N/A
	Tween 20	0.05%	N/A
	Sterile fetal calf serum	10%	N/A
Coating buffer	Sodium bicarbonate	8.4 g	8.2
	Distilled water	1 l	
Blocking buffer	PBS	1x	N/A
	Sterile fetal calf serum	10%	N/A
Citric acid buffer	Citric acid	5.19 g	5.0
	Na ₂ HPO ₄ ·2H ₂ O	9 g	
	Distilled water	1 l	

Mouse anti-human IFN γ capture antibody (BD Biosciences, US) was diluted with coating buffer to a concentration of 2 μ g/ml. 96 well-flat-bottomed plates were coated with 50 μ l/well of capture

antibody before incubating overnight at 4 °C. Plates were brought to room temperature and washed twice with PBS/tween. Plates were blocked by adding 200 µl/well of blocking buffer and incubated for 2 hours at room temperature. Plates were washed twice with PBS/tween. Supernatant samples were prepared at 1:4 and 1:8 dilutions using PBS/tween and 100 µl of diluted samples was added in duplicate to corresponding wells. 100 µl of PBS/tween was added in duplicate as negative controls. To generate the standard curve, 10 µl of IFN γ (1 µg/ml) was diluted in 1 ml PBS/tween/FCS to produce the top standard of 10,000 pg/ml. This top standard was double diluted 9 further times to produce a standard curve ranging from 10,000 pg/ml to 20 pg/ml. The plates were sealed and incubated overnight at 4 °C. The plates were brought to room temperature and washed 4 times with PBS/tween. Detection antibody (biotin labelled anti-cytokine IFN γ , BD Biosciences, US) was diluted to 1 µg/ml with PBS/tween/FCS and 100 µl was added to each well. The plates were sealed and incubated for 1 hour at room temperature. The plates were then washed 6 times with PBS/tween before coating. Avidin-peroxidase was diluted to 3.3 µg/ml with PBS/tween/FCS and 100 µl was added to each well of the plates. The plates were sealed and incubated for 30 minutes at room temperature. The plates were washed 8 times with PBS/tween. One 10 mg tablet of o-phenylenediamine-di-hydrochloride (OPD) was dissolved in 25 ml citrate buffer (4 mg/10 ml). 10 µl of hydrogen peroxide was added to the OPD/citrate buffer solution, mixed, and 100 µl of the solution was added to each well. The plates were sealed and incubated in the dark for 20 minutes at room temperature. 100 µl of sulphuric acid (2 M) was added to each well to stop the reaction. The plates were read using the Softmax Pro ELISA Microplate Reader (Molecular Devices, US) at 490 nm.

Tumour necrosis factor-alpha (TNF α) ELISA

TNF α ELISAs were performed on supernatant from the WBA or PPD/PHA assay. The same reagents mentioned in **Table 2.5** were used throughout this TNF α ELISA. Mouse anti-human TNF α capture antibody (BD Biosciences, US) was diluted with coating buffer to a concentration of 2 µg/ml. 96-flat bottomed plates were coated with 50 µl/well of capture antibody before incubating overnight at 4 °C. Plates were brought to room temperature and washed twice with PBS/tween. Plates were blocked by adding 200 µl/well of blocking buffer and incubated for 2 hours at room temperature. Plates were washed twice with PBS/tween. Supernatant samples were prepared at 1:2 and 1:4 dilutions using PBS/tween and 100 µl of diluted samples was added in duplicate to corresponding wells. 100 µl of PBS/tween was added in duplicate as negative controls. To generate the standard curve, 10 µl of TNF α (10 µg/ml) was diluted in 1 ml PBS/tween/FCS before 100 µl of this solution was diluted with 900 µl PBS/tween/FCS to produce the top standard of 10,000 pg/ml. This top standard was double diluted 9 further times to produce a standard curve ranging from 10,000 pg/ml to 20 pg/ml. The plates were

sealed and incubated overnight at 4°C. The plates were brought to room temperature and washed 4 times with PBS/tween. Detection antibody (biotin labelled anti-cytokine TNF α , BD Biosciences, US) was diluted to 1 μ g/ml with PBS/tween/FCS and 100 μ l was added to each well. The plates were sealed and incubated for 1 hour at room temperature. The plates were then washed 6 times with PBS/tween before coating. Avidin-peroxidase was diluted to 3.3 μ g/ml with PBS/tween/FCS and 100 μ l was added to each well in the plate. The plates were sealed and incubated for 30 minutes at room temperature. The plates were washed 8 times with PBS/tween. One 10 mg tablet of OPD was dissolved in 25 ml citrate buffer (4 mg/10 ml). 10 μ l of hydrogen peroxide was added to the OPD/citrate buffer solution, mixed, and 100 μ l of the solution was added to each well. The plates were sealed and incubated in the dark for 20 minutes at room temperature. 100 μ l of sulphuric acid (2 M) was added to each well to stop the reaction. The plates were read using the Softmax Pro ELISA Microplate Reader (Molecular Devices, US) at 490 nm.

2.4.6 Flow cytometry analysis for WBA cell proportion analysis

Antibody titration assay for antibody concentration optimisation

The correct concentration of each antibody used in each experiment needed to be identified before proceeding to perform flow cytometry analysis to enable the maximum fluorescence to be measured while maintaining low reagent costs.

Whole blood samples were diluted 1:1 with RPMI media supplemented with L-glutamine and aliquoted into 2 ml tubes. Samples were lysed at specified time-points using 10 ml lysis buffer (1x, BioLegend, US) for 10 minutes at room temperature whilst in the dark. Samples were centrifuged for 5 minutes at 500 x g, the supernatant discarded, and the pellet resuspended in 1 ml fluorescence-activated cell sorting (FACS) buffer (PBS supplemented with 1% fetal calf serum). 70 μ l of samples were aliquoted into a specified number of wells of a 96 well plate. The plate was centrifuged for 5 minutes at 500 x g and the supernatant discarded. 70 μ l of live-dead stain (Zombie-aqua) was added to specified wells and the plate was left for 15 minutes at 4 °C in the dark (**Table 2.6**). The plate was washed once in FACS buffer and the supernatant discarded. 70 μ l of the staining panel described (**Table 2.6**, BioLegend) at the concentrations shown was added to corresponding wells and the plate was left for 20 minutes at 4 °C in the dark. The plate was washed twice in FACS buffer, and the supernatant was discarded. 200 μ l of BD Fixation Solution (4% paraformaldehyde – BD BioSciences, US) was added to the plate and left for 1 hour at room temperature to fix the cells. The plate was

washed once in FACS buffer, and the pellet was resuspended in 1 μ l of FACS buffer. The plate was stored at 4 °C in the dark for further analysis.

Table 2.6: Fluorochrome panel used for flow cytometry analysis. NK = natural killer cells.

Marker	Target cell	Fluorochrome	LASER	Company	Type	Clone	Dilution
CD3	T- cells	PE-Dazzle594	561	BioLegend	Mouse IgG2a, κ	OKT3	1 in 50
CD19	B-cells	BV650	405	BioLegend	Mouse IgG1, κ	HIB19	1 in 100
CD56	NK cells	PE/Cy7	561	BioLegend	Mouse IgG1, κ	5.1H11	1 in 200
CD14	Monocytes	BV421	405	BioLegend	Mouse IgG2a, κ	M5E2	1 in 100
CD11b	Monocytes, neutrophils, NK cells, granulocytes and macrophages	FITC	488	BioLegend	Rat IgG2b, κ	M1/70	1 in 20
CD15	Neutrophils	BV785	405	BioLegend	Mouse IgG1, κ	W6D3	1 in 50
HLA-DR	Activation marker (B cells, activated T cells, monocytes/macrophages, dendritic cells)	APC/Fire750	633	BioLegend	Mouse IgG2a, κ	L243	1 in 20
Zombie-aqua	Live-Dead (Stains dead cells)	BV500	405	BioLegend	N/A	N/A	1 in 100

Samples were run on the LSRFortessa (BD BioSciences, US) at the St Mary's FACS Facility at Imperial College London with 5 laser lines (405, 488, 561, and 640 nm). Data were collected using the FACSDiva software (BD BioSciences, US). Unstained samples were used as a negative control, single-stained samples were used as a positive control following a 2-fold serial dilution (**Table 2.7**), and fluorescence minus one (FMO) controls were used to estimate positive cell populations.

Cells were gated according to the following gating strategy: all cells (forward-scatter-area [FSC-A]/side-scatter-area [SSC-A]), single cells (forward-scatter-height [FSC-H]/FSC-A), live cells (zombie-aqua/FSC-A), and marker of interest (**Figure 2.4**). FlowJo (Becton, et. al., 2021) was used for data analysis and GraphPad Prism (Version 8.0.0 for Windows, GraphPad Software, San Diego, California USA, www.graphpad.com) was used for visualisation of results.

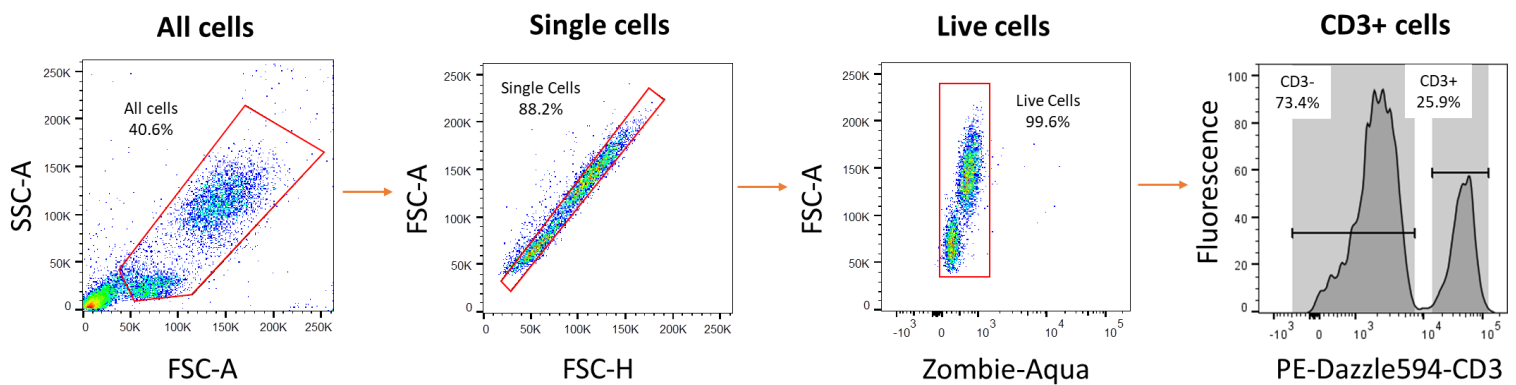


Figure 2.4: Example gating strategy of the CD3 antibody marker (PE-Dazzle594) at a 1 in 28 dilution in uninfected whole blood at 0 hours.

Table 2.7: Serial dilutions used for antibody titrations. Single-stained controls included CD3, CD19, CD56, CD14, CD11b, CD15, and HLA-DR antibodies. Live-dead control includes Zombie-aqua antibody.

Single-stained controls	Live-dead control
1:28	1:100
1:57	1:500
1:114	1:1000
1:228	1:2000

The separation index (SI) for each antibody was calculated as described previously (Maecker et al., 2004) according to

$$SI = \frac{MFI_{pos} - MFI_{neg}}{2 \times SD_{neg}},$$

where MFI denotes the median fluorescence intensity, SD denotes the standard deviation, pos denotes the positively stained cell population for the antibody of interest, and neg denotes the cell population without staining for the antibody of interest (i.e. negative staining).

Flow cytometry analysis of WBA samples

For flow cytometry analysis of WBA samples, samples were prepared as described for the antibody titration assay experiment. Unstained samples were used as a negative control and single-stained samples were used as a positive control following a 2-fold serial dilution. Where applicable, FMO controls were used to estimate positive cell populations. Samples were run on the LSRFortessa as previously described.

Cells were gated according to the following gating strategy: all cells (FSC-A/SSC-A), single cells (FSC-H/FSC-A), live cells (zombie-aqua/FSC-A), and marker of interest if relevant (**Figure 2.5**). FlowJo (BD BioSciences, US) was used for data analysis and GraphPad Prism (GraphPad Software, US) was used for visualisation of results. Compensation was calculated for experiments which used more than 1 stain by using the composition controls (unstained controls, single-stained controls, FMO controls) using the FlowJo software (Becton, et. al., 2021).

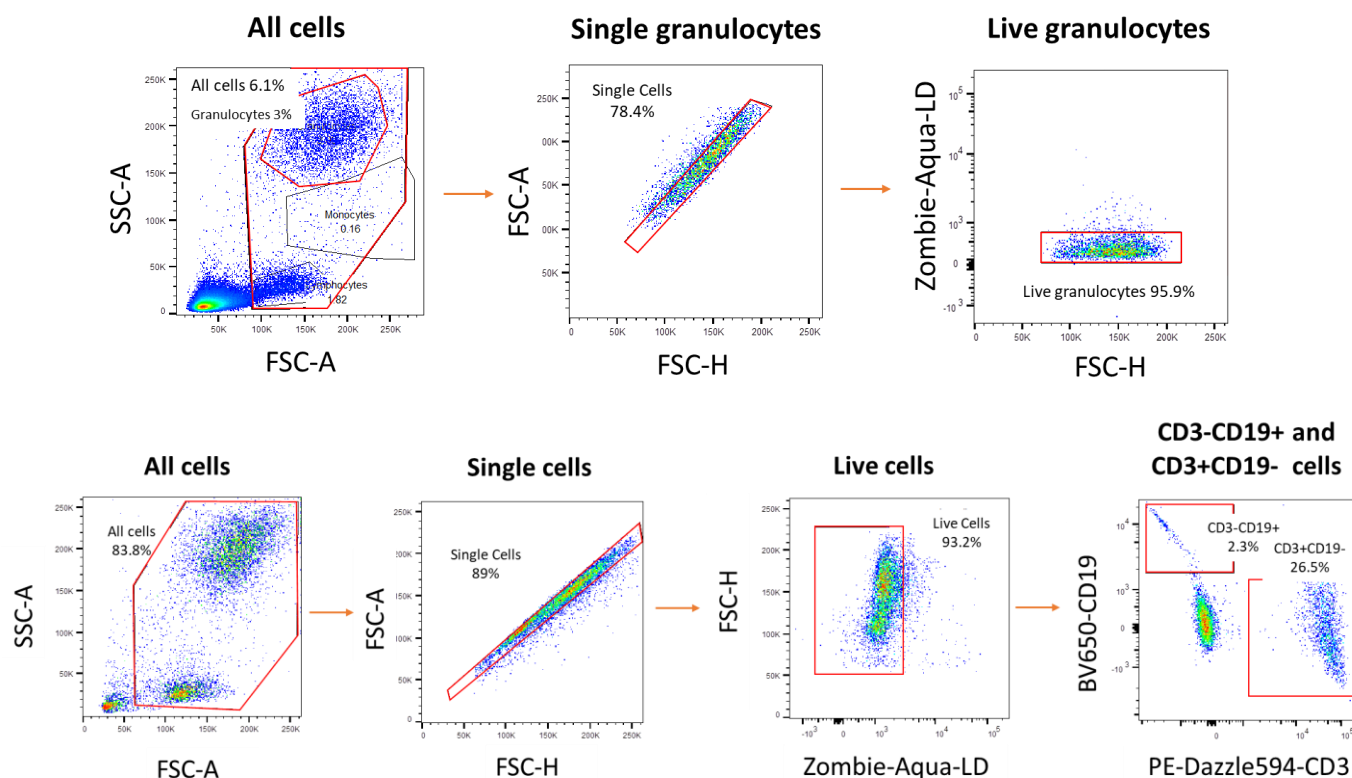


Figure 2.5: Example gating strategies for flow cytometry experiments with and without antibody staining. A) Example gating strategy for flow protocol optimisation and pilot dual RNA-seq analysis. In this example, the final plot shows the percentage of live granulocytes in infected whole blood at 0 hrs post-infection. B) Example gating strategy for final dual RNA-seq analysis with antibody markers, in this case for CD19 and CD3. In this example, CD3-CD19+ B cells and CD3+CD19- T cells are gated in the final plot from infected whole blood at 0 hrs post-infection. LD = live-dead.

2.4.7 RNA extraction optimisation and dual RNA-seq

WBA samples were diluted 3:1 with TRIzol LS (Invitrogen, US) and incubated for 10 minutes at 4 °C with manual mixing by inverting 3 times. Uninfected samples were then stored at -80 °C for further use. For infected samples, 200 µl of 0.1 mm silica beads (MP Biomedicals, US) were added and bead-beating was performed using a FastPrep-24 Ribolyser (MP Biomedicals). Bead-beating was performed for two 45 seconds runs at 6.5 m/s. After the first run, samples were cooled for 5 minutes at 4 °C. After bead-beating, samples were stored at -80 °C for further use.

Samples were thawed at 4 °C prior to extraction. Samples were centrifuged (12,000 x g, 10 minutes, 4 °C) and the supernatant was transferred to tubes containing 200 µl chloroform. Samples were mixed vigorously for 15 seconds and left to rotate for 30 minutes at room temperature. Samples were centrifuged (12,000 x g, 10 minutes, 4 °C) and the aqueous phase containing the RNA was transferred into a tube containing 750 µl chloroform. Samples were mixed vigorously for 15 seconds and centrifuged (12,000 x g, 10 minutes, 4 °C) and the aqueous phase was transferred to a new tube. Ethanol (100%) was added to samples at 1.5x volume before continuing with an adapted protocol from the miRNAeasy kit. Samples were transferred into miRNAeasy spin columns (700 µl at a time), centrifuged (8,000 x g, 30 seconds, room temperature) and the elution was discarded. 700 µl of the RWT buffer was added to each column before columns were centrifuged (8,000 x g, 30 seconds, room temperature) and the elution was discarded. 500 µl of RPE buffer was added to each column and each column was centrifuged (8,000 x g, 30 seconds, room temperature) and the elution was discarded. The previous step was repeated before the columns were centrifuged with open lids (12,000 x g, 2 minutes, room temperature) to dry. The columns were transferred into new elution tubes and 32 µl of RNase free water was added to each column and left for 1 minute at room temperature. The columns were centrifuged (12,000 x g, 1 minute, room temperature) to elute the RNA. RNA was either processed for quality and quantity analysis (described below) or stored at -80 °C for further use.

RNA quality control and quantity measurement

Prior to sequencing, RNA samples were treated to remove contaminating DNA using the TURBO DNA-free kit (Invitrogen, US) according to manufacturer's instructions. Samples were concentrated to meet concentrations requirements for sequencing using the Concentrator plus centrifuge concentrator (Eppendorf, UK) on the aqueous solution setting at 45 °C until reaching the desired concentration.

RNA quantity and quality was measured using the NanoDrop 1000 (PeqLab Biotechnologie, Germany) and the 2100 BioAnalyzer (Agilent Technologies, US). DNA concentration was measured using the Qubit 4 Fluorometer (ThermoFisher, US). Generally, there are ideal cut-offs for RNA quality and quantity using these measurements for standard human samples such as whole blood RNA. The ideal values as measured on the NanoDrop are those above 1.8 when measuring the A260/A280 ratio, which indicates RNA purity, and those above 1.5 when measuring the A230/A260 ratio, which indicates protein contamination. However, for the modified extraction protocol carried out in this thesis, a typical A230/A260 ratio is below 1 and this has previously been known to be adequate for sequencing whole blood samples from our laboratory (data not shown). The Bioanalyzer is mainly

used to measure the RNA integrity (RIN). An ideal RIN is above 7 but this can be affected by measuring RIN in bacterial samples.

Dual RNA sequencing

Extracted RNA which met cut-off criteria of appropriate concentration and RIN values, as described in **Chapter 5.2.6**, was sent to The Wellcome Centre for Human Genetics (UK) for RNA-sequencing. Prior to sequencing, ribosomal RNA (rRNA) was removed from samples using the Illumina Ribo-zero rRNA depletion kit (Illumina, US) according to the manufacturer’s instructions. mRNA was selected before cDNA sequencing was carried out. Paired-end libraries were generated through sequencing using a NovaSeq6000 (Illumina, US).

2.4.8 Transcriptomics for dual RNA-seq analysis

The bioinformatics workflow for the dual RNA-seq analysis is shown in **Figure 2.6**.

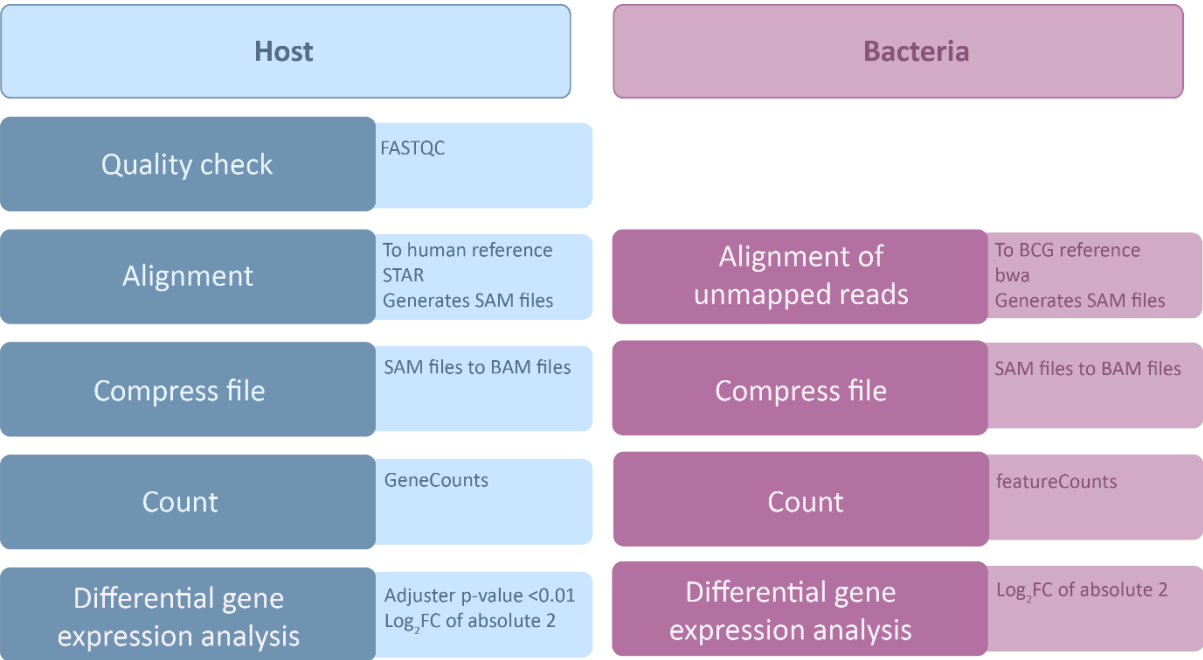


Figure 2.6: bioinformatics workflow for dual RNA-seq analysis.

Pre-processing

Quality control checks of paired end reads was performed using FastQC (Andrews, 2010). Reads were aligned to the GCh38 reference genome version 89 ensembl (Zerbino et al., 2018) using the STAR

aligner (Dobin et al., 2013) and annotation and read counting was performed using SAMtools (Quinlan & Hall, 2010) and GeneCounts (Dobin et al., 2013), respectively. Unmapped reads from the initial alignment to the GCh38 reference genome were used to map reads to the BCG reference genome (*Mycobacterium bovis*_af2122_97.ASM19583v1.37, NCBI; (Agarwala et al., 2016)) using the bwa aligner (H. Li & Durbin, 2010). Annotation and read counting was performed using SAMtools (Quinlan & Hall, 2010) and GeneCounts (Dobin et al., 2013) for human reads and SAMtools (H. Li et al., 2009) and FeatureCounts (Liao et al., 2014) for bacterial reads.

Differential gene expression and pathway analysis

For host reads, transcripts were log₂-transformed and normalised using the DESeq2 package (Love et al., 2014). Differential gene expression analysis was performed using DESeq2 and significant transcripts were filtered for using a false discovery rate cut-off of <0.01 and a log₂-fold change of absolute 2 when comparing time-points.

For bacterial reads, transcripts were DESeq2 normalised (Love et al., 2014). Normalised counts were increased in value by 1 before fold change between time-points and log₂ transformation was performed. This pseudo value increase of 1 was performed to ensure no transcripts had a value of between 0 and 1, as this could not be log₂-transformed.

2.4.9 Transcriptomics for comparison of *in vitro* and *in vivo* Mtb infection models

Transcriptomic analysis to compare the *in vitro* WBA with *in vivo* Mtb infection was performed by a student, Petra Bachanová (Levin group, Imperial College London), whom I supervised.

Datasets used

Publicly available datasets were downloaded from the Gene Expression Omnibus (GEO, (Barrett et al., 2013)). The Accession number for the *in vitro* dataset is GSE108363, the adult *in vivo* dataset is GSE37250, and the paediatric *in vivo* is GSE39941.

Quality control and pre-processing

All datasets were quantile normalised and $\log_2$ transformed. Batch correction was carried out using the ComBat_seq function from the sva package (Leek et al., 2012) in R to correct for the study site differences in the *in vivo* datasets and the cohort differences in the *in vitro* dataset. Only culture confirmed TB and HIV- samples were included in the analysis from the *in vivo* dataset. Sample GSM914578 from the GSE39941 *in vivo* dataset was excluded from analysis due to mislabelling. BCG-infected samples were excluded from the *in vitro* dataset. PCA analysis was used to visualise the clustering of samples based on their gene expression profiles.

Differential gene expression analysis and pathway analysis

The package Limma (Ritchie et al., 2015) was used for differential gene expression analysis. The variables for disease group, study site, sex, and age were accounted for in the linear models for the *in vivo* datasets, and the variables for treatment condition and treatment time were accounted for in the linear model for the *in vitro* dataset. Differentially expressed genes were visualised using a volcano plot with the $\log_2FC$ cut-off of absolute 0.5 and p-value cut-off of <0.05 . Pathway analysis was performed on the datasets using the TopGO package (Adrian & Jorg, 2021). The identified GO terms were modified for visualisation purposes using the rrvgo package (Sergi Sayols, 2020).

Concordance and discordance analysis

Genes which were statistically differentially expressed (according to a p-value cut-off of <0.05) in at least one of the *in vitro-in vivo* comparisons were used for the disco.score analysis using the disco package (Domaszewska et al., 2017). The disco.score is a measurement of discordance and concordance between different variables, in this case gene expression. The size, significance, and direction of the gene expression change are used to generate this score according to

$$disco.score = \log_2FC_{Hs} . \log_2FC_{Mm} . |(\log_{10}P_{Hs} + \log_{10}P_{Mm})|,$$

where $\log_2FC_{Hs}$ is the $\log_2$ fold change for the gene of interest in dataset 1 as calculated in the differential gene expression analysis, $\log_2FC_{Mm}$ is the $\log_2$ fold change for the same gene of interest from dataset 2 as calculated in the differential gene expression analysis, $\log_{10}P_{Hs}$ is the $\log_{10}$ of the p-value for the gene of interest in dataset 1 as calculated in the differential gene expression analysis, and $\log_{10}P_{Mm}$ is the $\log_{10}$ of the p-value of the gene of interest in dataset 2 as calculated in the differential gene expression analysis.

Disco.scores, and associated $\log_2FC$ values, were visualised using a disco plot. The Tmod package (January Weiner, 2020) as used to perform gene set enrichment analysis on each of the four corresponding quadrants of the disco plot. The gene modules identified using Tmod were summarised into superfamily terms for visualisation purposes and to reduced redundancy in category names.

Cibersort cell analysis

CibersortX (Newman et al., 2019) was used to estimate the cell fraction proportions in each sample from all datasets. The cell fractions identified were summarised into B cells, T cells, NK cells, dendritic cells, macrophages, and monocytes for visualisation purposes. The cell proportions were used to calculate a fold change between the control and the disease groups for all cell types. Cell pairings were identified between each *in vitro-in vivo* comparison and a cell disco.score was calculated using the *in vitro-in vivo* $\log_2FC$ and associated p-values. Cell disco.scores were visualised using cell disco plots.

Chapter 3: The host: Host blood RNA signatures for paediatric TB

3.1 Background

Paediatric TB accounts for 12% of the global TB mortality rates (WHO, 2020). Children are much more likely to develop active TB than adults, with those less than 5 years of age and adolescents being at greatest risk, with a dip in risk around 5 – 10 years of age, which increases again after 10 years (Mandal et al., 2017). For example, children aged 1 – 2 years have a 20 – 30% risk of disease following infection. Most paediatric cases also result in disease within the first year of infection, allowing paediatric transmission rates to be good indicators of community circulation (Ki & Shingadia, 2017). Furthermore, infected children also have a higher risk of extrapulmonary TB.

This increased risk of active disease reflects the immature immune response observed in children (Basu Roy et al., 2012). For example, both paediatric macrophage cells and DCs show less phagocytotic function in paediatric patients compared to adults and DCs are also less efficient at antigen presentation in children. These deficient innate responses hamper the initiation of adaptive immunity. In addition, there are several differences in CD4 T cell responses in children compared to adults. Studies suggested that paediatric CD4 T cells are biased towards the Th2 phenotype (Knolle et al., 2020), associated with tissue repair, and Th1 cells, which are key to controlling mycobacterial infection, effector functions are also reduced (Basu Roy et al., 2012). However, more recent analysis of T cell populations in children indicate that there are higher levels of Treg cells in infected children than healthy controls (Basu Roy et al., 2019). While there are many suggested reasons for the increased disease observed in paediatric patients, there are overall few studies with conclusive evidence explaining this.

Despite the higher morbidity and mortality rate, TB diagnosis is often missed in paediatric patients due to the low number of bacilli in sputum samples (Omotayo Togun et al., 2018). The most accurate diagnostic tests for both adult and paediatric TB include smear microscopy, culture, and the Gene Xpert tests. While these tests are effective in adult TB patients, all require a sputum sample which is

difficult to acquire from paediatric TB patients. A novel non-sputum based diagnostic test for paediatric TB is therefore critically needed to help tackle this burden.

Host gene expression signatures in blood have shown to reflect specific diseases, enabling the measurement of the expression of specific genes to indicate the presence of different types of infections and providing the potential of new diagnostic tests (Gliddon et al., 2018). These host blood transcriptomic biomarker signatures for TB do not require a sputum sample and can be reduced to a limited gene size (1 – 3 genes), meaning they can be converted into portable PCR tests and provide rapid results. As a result, host blood transcriptomic biomarker signatures for TB have the potential to be developed into novel diagnostic tests, for both paediatric and adult TB.

A biomarker signature test for TB would improve upon the sputum microscopy and culture tests by reducing time for diagnosis, would not require a sputum sample, and would also be more accessible than the Gene Xpert which requires laboratory space and trained staff to function. Despite this potential, and the high TB burden in children, the majority of current biomarker studies have focussed on the identification of gene signatures for adult TB. However, adult biomarker signatures have proven to be less effective in paediatric cohorts (Omotayo Togun et al., 2018). This is likely due to the different immune responses observed in adults and children, and therefore a separate paediatric TB biomarker test should be developed to allow for testing in children without the requirement of sputum samples. In contrast to the adult biomarker field, there are considerably fewer studies identifying biomarker signatures for paediatric TB, with only 3 studies being published so far (**Table 3.1**) (Anderson, et al., 2014; Gjøen et al., 2017; Verhagen, et al., 2013).

Table 3.1: Host blood transcriptomic signatures for paediatric TB.

HIV = human immunodeficiency virus; TB = tuberculosis; OD = other diseases; DRS = disease risk score; LASSO = least absolute shrinkage and selection operator; HC = healthy controls; dc-RT MLPA = dual-colour-Reverse-Transcriptase-Multiplex-Ligation-dependent-Probe-Amplification; LTBI = latent TB infection; RF = random forest model.

Signature name	Reference	Target group	Case	Control	Size	Platform	Signature use	Country
Anderson44	Anderson, et. al., 2014	Children (HIV+/-)	TB (n=87)	OD (n=43)	51 genes	Microarray	DRS	South Africa, Malawi
Anderson38	Anderson, et. al., 2014	Children (HIV+/-)	TB (n=87)	LTBI (n=134)	42 genes	Microarray	DRS	South Africa, Malawi
Gjoen11	Gjoen, et. al., 2017	Children (HIV-)	TB (n=47)	HC and OD (n=36)	10 genes	dc-RT MLPA	LASSO regression	India
Verhagen8	Verhagen, et. al., 2013	Children (HIV-)	TB (n=9)	HC and LTBI (n=18)	10 genes	Microarray	RF	Venezuela

These paediatric studies took different approaches in identifying a biomarker signature for paediatric TB (Anderson, et al., 2014; Gjøen et al., 2017; Verhagen, et al., 2013). They used both different controls groups to compare against TB patients and certain studies differed on the inclusion of HIV infected individuals in their cohorts. Using different control groups to identify biomarker signatures which are developed into diagnostic tests answers different biological questions. Their use will also depend on the intended country for the diagnostic test developed, as countries with high infectious disease burden may benefit more from a test which distinguishes TB from other childhood infections. While it is biologically interesting to understand the differences, for example between active TB, LTBI, healthy individuals, or asymptomatic household contacts, it is unlikely that a biomarker test would be required to distinguish between these groups in the clinic as active TB is symptomatic. Overall, these different approaches suggest that experts of the field need to define what experimental design is most relevant to diagnose paediatric TB, and what cohort groups are required to develop that test. Relevant groups of researchers, clinicians, and others impacted by the development of a childhood TB diagnostic test, must discuss what type of test is most appropriate to help identify what study design is required.

In 2014, the WHO made progress in defining this through describing guidelines regarding the development of a novel non-sputum based rapid diagnostic test for TB, setting a standard for diagnostic sensitivity and specificity for novel adult and paediatric TB tests (WHO, 2014). These guidelines specify targets for minimal and maximum sensitivity and specificity (**Table 1.2**) and can be used to assess the performance of current biomarker signatures. As these guidelines were not released until 2014, some studies did not use these to validate the performance of their biomarker signatures, including a study by Verhagen, et. al. While the Anderson, et. al., study was also published before the WHO guidelines could be included in their study design, they do describe the sensitivity and specificity measurements of their biomarker signatures (**Table 3.2**) et al., 2014). Their signatures performed well (meeting the WHO guidelines) on their test cohort which comprised of patients from South Africa and Malawi but performed poorly (below the WHO guidelines) when independently validated on a Kenyan cohort. The signature from Gjoen, et. al., also performed well on their test dataset, but was not independently validated on a dataset with a different country of origin (Gjøen et al., 2017).

However, a recent publication by Turner, et. al., has independently validated all published biomarker signatures on their independent cohort from South Africa comprising of 54 pulmonary TB patients and

127 other disease patients of which 24% were HIV positive (Turner et al., 2020). This study identified the AUCs, sensitivities, and specificities of all currently published adult biomarker signatures when stratifying by different measurements. The sensitivities and specificities were only discussed for the top 5 performing signatures, according to AUC, which did not include any of the three paediatric TB studies. These signatures included the Roe1, Roe3, Kaforou25, and Sweeney3 signatures. These signatures approximately met the minimum WHO guidelines for sensitivity and specificity for adult TB diagnostic tests (set at 90% and 70% respectively). However, none of the signatures met the optimal WHO guidelines for sensitivity or specificity (set at 95% and 80% respectively). In addition, all three paediatric signatures evaluated in this study performed relatively poorly with low AUC measurements in this independent dataset (Anderson39.OD performing with an AUC of 79.5%, Anderson39.LTBI performing with an AUC of 60.8%, and Gjoen8 performing with an AUC of 57.2%). While the performance of these signatures is underwhelming, it is expected given the validation cohort in this case was adult patients.

Table 3.2: Signature performances of paediatric biomarker signatures. No data were published describing sensitivity or specificity values in the Turner, et. al., study. dc-RT MLPA = dual colour reverse transcription multiplex ligation dependent probe amplification assay; N/A = not applicable as not described in the original publication; LTBI = latent TB infection; HC = healthy controls.

Signature name	Reference	Platform	AUC		Turner, et. al.	Sensitivity		Specificity	
			Original cohort	Validation cohort		Original publication	Validation cohort	Original publication	Validation cohort
Anderson44	Anderson, et. al., 2014	Microarray	86.2*	89.0**	79.5	78.3*	82.9**	73.5*	82.8**
Anderson38	Anderson, et. al., 2014	Microarray	98.4	100.0	60.8	95.7	94.3	90.9	100.0
Gjoen11	Gjoen, et. al., 2017	dc-RT MLPA	99.0	N/A	57.2	97.9	N/A	97.2	N/A
Verhagen8***	Verhagen, et. al., 2013	Microarray	78.0 (LTBI controls) 94.0 (LTBI, HC, and non-TB pneumonia controls)	N/A	N/A	100.0 (LTBI controls) 78.0 (LTBI, HC, and non-TB pneumonia controls)	N/A	N/A (LTBI controls) 96.0 (LTBI, HC, and non-TB pneumonia controls)	N/A

*Test set from original cohort where training set performance was not available

**Validation cohort including children with positive IGRA in group with other diseases

***Performance values from 5-gene signature as 10-gene signature values were not stated in the original publication

No previous study has validated the signatures derived from these paediatric studies, or indeed any adult signatures, on a paediatric cohort. While it has been suggested that adult signatures will not perform well on paediatric cohorts (Omotayo Togun et al., 2018), no study has aimed to evaluate this. A study evaluating the performances of all the available signatures, paediatric and adult, on an independent paediatric cohort is necessary to understand if separate signatures for these cohorts are required, and if the current paediatric signatures can meet or approximate the current WHO guidelines. As a result, my aims in this chapter are as follows:

- Identify differentially expressed genes in a heterologous cohort of paediatric TB patients from South Africa.
- Identify which biological pathways are significantly different within this cohort.
- Validate previously published adult and paediatric TB biomarker signatures on this cohort to identify if any meet the WHO guidelines for sensitivity and specificity.

3.2 Results

3.2.1 Cohort

Descriptive analysis of the paediatric cohort

Paediatric patients (<16 years old) presenting with suspected pulmonary TB between 2009 and 2013 were recruited to the study at Red Cross Children's Hospital, Cape Town, South Africa by the clinical team from Prof. Heather Zar's group at Red Cross Children's Hospital and the University of Cape Town. Patients were recruited according to the criteria outlined in **Chapter 2.2**.

Whole blood samples for RNA sequencing were provided from 127 patients. Clinical characteristics of this cohort can be seen in **Table 3.3**. Of the 127 patients, 70 (55%) presented with active TB infection, confirmed through sputum culture or a positive Xpert test, and 10 of these active TB patients (14%) presented with extrapulmonary TB. In addition to the active TB patients, 47 patients presented with other diseases, 10 patients (8%) presented with LTBI and 15 (12%) of the total number of patients were HIV positive. The other diseases group included patients with pneumonia (30/47, 64%), TB exposed (7/47, 15%), lower respiratory tract infection (2/47, 4%), upper and lower respiratory tract infection (2/47, 4%), and probably gastro-oesophageal reflux causing cough, non-MTB on sputum culture, severe lower airway obstruction, meningitis, viral infection with croupy cough and anaemia, and upper respiratory tract infection (1/47, 2% each).

Table 3.3: Patient cohort characteristics. TB = tuberculosis; LTBI = latent TB infection; EPTB = extrapulmonary TB; OD = other diseases; ICU = intensive care unit.

	All (n=127)	Confirmed TB (n=70)	LTBI (n=10)	EPTB (n=10)	OD (n=47)
Age in months	36 (18 – 82)	53 (22 – 91)	42 (29 – 57)	53 (35 – 105)	25 (15 – 55)
Female	N (46%)	N (40%)	N (60%)	N (30%)	N (53%)
HIV infected	N (22%)	N (23%)	0 (0%)	N (10%)	N (23%)
Previous TB treatment	Yes (43%) Unknown (13%)	Yes (76%) Unknown (24%)	Yes (20%)	Yes (100%)	No (100%)
Lymphocyte count	4 (3 – 6)	4 (3 – 5)	1 (1 – 1)	3 (2 – 3)	5 (3 – 6)
Lymphocyte %	30 (21 – 46)	30 (22 – 38)	10 (10 – 10)	23 (23 – 24)	42 (22 – 53)
White blood cell count	12 (9 – 15)	12 (9 – 15)	11 (9 – 13)	12 (9 – 13)	13 (10 – 15)
CD4 count*	733 (319 – 1210)	732 (316 – 1134)	NA	230 (230 – 230)	1147 (511 – 1475)
CD4 %*	17 (14 – 25)	16 (14 – 23)	NA	16 (16 – 16)	22 (16 – 32)
Haemoglobin level	10 (9 – 11)	9 (8 – 11)	11 (10 – 12)	11 (8.5 – 11.5)	11 (10 – 12)
Admission to ICU	N (100%)	N (100%)	N (100%)	N (0%)	N (100%)

* HIV positive patients only.

3.2.2 Pre-processing and data exploration

Before performing downstream analyses several pre-processing steps must be performed to detect outliers or confounding variables; these steps are described in **Chapter 2**. First, the transcripts were filtered to exclude those with low to no contribution to the dataset by removing those with less than 15 counts across all samples.

To account for experimental errors such as differences in read depth on different samples, RNA-seq data must be normalised to eliminate noise. It is also beneficial to visualise the effect of the normalisation using PCA plots to identify any other confounding factors such as batch effects, and any outliers, such as samples with mistyped sex when inputting data, before progressing. Data were normalised and visualised using PCA plots (**Figure 3.1**). Data which was not normalised showed five outliers in both the TB and OD, and TB and LTBI datasets (**Figure 3.1A, 3.1C**) which contributed to the large variation observed in PC1 and was removed after normalisation (**Figure 3.1B, 3.1D**).

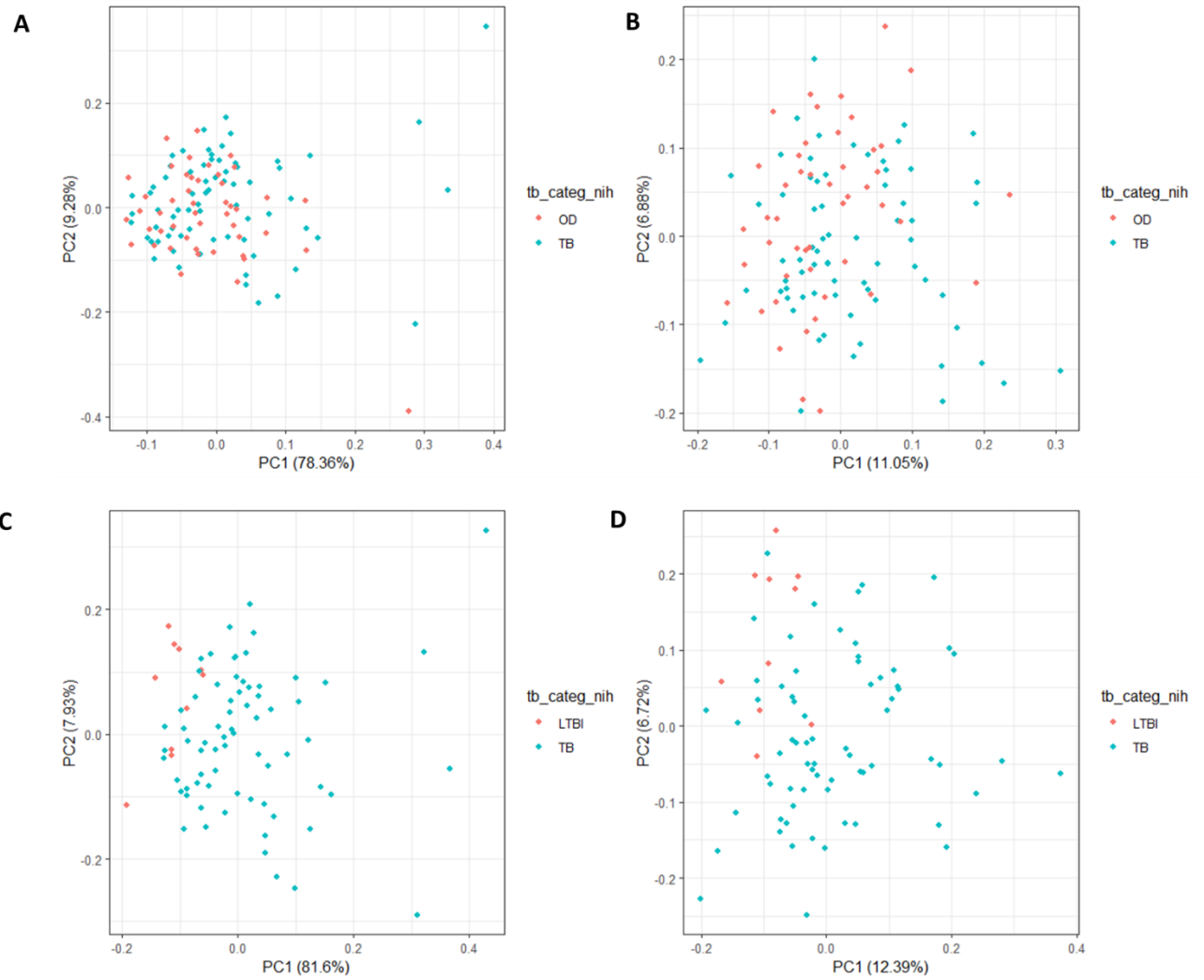


Figure 3.1: Visualisation of dataset before and after normalisation. A.) PCA plot of TB and OD samples before normalisation. **B.)** PCA plot of TB and OD patients after normalisation. **C.)** PCA plot of TB and LTBI patients before normalisation. **D.)** PCA plot of TB and LTBI patients after normalisation.

In addition, batch correction was performed to remove any confounding effects due to the samples being sequenced in batches. Batch correction removed variation observed in PC2 from the TB and OD dataset (**Figure 3.2**), and from PC1 of the TB and LTBI dataset (**Figure 3.3**).

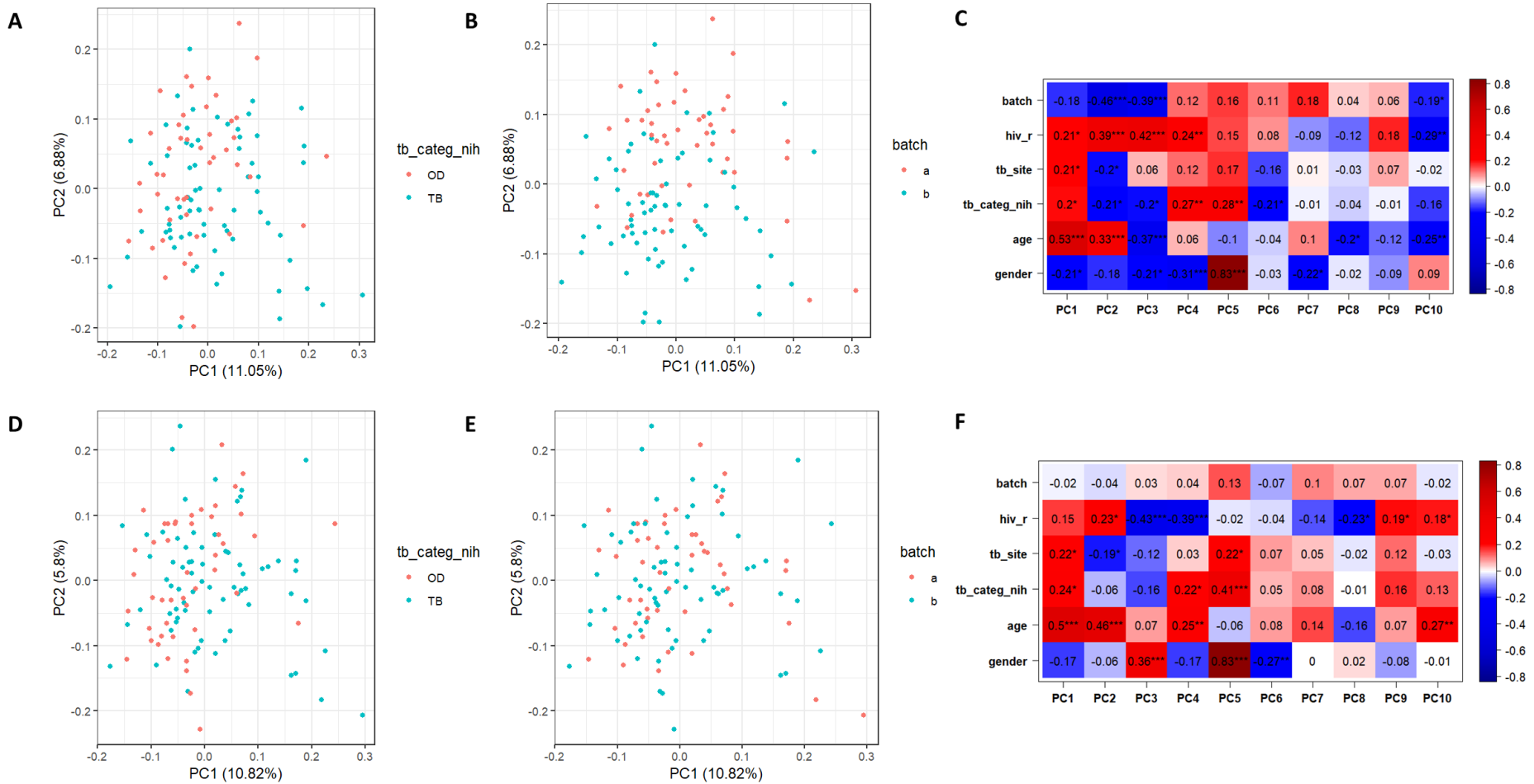


Figure 3.2: Visualisation of the effect of batches on the sample cohort comparing TB and OD patients. A.) PCA plot of TB and OD samples before batch correction, coloured by disease status. **B.)** PCA plot of TB and OD patients before batch correction, coloured by batch. **C.)** Eigencor plot of the first 10 principal components of TB and OD patients before batch correction. **D.)** PCA plot of TB and OD patients after batch correction, coloured by disease status. **E.)** PCA plot of TB and OD patients after batch correction, coloured by batch. **F.)** Eigencor plot of the first 10 principal components of TB and OD patients after batch correction. *tb_categ_nih* = TB status (TB or OD); TB = tuberculosis; OD = other diseases; *hiv_r* = HIV status; *tb_site* = TB status (pulmonary TB, extrapulmonary TB, or pulmonary and extrapulmonary TB).

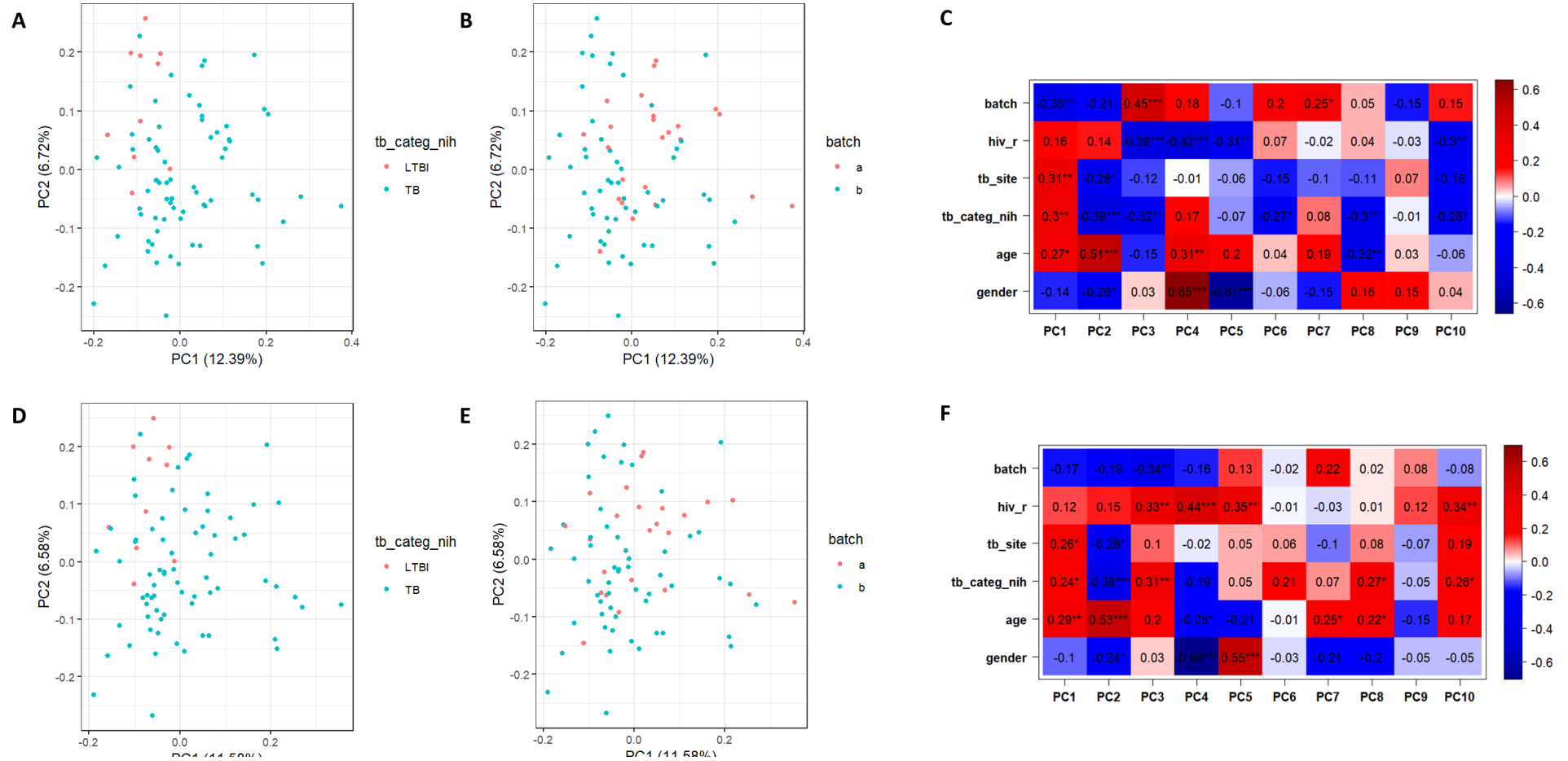


Figure 3.3: Visualisation of the effect of batches on the sample cohort comparing TB and LTBI patients. A.) PCA plot of TB and LTBI samples before batch correction, coloured by disease status. **B.)** PCA plot of TB and LTBI patients before batch correction, coloured by batch. **C.)** Eigencor plot of the first 10 principal components of TB and LTBI patients before batch correction. **D.)** PCA plot of TB and LTBI patients after batch correction, coloured by disease status. **E.)** PCA plot of TB and LTBI patients after batch correction, coloured by batch. **F.)** Eigencor plot of the first 10 principal components of TB and LTBI patients after batch correction. *tb_categ_nih* = TB status (TB or OD); TB = tuberculosis; OD = other diseases; *hiv_r* = HIV status; *tb_site* = TB status (pulmonary TB, extrapulmonary TB, or pulmonary and extrapulmonary TB).

3.2.3 Differential gene expression analysis

TB vs OD

Differentially expressed genes in TB compared to OD patients were identified using a linear model fitted through the DESeq2 package (Love et al., 2014). Patient age and sex was included in the model to account for differences in gene expression attributed to these variables. A total of 1,274 genes were overexpressed and 491 genes were downregulated, according to an adjusted p -value cut-off of < 0.1 . The most significant gene which was differentially expressed was *GBP5* (adjusted p -value of 5.08×10^{-8} and $\log_2$ fold change of 1.5) (**Figure 3.4A**). The statistically significant and most highly up or downregulated genes can be seen in **Figure 3.4B**. There was a majority upregulation of genes observed. The top 50 differentially expressed genes can be found in **Table 3.4**.

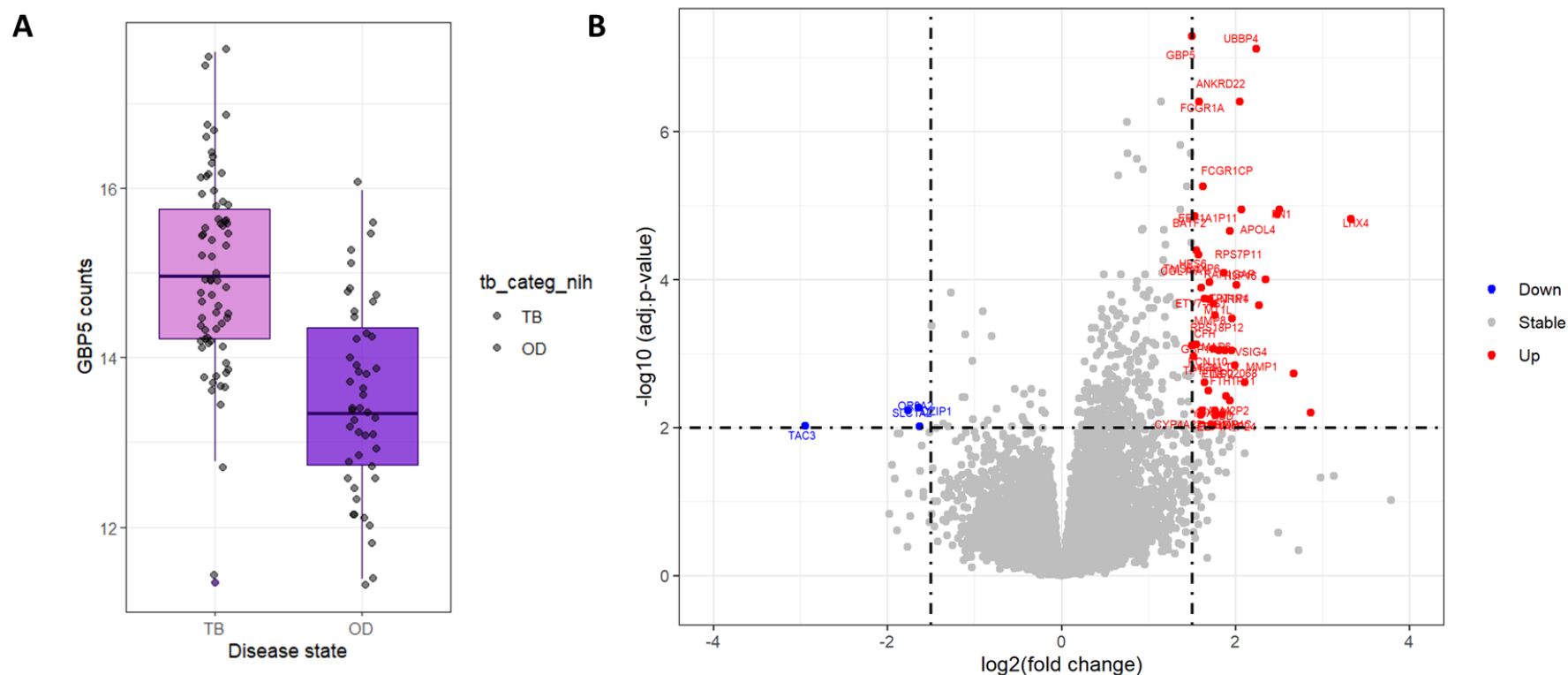


Table 3.4: Top 50 differentially expressed genes which differentiate between TB vs OD patients. NA = not applicable.

Transcript ID	Log ₂ FoldChange	P-value	Adjusted p-value	Gene ID	Gene name	GO term
ENSG00000154451	1.5	1.87E-12	5.08E-08	<i>GBP5</i>	guanylate binding protein 5	cytoplasm
ENSG00000263563	2.2	5.58E-12	7.57E-08	<i>UBBP4</i>	ubiquitin B pseudogene 4	NA
ENSG00000150337	1.6	7.21E-11	3.91E-07	<i>FCGR1A</i>	Fc gamma receptor 1a	membrane
ENSG00000145685	1.1	5.86E-11	3.91E-07	<i>LHFPL2</i>	LHFPL tetraspan subfamily member 2	membrane
ENSG00000152766	2.1	4.75E-11	3.91E-07	<i>ANKRD22</i>	ankyrin repeat domain 22	protein binding
ENSG00000008283	0.8	1.64E-10	7.43E-07	<i>CYB561</i>	cytochrome b561	membrane
					calcium homeostasis modulator family	
ENSG00000188820	1.4	3.96E-10	1.53E-06	<i>CALHM6</i>	member 6	membrane
ENSG00000184500	1.5	6.56E-10	1.98E-06	<i>PROS1</i>	protein S	extracellular space
					uveal autoantigen with coiled-coil	
ENSG00000137831	0.8	6.19E-10	1.98E-06	<i>UACA</i>	domains and ankyrin repeats	nucleus
ENSG00000163251	0.9	8.63E-10	2.34E-06	<i>FZD5</i>	frizzled class receptor 5	protein binding
ENSG00000133112	0.9	1.31E-09	3.24E-06	<i>TPT1</i>	tumor protein, translationally-controlled 1	RNA binding
ENSG00000070501	0.7	1.72E-09	3.88E-06	<i>POLB</i>	DNA polymerase beta	DNA binding
ENSG00000265531	1.6	2.72E-09	5.52E-06	<i>FCGR1CP</i>	Fc gamma receptor 1c, pseudogene	NA
ENSG00000174944	1.4	2.85E-09	5.52E-06	<i>P2RY14</i>	purinergic receptor P2Y14	membrane
ENSG00000115414	2.5	6.84E-09	1.12E-05	<i>FN1</i>	fibronectin 1	protein binding
ENSG00000120217	1.4	7.04E-09	1.12E-05	<i>CD274</i>	CD274 molecule	membrane
ENSG00000100336	2.1	6.86E-09	1.12E-05	<i>APOL4</i>	apolipoprotein L4	NA
ENSG00000243829	2.5	8.72E-09	1.31E-05	N/A	ribosomal protein S9 (RPS9) pseudogene	NA
					basic leucine zipper ATF-like transcription	
ENSG00000168062	1.5	9.67E-09	1.38E-05	<i>BATF2</i>	factor 2	nucleus
ENSG00000121454	3.3	1.12E-08	1.52E-05	<i>LHX4</i>	LIM homeobox 4	nucleus
ENSG00000143776	0.9	1.59E-08	2.05E-05	<i>CDC42BPA</i>	CDC42 binding protein kinase alpha	cytoplasm
						integral
						component of
ENSG00000138119	1.2	1.72E-08	2.12E-05	<i>MYOF</i>	myoferlin	membrane
ENSG00000107566	0.9	1.80E-08	2.13E-05	<i>ERLIN1</i>	ER lipid raft associated 1	membrane
					eukaryotic translation elongation factor 1	
ENSG00000228502	1.9	1.96E-08	2.22E-05	<i>EEF1A1P11</i>	alpha 1 pseudogene 11	
ENSG00000081181	1.5	2.92E-08	3.17E-05	<i>ARG2</i>	arginase 2	metal ion binding
ENSG00000198019	1.2	3.25E-08	3.38E-05	<i>FCGR1B</i>	Fc gamma receptor 1b, pseudogene	NA
ENSG00000144485	1.6	3.98E-08	3.99E-05	<i>HES6</i>	hes family bHLH transcription factor 6	nucleus

ENSG00000230043	1.6	4.77E-08	4.62E-05	<i>TMSB4XP6</i>	TMSB4X pseudogene 6	NA
ENSG00000117228	1.2	5.86E-08	5.48E-05	<i>GBP1</i>	guanylate binding protein 1	NA
ENSG00000213326	1.9	8.96E-08	8.10E-05	<i>RPS7P11</i>	ribosomal protein S7 pseudogene 11	NA
ENSG00000181788	1.2	9.33E-08	8.16E-05	<i>SIAH2</i>	siah E3 ubiquitin protein ligase 2	nucleus
ENSG00000073737	1.0	1.02E-07	8.49E-05	<i>DHRS9</i>	dehydrogenase/reductase 9	membrane
ENSG00000152056	0.5	1.03E-07	8.49E-05	<i>AP1S3</i>	adaptor related protein complex 1 subunit sigma 3	NA
ENSG00000110852	0.7	1.07E-07	8.52E-05	<i>CLEC2B</i>	C-type lectin domain family 2 member B	membrane
ENSG00000178458	2.3	1.28E-07	9.95E-05	<i>H3P16</i>	H3 histone pseudogene 16	NA
ENSG00000143162	0.9	1.39E-07	1.04E-04	<i>CREG1</i>	cellular repressor of E1A stimulated genes 1	extracellular region
ENSG00000217027	1.7	1.49E-07	1.09E-04	<i>TPT1P4</i>	TPT1 pseudogene 4	NA
ENSG00000076864	2.0	1.65E-07	1.17E-04	<i>RAP1GAP</i>	RAP1 GTPase activating protein	regulation of catalytic activity
ENSG00000123500	1.6	1.91E-07	1.30E-04	<i>COL10A1</i>	collagen type X alpha 1 chain	metal ion binding
ENSG00000111666	0.9	1.91E-07	1.30E-04	<i>CHPT1</i>	choline phosphotransferase 1	NA
ENSG00000196205	1.0	2.12E-07	1.40E-04	<i>EEF1A1P5</i>	eukaryotic translation elongation factor 1 alpha 1 pseudogene 5	NA
ENSG00000178222	-1.3	2.32E-07	1.49E-04	<i>RNF212</i>	ring finger protein 212	NA
ENSG00000107960	0.6	2.56E-07	1.61E-04	<i>STN1</i>	STN1 subunit of CST complex	nucleic acid binding
ENSG00000152229	0.9	2.64E-07	1.62E-04	<i>PSTPIP2</i>	proline-serine-threonine phosphatase interacting protein 2	cytoplasm
ENSG00000123836	1.2	2.93E-07	1.73E-04	<i>PFKFB2</i>	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2	nucleotide binding
ENSG00000140105	1.0	2.90E-07	1.73E-04	<i>WARS1</i>	tryptophanyl-tRNA synthetase 1	nucleus
ENSG00000260549	1.6	3.14E-07	1.81E-04	<i>MT1L</i>	metallothionein 1L, pseudogene	NA
ENSG00000124785	1.7	3.29E-07	1.86E-04	<i>NRN1</i>	neuritin 1	nervous system development
ENSG00000260943	1.2	3.43E-07	1.87E-04	<i>LINC02555</i>	long intergenic non-protein coding RNA 2555	NA
ENSG00000197301	1.1	3.44E-07	1.87E-04	<i>HMGA2-AS1</i>	HMGA2 antisense RNA 1	NA

TB vs LTBI

Differentially expressed genes in TB compared to LTBI patients were identified using a linear model fitted through the DESeq2 package (Love et al., 2014). Patient age and sex were included in the model to account for differences in gene expression attributed to these variables. A total of 3,425 genes were overexpressed and 2,608 genes were downregulated, according to an adjusted p -value cut-off of < 0.1 . The most significant gene which was differentially expressed was *SNORD17* (adjusted p -value of 4.07×10^{-24} and $\log_2$ fold change of 2.6) (**Figure 3.4A**). The statistically significant and most highly up or downregulated genes can be seen in **Figure 3.4B**. There was a majority upregulation of genes observed. The top 50 differentially expressed genes can be found in **Table 3.5**.

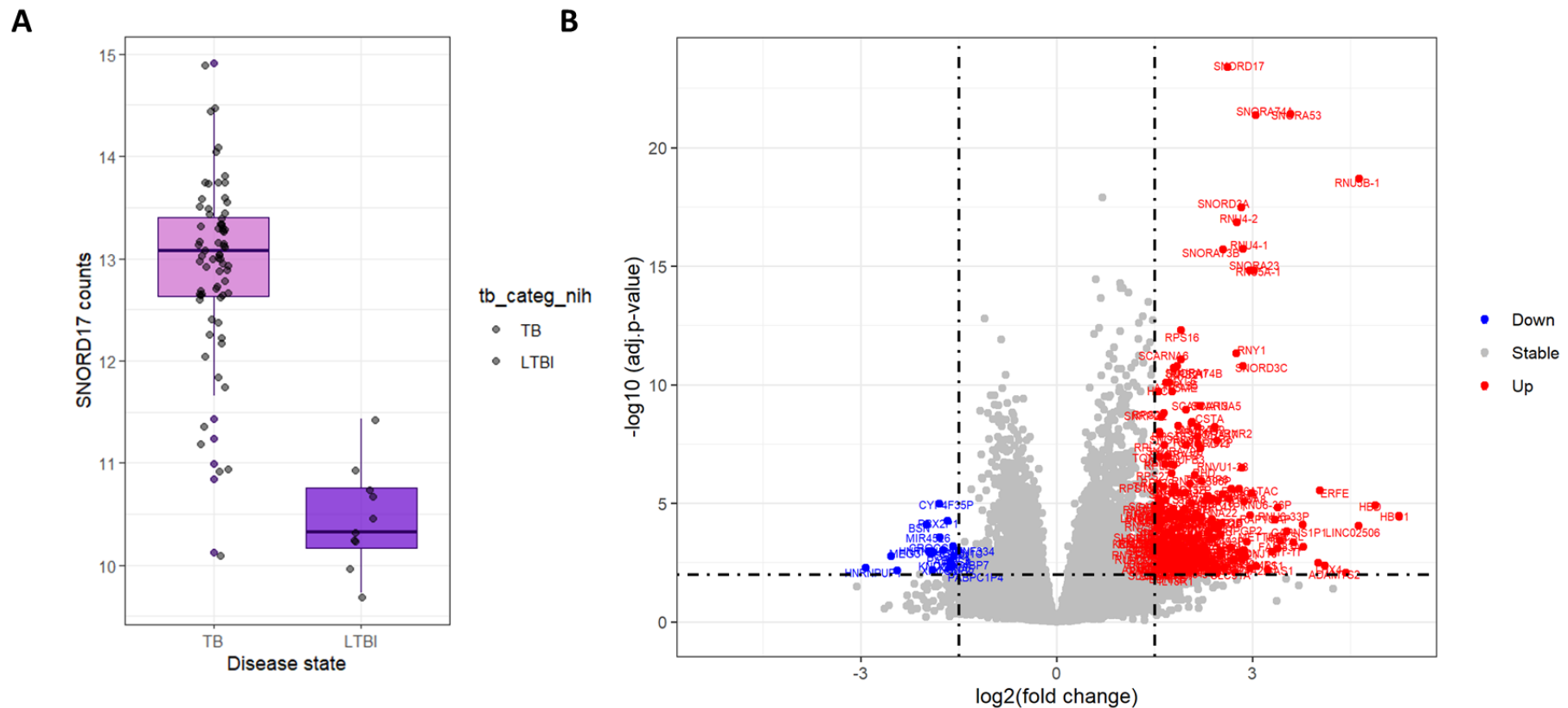


Table 3.5: Top 50 differentially expressed genes which differentiate between TB vs LTBI patients. NA = not applicable.

Transcript ID	Log2FoldChange	P-value	Adjusted p-value	Gene ID	Gene name	GO term
ENSG00000212232	2.6	1.78E-28	4.08E-24	<i>SNORD17</i>	small nucleolar RNA, C/D box 17	RNA processing
ENSG00000212443	3.6	3.33E-26	3.82E-22	<i>SNORA53</i>	small nucleolar RNA, H/ACA box 53	RNA processing
ENSG00000200959	3.1	5.79E-26	4.42E-22	<i>SNORA74A</i>	small nucleolar RNA, H/ACA box 74A	RNA processing
ENSG00000200156	4.6	3.68E-23	2.11E-19	<i>RNU5B-1</i>	RNA, U5B small nuclear 1	spliceosomal tri-snRNP complex assembly
ENSG00000135390	0.7	2.78E-22	1.27E-18	<i>ATP5MC2</i>	ATP synthase membrane subunit c locus 2	NA
ENSG00000263934	2.8	8.92E-22	3.41E-18	<i>SNORD3A</i>	small nucleolar RNA, C/D box 3A	RNA processing
ENSG00000202538	2.8	4.36E-21	1.43E-17	<i>RNU4-2</i>	RNA, U4 small nuclear 2	spliceosomal tri-snRNP complex assembly
ENSG00000200795	2.9	6.72E-20	1.92E-16	<i>RNU4-1</i>	RNA, U4 small nuclear 1	spliceosomal tri-snRNP complex assembly
ENSG00000200087	2.6	7.90E-20	2.01E-16	<i>SNORA73B</i>	small nucleolar RNA, H/ACA box 73B	RNA processing
ENSG00000201998	3.0	7.43E-19	1.56E-15	<i>SNORA23</i>	small nucleolar RNA, H/ACA box 23	RNA processing
ENSG00000199568	3.0	7.47E-19	1.56E-15	<i>RNU5A-1</i>	RNA, U5A small nuclear 1	spliceosomal tri-snRNP complex assembly
ENSG00000168090	0.6	1.85E-18	3.52E-15	<i>COPS6</i>	COP9 signalosome subunit 6	NA
ENSG00000170906	1.0	3.02E-18	5.33E-15	<i>NDUFA3</i>	NADH:ubiquinone oxidoreductase subunit A3	membrane
ENSG00000134825	1.0	5.10E-18	8.34E-15	<i>TMEM258</i>	transmembrane protein 258	NA
ENSG00000104979	1.0	5.66E-18	8.64E-15	<i>C19orf53</i>	chromosome 19 open reading frame 53	NA
ENSG00000106153	1.1	9.08E-18	1.30E-14	<i>CHCHD2</i>	coiled-coil-helix-coiled-coil-helix domain containing 2	nucleus
ENSG00000178982	0.7	1.68E-17	2.26E-14	<i>EIF3K</i>	eukaryotic translation initiation factor 3 subunit K	nucleus
ENSG00000118640	1.4	2.50E-17	3.19E-14	<i>VAMP8</i>	vesicle associated membrane protein 8	membrane
ENSG00000149806	1.3	1.11E-16	1.34E-13	<i>FAU</i>	FAU ubiquitin like and ribosomal protein S30 fusion	nucleus
ENSG00000070047	-1.1	1.46E-16	1.68E-13	<i>PHRF1</i>	PHD and ring finger domains 1	membrane
ENSG00000198258	1.5	1.78E-16	1.94E-13	<i>UBL5</i>	ubiquitin like 5	cytoplasm
ENSG00000125995	1.3	2.35E-16	2.45E-13	<i>ROMO1</i>	reactive oxygen species modulator 1	membrane
ENSG00000128463	0.7	4.16E-16	4.14E-13	<i>EMC4</i>	ER membrane protein complex subunit 4	membrane
ENSG00000105193	1.9	5.23E-16	5.00E-13	<i>RPS16</i>	ribosomal protein S16	RNA binding
ENSG00000180879	1.0	6.17E-16	5.65E-13	<i>SSR4</i>	signal sequence receptor subunit 4	membrane
ENSG00000171135	0.6	8.44E-16	7.44E-13	<i>JAGN1</i>	jagunal homolog 1	membrane

ENSG00000167283	1.3	1.37E-15	1.16E-12	<i>ATP5MG</i>	ATP synthase membrane subunit g	ATP synthesis coupled proton transport
ENSG00000100242	-0.8	1.52E-15	1.25E-12	<i>SUN2</i>	Sad1 and UNC84 domain containing 2	integral component of nuclear inner membrane
ENSG00000137818	1.4	2.11E-15	1.67E-12	<i>RPLP1</i>	ribosomal protein lateral stalk subunit P1	translation mitochondrial electron transport, cytochrome c
ENSG00000131143	0.9	3.51E-15	2.68E-12	<i>COX4I1</i>	cytochrome c oxidase subunit 4I1	to oxygen
ENSG00000103363	1.4	3.98E-15	2.94E-12	<i>ELOB</i>	elongin B	nucleus
ENSG00000201098	2.8	6.68E-15	4.79E-12	<i>RNY1</i>	RNA, Ro60-associated Y1	NA chromosome segregation
ENSG00000166595	0.8	7.29E-15	5.06E-12	<i>CIAO2B</i>	cytosolic iron-sulfur assembly component 2B	cytoplasm
ENSG00000034510	1.3	8.94E-15	6.03E-12	<i>TMSB10</i>	thymosin beta 10	cytoplasm
ENSG00000111481	0.9	1.27E-14	8.35E-12	<i>COPZ1</i>	COPI coat complex subunit zeta 1	cytoplasm
ENSG00000251791	1.9	1.35E-14	8.62E-12	<i>SCARNA6</i>	small Cajal body-specific RNA 6	RNA processing
ENSG00000111775	1.2	1.97E-14	1.19E-11	<i>COX6A1</i>	cytochrome c oxidase subunit 6A1	NA
ENSG00000119632	1.3	1.97E-14	1.19E-11	<i>IFI27L2</i>	interferon alpha inducible protein 27 like 2	NA
ENSG00000143612	0.8	3.00E-14	1.68E-11	<i>C1orf43</i>	chromosome 1 open reading frame 43	membrane
ENSG00000212402	1.9	2.99E-14	1.68E-11	<i>SNORA74B</i>	small nucleolar RNA, H/ACA box 74B	RNA processing
ENSG00000264940	2.9	2.87E-14	1.68E-11	<i>SNORD3C</i>	small nucleolar RNA, C/D box 3C	RNA processing
ENSG00000171858	1.8	3.55E-14	1.94E-11	<i>RPS21</i>	ribosomal protein S21	RNA binding
ENSG00000197958	1.1	3.71E-14	1.98E-11	<i>RPL12</i>	ribosomal protein L12	NA
ENSG00000142507	1.0	4.04E-14	2.11E-11	<i>PSMB6</i>	proteasome 20S subunit beta 6	nucleus
ENSG00000106803	0.9	4.17E-14	2.12E-11	<i>SEC61B</i>	SEC61 translocon subunit beta	membrane
ENSG00000114391	1.5	4.32E-14	2.15E-11	<i>RPL24</i>	ribosomal protein L24	NA
ENSG00000125356	1.8	4.45E-14	2.17E-11	<i>NDUFA1</i>	NADH:ubiquinone oxidoreductase subunit A1	membrane
ENSG00000011304	-0.8	7.96E-14	3.80E-11	<i>PTBP1</i>	polypyrimidine tract binding protein 1	nucleus
ENSG00000112695	1.5	8.55E-14	4.00E-11	<i>COX7A2</i>	cytochrome c oxidase subunit 7A2	membrane
ENSG00000131495	1.1	9.06E-14	4.15E-11	<i>NDUFA2</i>	NADH:ubiquinone oxidoreductase subunit A2	NA

3.2.4 Pathway analysis of significantly differentially expressed genes

To identify the biological pathways which are different between the groups in this study, pathway analysis was performed using Reactome (Gillespie et al., 2022) with batch-corrected vst normalised values. While the pathways identified are not necessarily important for biomarker studies, as some algorithms may choose genes for their signatures which do not have a significant biological impact but do contribute to separate the case and control groups, they can still be useful to characterise and understand the dataset used.

When comparing TB with OD, 2496 pathways were identified and the significant pathways affected (false-discovery rate, FDR, < 0.01) were cell cycle, DNA repair and DNA replication. Those nearing significance (FDR < 0.1) included autophagy (FDR 0.07), chromatin organisation (FDR 0.08), the immune system (FDR 0.07) and reproduction (FDR 0.07). Of these, the immune system showed an overall up-regulation, whereas the other top pathways showed a down-regulation (**Figure 3.6**). The top 50 pathways, by FDR, can be seen in **Table 3.6**. As expected, among the top significantly overexpressed pathways was IFN γ signalling, however IFN α/β signalling and neutrophil degranulation were also increased alongside a deactivation of monocyte inflammatory pathways (inactivation of CSF3 (G-CSF) signalling) (Hamilton, 2019). The majority of the top significant pathways were underexpressed including those involved in B cell pathways (CD22 mediated BCR signalling and antigen activates B cell receptor (BCR) leading to generation of second messengers), complement pathways (classical antibody-mediated complement activation, creation of C4 and C2 activators, and initial triggering of complement), and cell cycle pathways and DNA repair (homology directed repair, cell cycle checkpoints, mitotic spindle checkpoint, DNA double-strand break repair, mitotic prometaphase, G2/M checkpoints, and cell cycle).

When comparing TB with LTBI, again 2496 pathways were identified and the significant pathways affected (FDR < 0.01) were cellular responses to stimuli and DNA replication, with the immune system having a FDR of 0.55. The direction of these pathways was less clear, with most involving both up and down-regulation of pathways within these groups (**Figure 3.7**), which may be driving the lower significance rates observed in this comparison. The top 50 pathways, by FDR, can be seen in **Table 3.7**. Unlike the previous comparison, the top significant pathways in TB vs latent TB patients were all overregulated. Of those, several pathways were involved in translation (viral mRNA translation, eukaryotic translation elongation, eukaryotic translation termination, eukaryotic translation initiation, cap-dependent translation initiation, translation, translation initiation complex formation,

mitochondrial translation termination, mitochondrial translation, and mitochondrial translation elongation), reflecting the high differential expression of *SNOR* genes, and protein assembly (peptide chain elongation, SRP-dependent cotranslational protein targeting to membrane, formation of a pool of free 40S subunits, GTP hydrolysis and joining of the 60S ribosomal subunit, and ribosomal scanning and start codon recognition). However, pathways involved in viral responses were also identified (viral mRNA translation, influenza viral RNA transcription and replication, and influenza infection), which may reflect an immune response to intracellular pathogens.

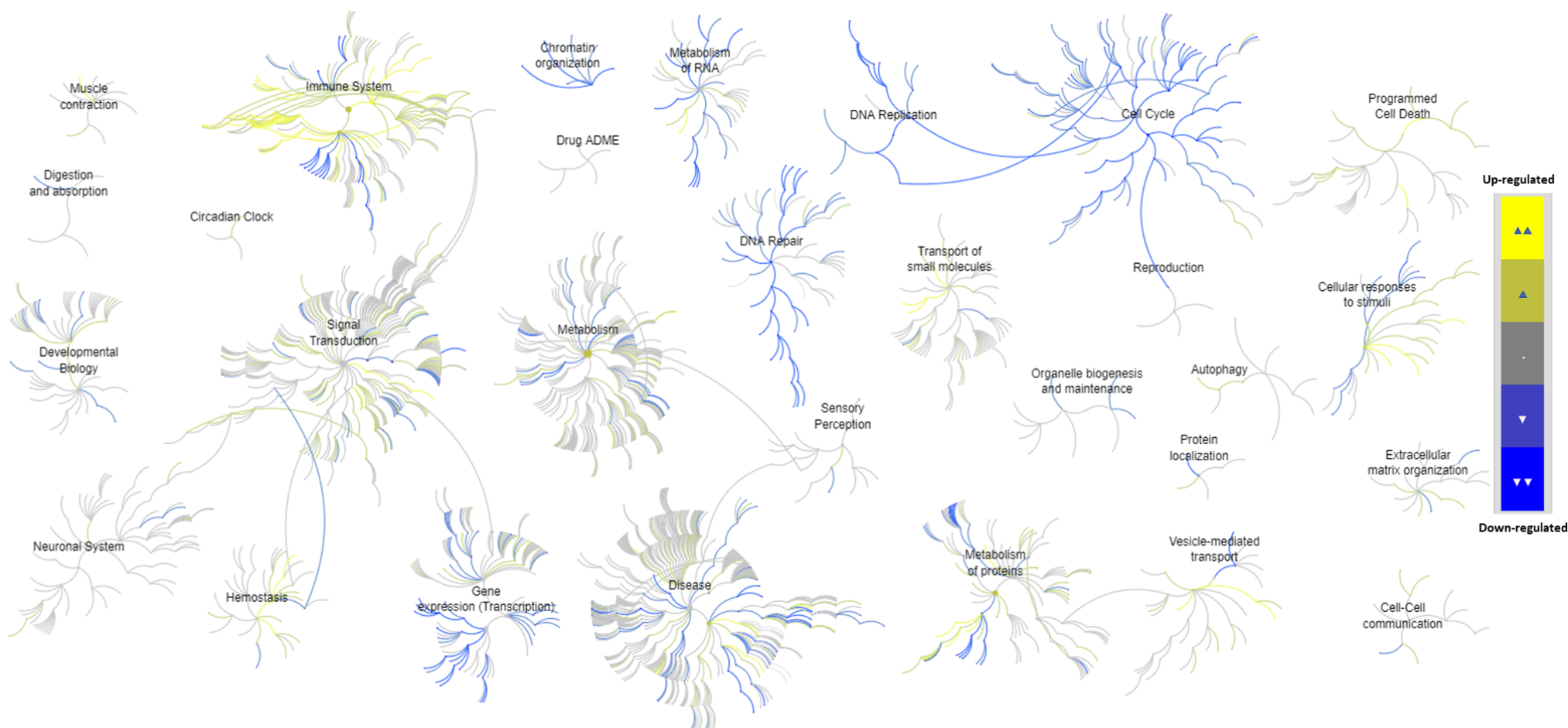


Figure 3.6: Pathway analysis of TB vs OD patients using Reactome (Gillespie et al., 2022). Each line represents a sub-pathway. Significantly overexpressed pathways are shown in deep yellow; non-significant overexpressed pathways are shown in grey-yellow; non-significant pathways with no direction change are shown in grey; non-significant down-regulated pathways are shown in grey-blue; significant down-regulated pathways are shown in deep blue. Significance is measured as adjusted p -value < 0.05 .

Table 3.6: Top 50 significant pathways when comparing TB and OD groups from Reactome pathway analysis. Ordered and filtered by FDR < 0.01.

Name	Direction	FDR	Number of genes	Average fold change
Interferon gamma signaling	Up	8.3E-09	91	2.4E-01
Inactivation of CSF3 (G-CSF) signaling	Up	4.3E-04	21	2.3E-01
Interferon alpha/beta signaling	Up	5.1E-04	70	1.4E-01
Neutrophil degranulation	Up	6.4E-04	478	1.0E-01
CD22 mediated BCR regulation	Down	4.7E-08	67	-3.7E-01
Unwinding of DNA	Down	4.7E-08	12	-3.4E-01
DNA strand elongation	Down	4.7E-08	30	-2.0E-01
Classical antibody-mediated complement activation	Down	5.2E-08	84	-3.3E-01
Activation of ATR in response to replication stress	Down	6.3E-08	36	-2.1E-01
FCGR activation	Down	9.0E-08	90	-3.2E-01
Activation of the pre-replicative complex	Down	1.9E-07	33	-2.0E-01
Scavenging of heme from plasma	Down	1.9E-07	88	-3.1E-01
Creation of C4 and C2 activators	Down	5.2E-07	91	-3.0E-01
Role of phospholipids in phagocytosis	Down	6.3E-07	103	-2.9E-01
Role of LAT2/NTAL/LAB on calcium mobilization	Down	4.3E-06	91	-3.0E-01
FCERI mediated Ca ²⁺ mobilization	Down	4.3E-06	105	-2.7E-01
Homology Directed Repair	Down	9.4E-06	115	-1.1E-01
Amplification of signal from the kinetochores	Down	1.0E-05	93	-1.3E-01
Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal	Down	1.0E-05	93	-1.3E-01
Cell Cycle Checkpoints	Down	1.2E-05	269	-9.9E-02
FCERI mediated MAPK activation	Down	1.7E-05	106	-2.6E-01
Chromosome Maintenance	Down	1.9E-05	116	-1.1E-01
Mitotic Spindle Checkpoint	Down	2.9E-05	110	-1.1E-01
FCGR3A-mediated IL10 synthesis	Down	3.0E-05	114	-2.4E-01
G1/S-Specific Transcription	Down	3.0E-05	29	-2.1E-01
Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	Down	3.4E-05	92	-2.6E-01
Deposition of new CENPA-containing nucleosomes at the centromere	Down	3.5E-05	52	-1.7E-01
Nucleosome assembly	Down	3.5E-05	52	-1.7E-01
DNA Double-Strand Break Repair	Down	3.5E-05	144	-9.2E-02
Initial triggering of complement	Down	3.6E-05	99	-2.6E-01
HDR through Homologous Recombination (HRR) or Single Strand Annealing (SSA)	Down	5.3E-05	109	-1.0E-01

Condensation of Prophase Chromosomes	Down	7.8E-05	41	-1.8E-01
Condensation of Prometaphase Chromosomes	Down	1.6E-04	11	-2.8E-01
Cyclin A/B1/B2 associated events during G2/M transition	Down	1.6E-04	25	-2.0E-01
ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression	Down	2.4E-04	43	-1.5E-01
Resolution of Sister Chromatid Cohesion	Down	2.4E-04	125	-1.0E-01
Mitotic Prometaphase	Down	2.4E-04	202	-8.4E-02
Binding and Uptake of Ligands by Scavenger Receptors	Down	2.5E-04	117	-2.0E-01
Defective pyroptosis	Down	3.5E-04	40	-1.6E-01
G2/M Checkpoints	Down	3.5E-04	146	-9.2E-02
Processing of DNA double-strand break ends	Down	4.0E-04	76	-1.0E-01
PRC2 methylates histones and DNA	Down	5.2E-04	40	-1.6E-01
Mitochondrial translation elongation	Down	6.3E-04	88	-6.9E-02
Diseases associated with N-glycosylation of proteins	Down	6.4E-04	17	-9.9E-02
Polo-like kinase mediated events	Down	7.9E-04	16	-2.4E-01
Telomere Maintenance	Down	8.0E-04	89	-9.8E-02
EML4 and NUDC in mitotic spindle formation	Down	8.0E-04	116	-9.1E-02
Transcriptional regulation by small RNAs	Down	8.2E-04	73	-1.0E-01
RNA Polymerase I Promoter Clearance	Down	8.7E-04	77	-9.9E-02
Cell Cycle	Down	8.7E-04	653	-6.7E-02

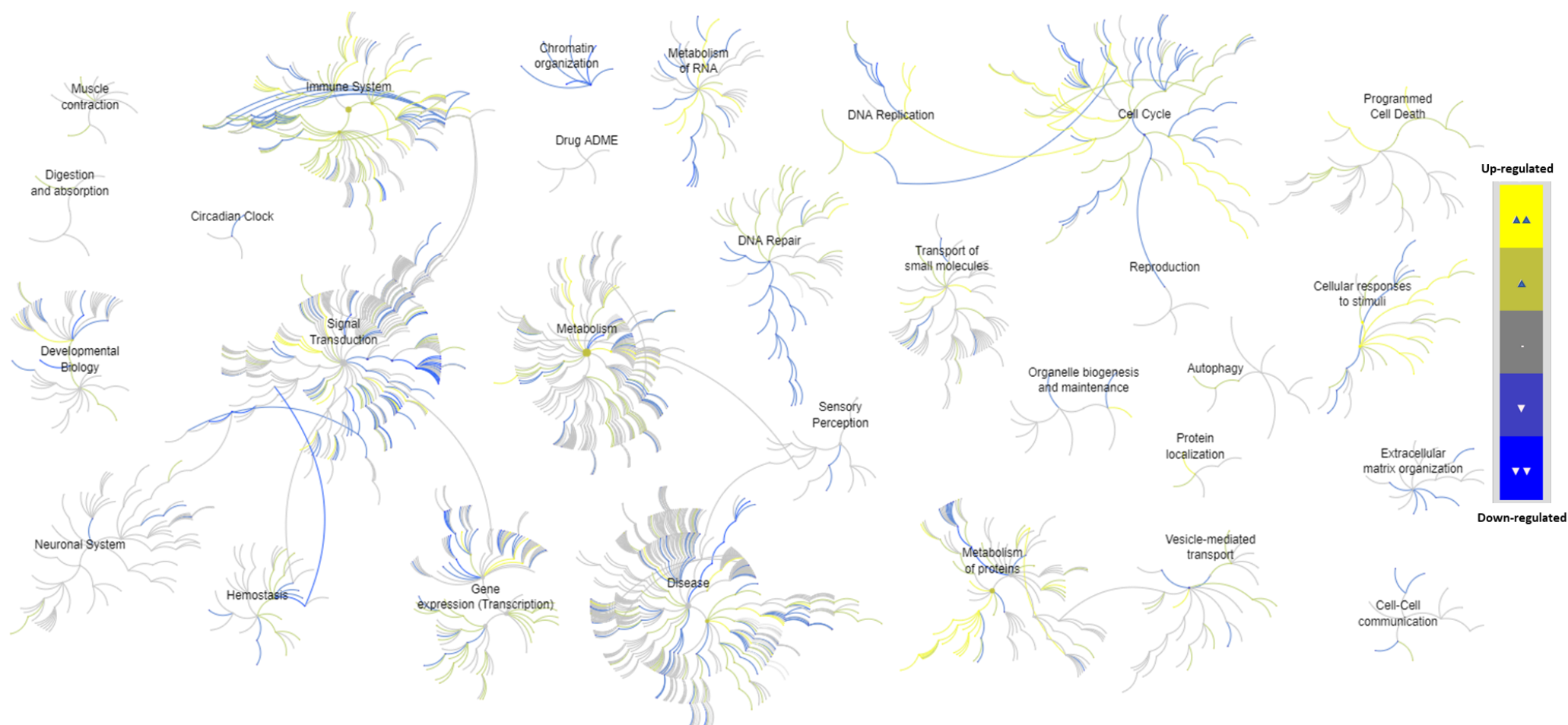


Figure 3.7: Pathway analysis of TB vs latent TB patients using Reactome (Gillespie et al., 2022). Each line represents a sub-pathway. Significantly overexpressed pathways are shown in deep yellow; non-significant overexpressed pathways are shown in grey-yellow; non-significant pathways with no direction change are shown in grey; non-significant down-regulated pathways are shown in grey-blue; significant down-regulated pathways are shown in deep blue. Significance is measured as adjusted p -value < 0.05 .

Table 3.7: Top 50 significant pathways when comparing TB and LTBI groups from Reactome pathway analysis. Ordered and filtered by FDR < 0.01.

Name	Direction	FDR	Number of genes	Average fold change
Viral mRNA Translation	Up	1.3E-28	88	8.1E-01
Peptide chain elongation	Up	1.3E-28	88	8.2E-01
SRP-dependent cotranslational protein targeting to membrane	Up	3.7E-28	111	7.1E-01
Formation of a pool of free 40S subunits	Up	9.7E-28	100	7.6E-01
Eukaryotic Translation Elongation	Up	1.2E-27	93	7.8E-01
Eukaryotic Translation Termination	Up	1.6E-27	92	7.8E-01
Selenocysteine synthesis	Up	2.5E-27	92	7.8E-01
Response of EIF2AK4 (GCN2) to amino acid deficiency	Up	3.7E-25	100	7.3E-01
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	Up	9.8E-25	94	7.5E-01
L13a-mediated translational silencing of Ceruloplasmin expression	Up	4.8E-24	110	6.9E-01
GTP hydrolysis and joining of the 60S ribosomal subunit	Up	1.3E-23	111	6.8E-01
Eukaryotic Translation Initiation	Up	2.1E-22	118	6.5E-01
Cap-dependent Translation Initiation	Up	2.1E-22	118	6.5E-01
Regulation of expression of SLITs and ROBOs	Up	4.1E-22	167	5.3E-01
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	Up	3.0E-21	126	4.3E-01
Selenoamino acid metabolism	Up	3.6E-20	116	6.2E-01
Respiratory electron transport	Up	1.0E-18	103	4.2E-01
Cellular response to starvation	Up	3.0E-18	154	5.4E-01
Nonsense-Mediated Decay (NMD)	Up	4.5E-18	113	6.2E-01
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	Up	4.5E-18	113	6.2E-01
Formation of the ternary complex, and subsequently, the 43S complex	Up	6.2E-16	52	7.3E-01
Translation	Up	7.9E-16	291	3.8E-01
Influenza Viral RNA Transcription and Replication	Up	2.9E-14	135	5.3E-01
Complex I biogenesis	Up	6.8E-14	57	4.3E-01
Signaling by ROBO receptors	Up	2.7E-13	211	4.0E-01
Mitochondrial translation initiation	Up	3.0E-13	88	3.2E-01
Vif-mediated degradation of APOBEC3G	Up	3.0E-13	51	3.8E-01

The citric acid (TCA) cycle and respiratory electron transport	Up	3.3E-13	177	3.2E-01
Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S	Up	1.0E-12	60	6.3E-01
Ribosomal scanning and start codon recognition	Up	1.2E-12	59	6.3E-01
Translation initiation complex formation	Up	1.4E-12	59	6.3E-01
Mitochondrial translation termination	Up	2.1E-12	88	3.1E-01
Mitochondrial translation	Up	2.2E-12	94	3.0E-01
Mitochondrial translation elongation	Up	3.0E-12	88	3.1E-01
Regulation of ornithine decarboxylase (ODC)	Up	9.4E-12	51	3.4E-01
Negative regulation of NOTCH4 signaling	Up	1.1E-11	51	3.5E-01
FBXL7 down-regulates AURKA during mitotic entry and in early mitosis	Up	2.2E-11	52	3.4E-01
Regulation of activated PAK-2p34 by proteasome mediated degradation	Up	2.7E-11	47	3.5E-01
Cross-presentation of soluble exogenous antigens (endosomes)	Up	3.1E-11	50	3.7E-01
Ubiquitin-dependent degradation of Cyclin D	Up	4.3E-11	49	3.4E-01
Vpu mediated degradation of CD4	Up	6.7E-11	49	3.3E-01
SCF-beta-TrCP mediated degradation of Emi1	Up	7.4E-11	52	3.3E-01
Influenza Infection	Up	7.8E-11	156	4.6E-01
Ubiquitin Mediated Degradation of Phosphorylated Cdc25A	Up	1.1E-10	50	3.2E-01
p53-Independent DNA Damage Response	Up	1.1E-10	50	3.2E-01
p53-Independent G1/S DNA damage checkpoint	Up	1.1E-10	50	3.2E-01
Formation of ATP by chemiosmotic coupling	Up	1.4E-10	17	6.4E-01
SCF(Skp2)-mediated degradation of p27/p21	Up	2.0E-10	57	3.2E-01
Autodegradation of Cdh1 by Cdh1:APC/C	Up	2.0E-10	61	3.1E-01
Regulation of Apoptosis	Up	2.5E-10	50	3.2E-01
Metabolism of polyamines	Up	3.3E-10	59	3.1E-01

3.2.5 Performance of published signatures on the paediatric cohort

Several transcriptomic biomarkers have been published as potential diagnostic tests for TB. These signatures have several limitations, one of which is their almost exclusive discovery from adult patient cohorts. A total of 27 published transcriptomic biomarker signatures were evaluated in this study using an independent paediatric cohort from South Africa.

Selection of previously published signatures

Signatures were selected from the literature based on the 5 specific requirements listed below (**Figure 3.8**):

1. They were derived with the intention of discovering a transcriptomic biomarker signature for TB diagnosis
2. They were derived from whole blood samples
3. Training data was available for signatures which required reconstructing of their models
4. The original publication must have independently tested their signature
5. Signatures which compared TB vs OD and TB vs LTBI were selected above TB vs OD and LTBI together if these comparisons were available.

This selection process resulted in 27 signatures being used in this study (**Table 1.3**). Focussing on the cohort country of origin, 11/27 of signatures were derived from South African cohorts which was the most common cohort used. Following South Africa was the UK (8/27, 30%); Malawi (7/27, 26%); India and the USA (3/27, 11%); The Gambia (2/27, 7%); and Ethiopia, France, and Venezuela (1/27, 4%). The majority of signatures were derived from adult cohorts (21/27); then paediatric cohorts (4/27); followed by adolescent cohorts (2/27). The selected publications also used different control groups when comparing gene expression against active TB patients. These control groups include other diseases (5/27); LTBI (5/27); healthy controls or healthy household contacts (3/27); other diseases and healthy controls together (3/27); LTBI and healthy controls together (4/27); and other diseases and LTBI together (3/27). The signatures were also derived from different transcriptomic platforms, with the majority using microarrays (19/27); then RNA-seq (6/27); then RT-PCR (2/27); dc-RT MPLA (1/27); and 1/27 studies using a multicohort approach. The genes from each signature were then filtered further, where necessary (**Figure 3.8, Chapter 2.2**).

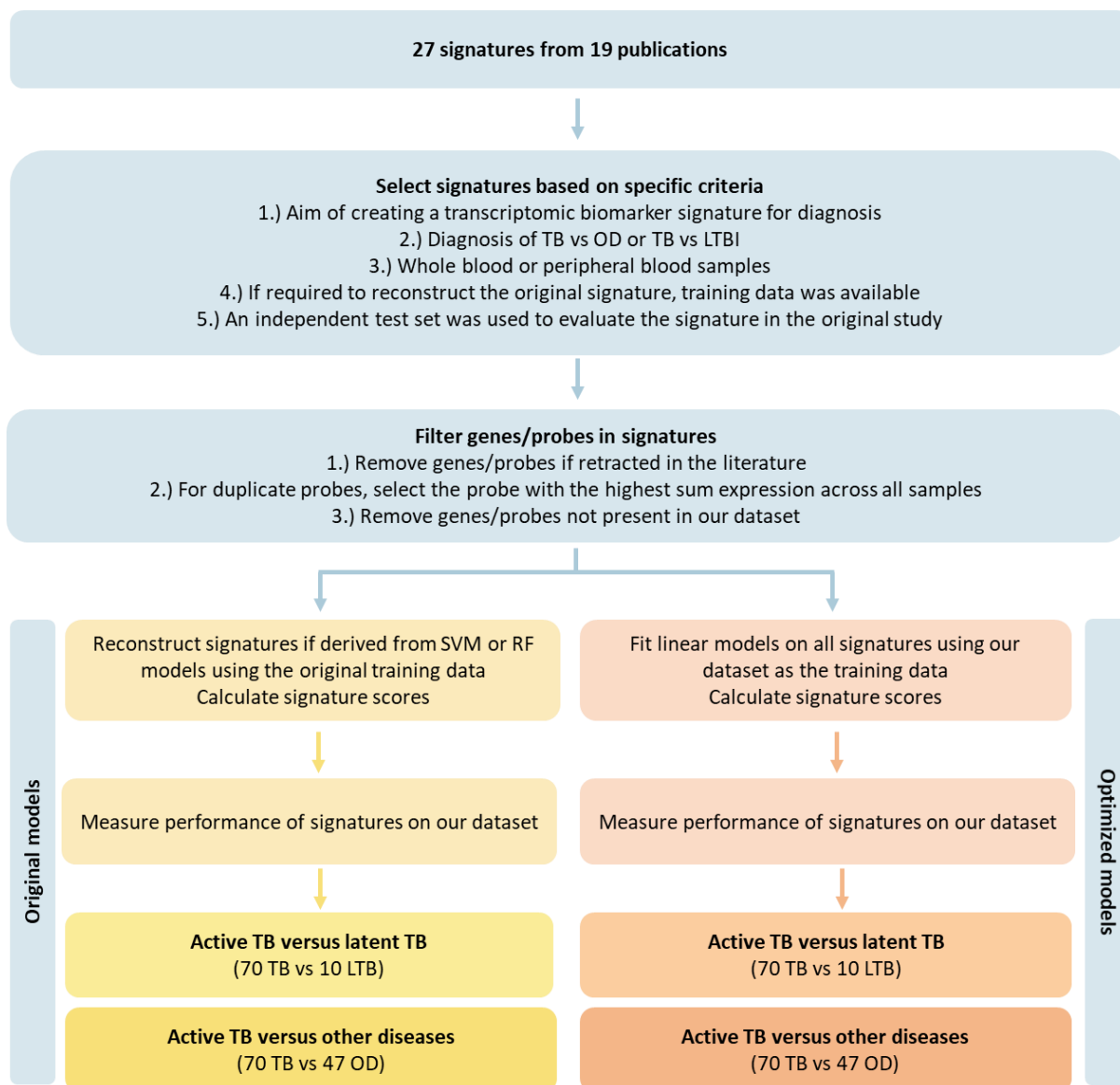


Figure 3.8: Literature and bioinformatic workflow of the validation of published biomarker signatures for TB on an independent paediatric cohort. 27 signatures were identified from 19 publications based on search criteria. The signature genes were filtered for further analysis. Downstream analysis included two branches: the first using the original models used to derive the signatures and the second using optimised models of the signatures through logistic regression analysis. Both branches of analysis were validated by their performance in discriminating TB from latent TB infection and TB from other disease patients. TB = tuberculosis; LTBI = latent tuberculosis infection; OD = other diseases; SVM = support vector machine; RF = random forest.

Signature performance on the paediatric dataset

The accuracy (measured by area under the ROC curve), sensitivity, and specificity was measured for each of the 27 signatures first using their original models to identify if these signatures met the WHO guidelines for a novel non-sputum based diagnostic test for TB in this paediatric dataset. Two comparisons were evaluated based on the disease status of the paediatric cohort: active TB vs LTBI and active TB vs OD.

When comparing active TB vs LTBI, two signatures achieved an AUC of above 90% (Araujo1 and Kaforou33, **Table 3.8**). The top 5 performing signatures based on AUC were Araujo1, Kaforou33, Kaforou24, Gliddon3, and Anderson44 (**Figure 3.9**). Only 4 signatures met the WHO guidelines for sensitivity or specificity (**Figure 3.9B**, **Table 3.8**). These included Araujo1 (meeting the optimal sensitivity and specificity), Kaforou33 (meeting the optimal specificity when stratifying for the minimum sensitivity, and vice versa), Gliddon3 (meeting the minimum sensitivity and specificity requirements), and Anderson44 (meeting the minimum sensitivity and specificity requirements).

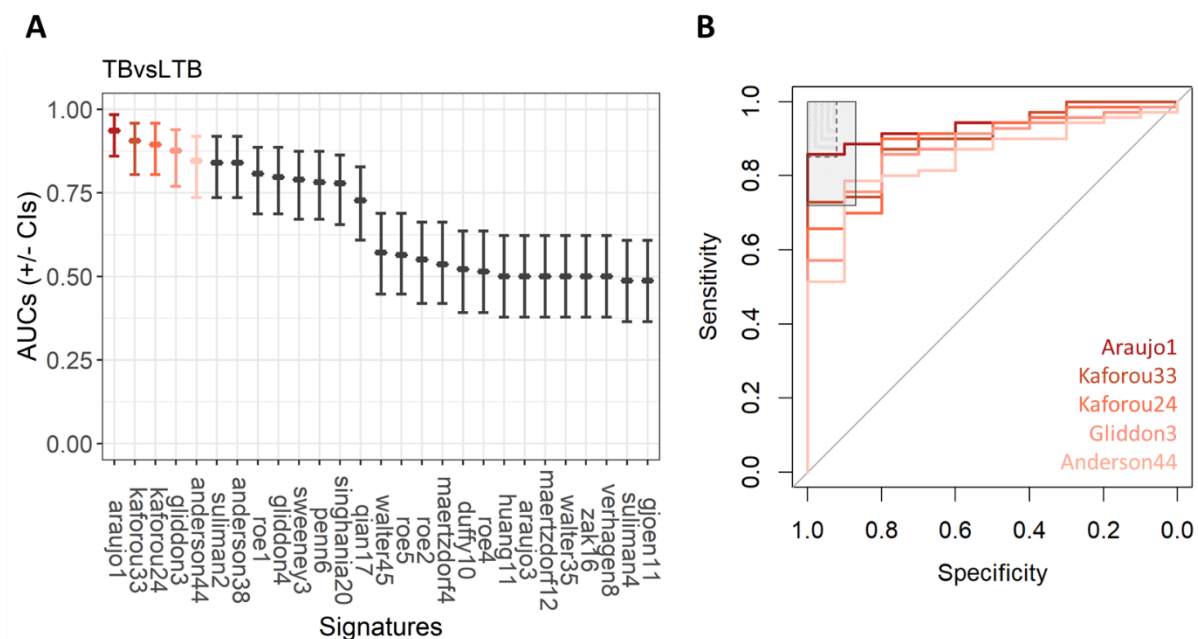


Figure 3.9: Performance of published host transcriptomic signatures using their original classification models to diagnose TB from LTBI. A.) AUC measurements of published signatures in decreasing order with the confidence intervals (CI) indicated by bars. **B.)** The ROC curves of the top 5 signatures, as measured by AUC. The solid line box indicates the minimum WHO guideline values for sensitivity and specificity and the dotted line box indicates the maximum WHO guideline values for sensitivity and specificity. The top 5 signatures as measured by AUC are coloured throughout.

Table 3.8: The performance of published biomarker signatures using their original models in discriminating between TB and LTBI patients. Brackets under the sensitivity and specificity measurements indicate the sensitivity or specificity measurements when stratifying the cohort to meet WHO minimum guidelines or WHO optimal guidelines. Bold indicates those signatures which meet the maximum WHO guidelines.

Signature name	AUC (95% CI)	Specificity (sensitivity at 72%)	Specificity (sensitivity at 85%)	Sensitivity (specificity at 87%)	Sensitivity (specificity at 92%)
Araujo1	0.936 (0.86-0.984)	1	1	0.89	0.86
Kaforou33	0.906 (0.805-0.959)	1	0.8	0.74*	0.73*
Kaforou24	0.894 (0.805-0.959)	0.8	0.8	0.70	0.66
Gliddon3	0.876 (0.77-0.939)	0.9*	0.8	0.76*	0.57
Anderson44	0.846 (0.736-0.919)	0.9*	0.6	0.79*	0.51
Suliman2	0.84 (0.736-0.919)	0.8	0.7	0.47	0.47
Anderson38	0.84 (0.736-0.919)	0.8	0.6	0.57	0.50
Roe1	0.807 (0.687-0.886)	0.8	0.5	0.54	0.37
Gliddon4	0.797 (0.687-0.886)	0.7	0.6	0.66	0.17
Sweeney3	0.79 (0.671-0.875)	0.6	0.6	0.54	0.44
Penn6	0.781 (0.671-0.875)	0.7	0.4	0.57	0.39
Singhania20	0.779 (0.656-0.863)	0.5	0.3	0.61	0.60
Qian17	0.727 (0.609-0.828)	0.6	0.6	0.31	0.19
Walter45	0.571 (0.447-0.689)	0.2	0.2	0.00	0.00
Roe5	0.564 (0.447-0.689)	0	0	0.23	0.00
Roe2	0.55 (0.419-0.663)	0	0	0.00	0.00
Maertzdorf4	0.536 (0.419-0.663)	0	0	0.17	0.00
Duffy10	0.521 (0.392-0.636)	0	0	0.04	0.04
Roe4	0.514 (0.392-0.636)	0	0	0.00	0.00
Huang11	0.5 (0.378-0.622)	0	0	0.00	0.00
Araujo3	0.5 (0.378-0.622)	0	0	0.00	0.00
Maertzdorf12	0.5 (0.378-0.622)	0	0	0.00	0.00
Walter35	0.5 (0.378-0.622)	0	0	0.00	0.00
Zak16	0.5 (0.378-0.622)	0	0	0.00	0.00
Verhagen8	0.5 (0.378-0.622)	0	0	0.00	0.00
Suliman4	0.487 (0.364-0.608)	0.1	0	0.21	0.10
Gjoen11	0.487 (0.364-0.608)	0.1	0	0.23	0.23

* Meets the WHO minimum guidelines

When comparing active TB with OD, no signatures achieved an AUC of above 90% (**Table 3.9**). The top 5 performing signatures based on AUC were Sweeney3, Kaforou33, Anderson38, Gjoen11, and Roe1 (**Figure 3.10**). None of the signatures met the optimal or minimum WHO guidelines.

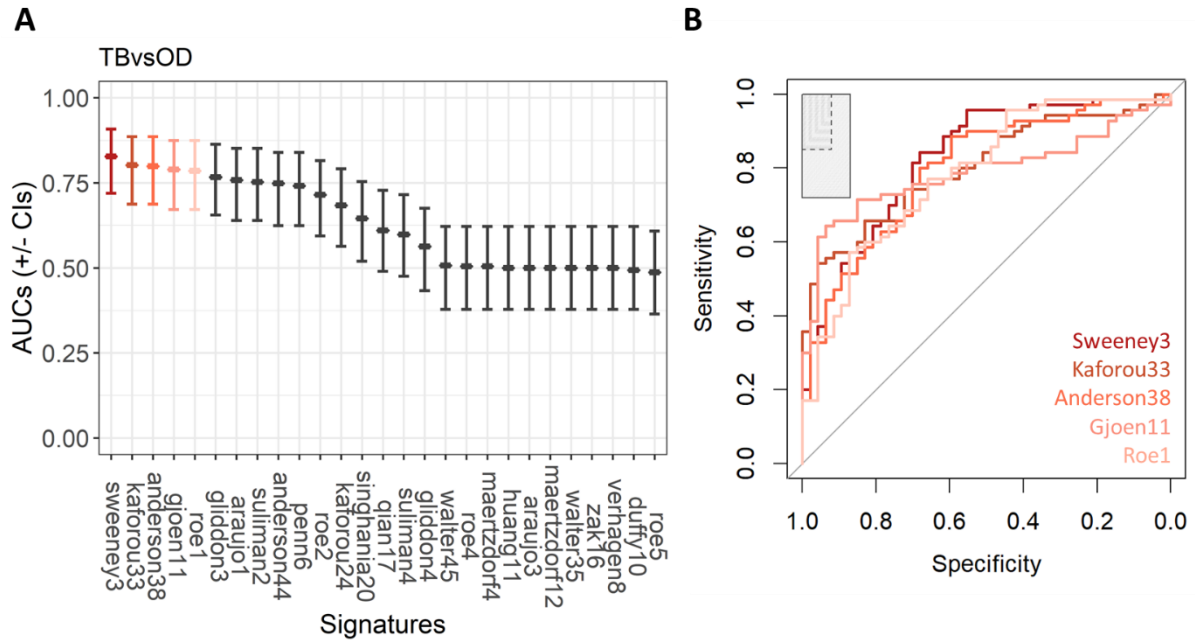


Figure 3.10: Performance of published host transcriptomic signatures using their original classification models to diagnose TB from OD. A.) AUC measurements of published signatures in decreasing order with the confidence intervals (CI) indicated by bars. **B.)** The ROC curves of the top 5 signatures, as measured by AUC. The solid line box indicates the minimum WHO guideline values for sensitivity and specificity and the dotted line box indicates the maximum WHO guideline values for sensitivity and specificity. The top 5 signatures as measured by AUC are coloured throughout.

Table 3.9: The performance of published biomarker signatures using their original models in discriminating between TB and OD patients. Brackets under the sensitivity and specificity measurements indicate the sensitivity or specificity measurements when stratifying the cohort to meet the WHO minimum guidelines or WHO optimal guidelines.

Signature name	AUC (95% CI)	Specificity (sensitivity at 72%)	Specificity (sensitivity at 85%)	Sensitivity (specificity at 87%)	Sensitivity (specificity at 92%)
Sweeney3	0.827 (0.72-0.908)	0.74	0.62	0.57	0.44
Kaforou33	0.802 (0.687-0.886)	0.72	0.49	0.57	0.56
Anderson38	0.798 (0.687-0.886)	0.70	0.60	0.51	0.44
Gjoen11	0.789 (0.671-0.875)	0.79	0.26	0.66	0.64
Roe1	0.785 (0.671-0.875)	0.66	0.49	0.57	0.34
Gliddon3	0.767 (0.656-0.863)	0.62	0.43	0.60	0.44
Araujo1	0.758 (0.64-0.852)	0.68	0.40	0.56	0.51
Suliman2	0.753 (0.64-0.852)	0.68	0.53	0.39	0.26
Anderson44	0.749 (0.624-0.84)	0.62	0.36	0.54	0.37
Penn6	0.741 (0.624-0.84)	0.57	0.47	0.46	0.26
Roe2	0.715 (0.594-0.816)	0.00	0.00	0.56	0.00
Kaforou24	0.683 (0.564-0.791)	0.57	0.34	0.36	0.20
Singhania20	0.646 (0.519-0.754)	0.45	0.30	0.30	0.21
Suliman4	0.598 (0.476-0.715)	0.47	0.26	0.11	0.09
Qian17	0.595 (0.476-0.715)	0.40	0.19	0.21	0.20
Gliddon4	0.563 (0.433-0.676)	0.28	0.15	0.21	0.16
Walter45	0.507 (0.378-0.622)	0.09	0.09	0.00	0.00
Roe4	0.504 (0.378-0.622)	0.00	0.00	0.00	0.00
Maertzdorf4	0.504 (0.378-0.622)	0.00	0.00	0.00	0.00
Huang11	0.5 (0.378-0.622)	0.00	0.00	0.00	0.00
Araujo3	0.5 (0.378-0.622)	0.00	0.00	0.00	0.00
Maertzdorf12	0.5 (0.378-0.622)	0.00	0.00	0.00	0.00
Walter35	0.5 (0.378-0.622)	0.00	0.00	0.00	0.00
Zak16	0.5 (0.378-0.622)	0.00	0.00	0.00	0.00
Verhagen8	0.5 (0.378-0.622)	0.00	0.00	0.00	0.00
Duffy10	0.493 (0.378-0.622)	0.00	0.00	0.03	0.03
Roe5	0.487 (0.364-0.608)	0.00	0.00	0.00	0.00

Optimised signature performance on the paediatric dataset – active TB vs LTBI

As the signatures performed poorly on the paediatric dataset using their original models, I sought to improve their performance to their highest potential by optimising their models. Optimisation was performed through fitting logistic regression models to each signature using the paediatric dataset as the training data. Gene counts were used for the model construction for each gene/probe in the original signatures (unless unavailable). The performance of these optimised models was evaluated using the paediatric dataset as the testing data.

The performance of the optimised models in discriminating active TB from LTBI was significantly improved from their original models (**Figure 3.11, Table 3.10**). Almost all signatures performed with higher accuracy using the optimised models; the exceptions being Araujo1, Gliddon3, Roe1, and Suliman2. The top 5 performing signatures based on AUC differed from the original models, with Walter45, Anderson38, Anderson44, Walter35, and Kaforou33 being in the top 5 optimised models (**Figure 3.11B**).

Significantly more signatures achieved the WHO guidelines when stratifying the sensitivity and specificity using the optimised models. Seventeen signatures met the WHO guidelines for minimum specificity (>87%), and fifteen signatures met the WHO guidelines for optimal specificity (>92%), after stratifying the sensitivity to the minimum guideline value of >72% (**Figure 3.11C, Table 3.10**). Eleven signatures met the minimum guidelines for specificity, and eight signatures met the optimal guidelines for specificity, after stratifying the sensitivity to the optimal guideline value of >85%.

Similarly, seventeen signatures met the guidelines for minimum sensitivity (>72%), and eleven signatures met the optimal sensitivity (>85%), after stratifying the specificity to the minimum guideline value of >87% (**Table 3.10**). Fifteen signatures met the minimum guidelines for sensitivity, and eight signatures met the optimal sensitivity, after stratifying the specificity to the optimal guideline value of >92%.

Table 3.10: The performance of published biomarker signatures using optimised classification models in discriminating between TB and LTBI patients. Brackets under the sensitivity and specificity measurements indicate the sensitivity or specificity measurements when stratifying the cohort to meet the WHO minimum guidelines or WHO optimal guidelines. Bold indicates those signatures which meet the maximum WHO guidelines.

Signature name	AUC (95% CI)	Specificity (sensitivity at 72%)	Specificity (sensitivity at 85%)	Sensitivity (specificity at 87%)	Sensitivity (specificity at 92%)
Walter45	1 (0.949-1)	1.00	1.00	1.00	1.00
Anderson38	1 (0.949-1)	1.00	1.00	1.00	1.00
Anderson44	1 (0.949-1)	1.00	1.00	1.00	1.00
Walter35	0.996 (0.949-1)	1.00	1.00	0.99	0.97
Kaforou33	0.991 (0.923-1)	1.00	1.00	1.00	0.91
Qian17	0.986 (0.923-1)	1.00	1.00	0.96	0.96
Kaforou24	0.977 (0.901-0.997)	1.00	1.00	0.94	0.90
Huang11	0.963 (0.88-0.991)	1.00	0.90*	0.94	0.80*
Penn6	0.96 (0.88-0.991)	1.00	0.80	0.83*	0.83*
Singhania20	0.941 (0.86-0.984)	1.00	0.90*	0.86	0.83*
Araujo1	0.936 (0.86-0.984)	1.00	1.00	0.89	0.86
Zak16	0.929 (0.841-0.976)	1.00	0.90	0.86	0.81*
Gjoen11	0.924 (0.841-0.976)	1.00	0.80	0.84*	0.73*
Duffy10	0.917 (0.823-0.968)	1.00	0.80	0.84*	0.79*
Sweeney3	0.9 (0.805-0.959)	1.00	0.70	0.77*	0.76*
Roe5	0.891 (0.787-0.949)	0.80	0.80	0.70	0.64
Araujo3	0.859 (0.753-0.929)	0.90*	0.60	0.80*	0.46
Roe4	0.847 (0.736-0.919)	0.90*	0.70	0.77*	0.47
Gliddon3	0.844 (0.736-0.919)	0.80	0.60	0.69	0.53
Maertzdorf4	0.83 (0.72-0.908)	0.70	0.60	0.66	0.37
Maertzdorf12	0.816 (0.703-0.897)	0.70	0.60	0.61	0.37
Gliddon4	0.811 (0.703-0.897)	0.70	0.50	0.54	0.51
Suliman4	0.811 (0.703-0.897)	0.70	0.40	0.61	0.61
Roe1	0.807 (0.687-0.886)	0.80	0.50	0.54	0.37
Roe2	0.793 (0.687-0.886)	0.60	0.50	0.57	0.36
Suliman2	0.773 (0.656-0.863)	0.70	0.40	0.54	0.54
Verhagen8	0.5 (0.378-0.622)	0	0	0	0

* Meets the WHO minimum guidelines

It was hypothesised that signatures with greater number of genes were more likely to achieve the WHO guidelines for sensitivity and specificity. To test this, the relationship between signature size and AUC, sensitivity, or specificity was visualised (**Figure 3.12**). A positive trend of increasing accuracy (AUC) for signatures with a larger size (in number of genes) was observed. This plateaued at around 20 – 30 genes. Despite this positive correlation, some signatures which had lower numbers of genes, outperformed signatures with a greater number of genes. Araujo1 outperformed the most signatures with a greater number of genes, including Roe2, Suliman2, Sweeney3, Araujo3, Gliddon3, Roe4, Suliman4, Gliddon4, Maertz4, Roe5, Verhagen8, Duffy10, Gjoen11, Maertz12, and Zak16. Of the signatures with less than 20 genes, Qian17, Huang11, Penn6, and Araujo1 had the highest AUCs. An identical trend was observed when comparing signature size with sensitivity (after stratifying for minimum or optimal specificity of >85% and >92%, respectively) or specificity (after stratifying for a minimum or optimal sensitivity of >72% and >87%, respectively). Signatures with the greatest number of genes had the highest sensitivity and specificity.

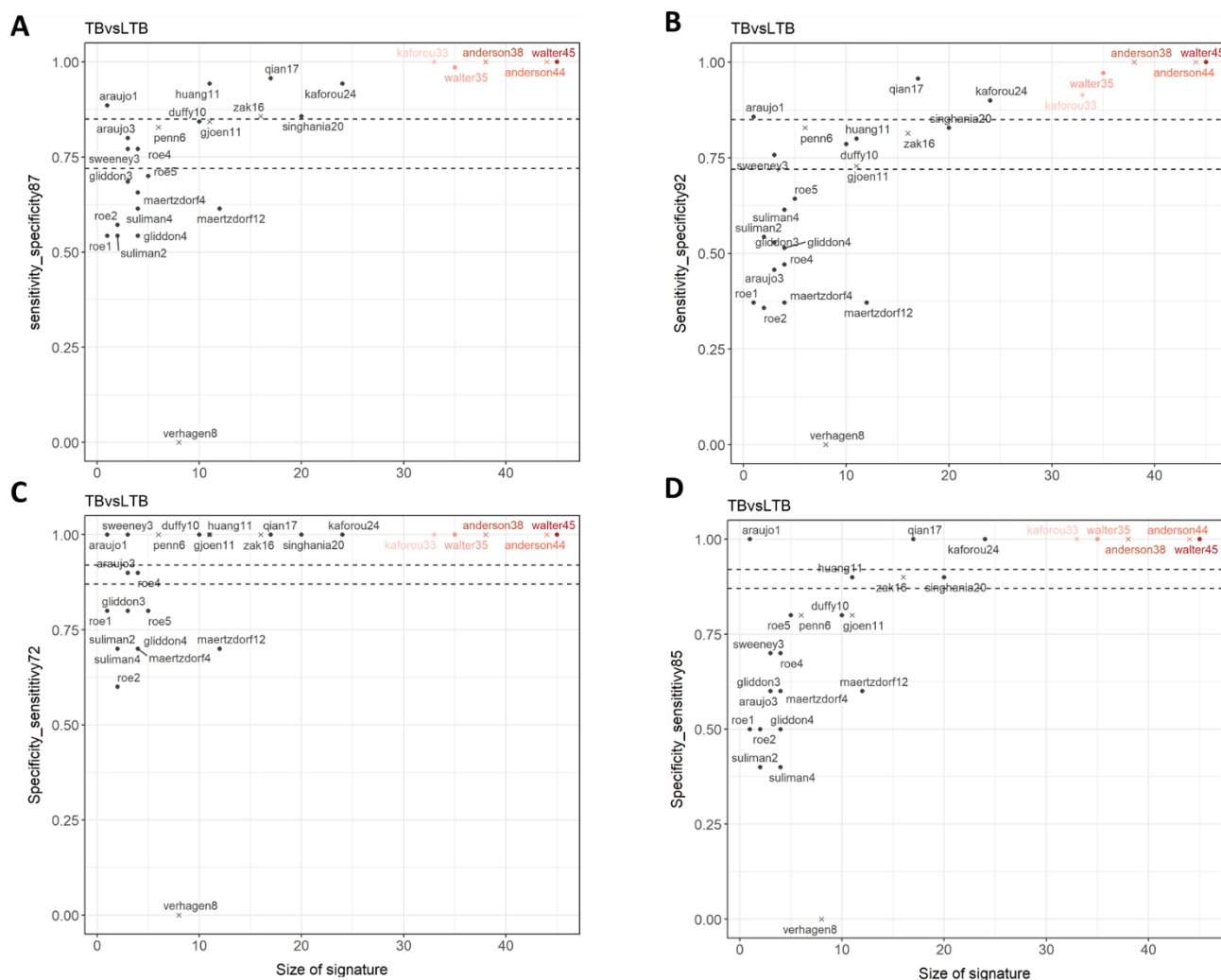


Figure 3.12: Relationship between signature size and sensitivity and specificity after stratification towards the minimum or maximum WHO guidelines to diagnose TB from LTB. A.) Visualisation of signature sensitivity, after stratifying specificity to 87%, with increasing signature size. The top dotted line indicates the cut-off for the maximum WHO guidelines for the sensitivity and the bottom dotted line indicates the cut-off for the minimum WHO guidelines for the sensitivity. **B.)** Visualisation of signature sensitivity, after stratifying specificity to 92%, with increasing signature size. The top dotted line indicates the cut-off for the maximum WHO guidelines for the sensitivity and the bottom dotted line indicates the cut-off for the minimum WHO guidelines for the sensitivity. **C.)** Visualisation of signature specificity, after stratifying sensitivity to 72%, with increasing signature size. The top dotted line indicates the cut-off for the maximum WHO guidelines for the specificity and the bottom dotted line indicates the cut-off for the minimum WHO guidelines for the sensitivity. **D.)** Visualisation of signature specificity, after stratifying sensitivity to 85%, with increasing signature size. The top dotted line indicates the cut-off for the maximum WHO guidelines for the specificity and the bottom dotted line indicates the cut-off for the minimum WHO guidelines for the specificity. Signatures derived from adult cohorts are indicated with a solid dot while signatures derived from adolescent or paediatric cohorts are indicated with a cross. The top 5 signatures as measured by AUC are coloured throughout.

Optimised signature performance on the paediatric dataset – active TB vs OD

Consistent with the previous comparison, the performance of optimised models in discriminating active TB from OD was significantly improved from the models of the original studies (**Figure 3.13**, **Table 3.11**). Optimised models had a higher AUC than original models, with the exception of Araujo1, Gliddon3, Roe1, Suliman2, and Sweeney3. The top performing signatures based on the AUC differed from the previous comparison, with Anderson38, Anderson44, Kaforou33, Walter45, and Walter35 performing with the highest accuracy (**Figure 3.13B**).

In contrast to the previous comparison, fewer signatures met the WHO guidelines for sensitivity and specificity. Fourteen signatures met the WHO guidelines for a minimum specificity (>87%), and Five signatures met the optimal specificity (>92%), after stratifying for a minimum sensitivity of >72% (**Table 3.11**). Six signatures met the minimum guidelines for specificity, and three signatures met the optimum guidelines for specificity, when stratifying for an optimal sensitivity of >85%.

Eight signatures met the guidelines for a minimum sensitivity (>72%), and three signatures met the optimal guidelines for sensitivity (>85%), after stratifying for a minimal specificity of >87%. Five signatures met the minimum guidelines for a minimum sensitivity, and two met the optimal guidelines for sensitivity, after stratifying for an optimal sensitivity of >92%.

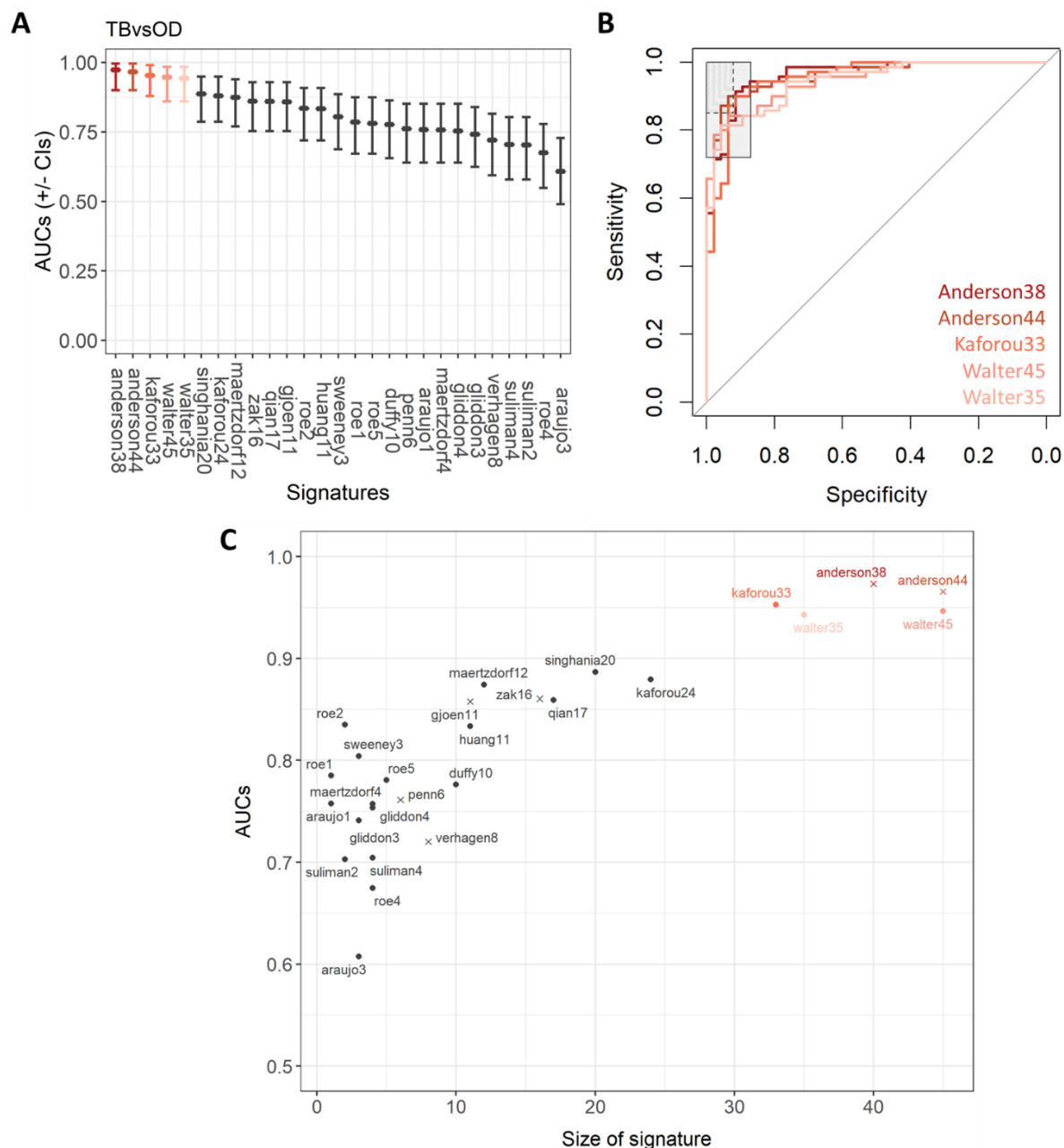


Figure 3.13: Performance of published host transcriptomic signatures using optimised classification models to diagnose TB from OD. A.) AUC measurements of published signatures in decreasing order with the confidence intervals (CI) indicated by bars. **B.)** The ROC curves of the top 5 signatures, as measured by AUC. The solid line box indicates the minimum WHO guideline values for sensitivity and specificity and the dotted line box indicates the maximum WHO guideline values for sensitivity and specificity. **C.)** Visualisation of the effect of signature size on AUC. Signatures derived from adult cohorts are indicated with a solid dot while signatures derived from adolescent or paediatric cohorts are indicated with a cross. The top 5 signatures as measured by AUC are coloured throughout.

Table 3.11: The performance of published biomarker signatures using optimised classification models in discriminating between TB and OD patients. Brackets under the sensitivity and specificity measurements indicate the sensitivity or specificity measurements when stratifying the cohort to meet the WHO minimum guidelines or WHO optimal guidelines. Bold indicates those signatures which meet the maximum WHO guidelines.

Signature name	AUC (95% CI)	Specificity (sensitivity at 72%)	Specificity (sensitivity at 85%)	Sensitivity (specificity at 87%)	Sensitivity (specificity at 92%)
Anderson38	0.962 (0.88-0.991)	0.96	0.91	0.94	0.83*
Anderson44	0.961 (0.88-0.991)	0.98	0.96	0.91	0.90
Kaforou33	0.953 (0.88-0.991)	0.94	0.94	0.93	0.87
Walter45	0.947 (0.86-0.984)	0.98	0.85*	0.84*	0.84*
Walter35	0.943 (0.86-0.984)	0.98	0.83*	0.84*	0.81*
Singhania20	0.887 (0.787-0.949)	0.87*	0.74*	0.73*	0.47
Kaforou24	0.879 (0.787-0.949)	0.87*	0.70	0.77*	0.70
Maertzdorf12	0.874 (0.77-0.939)	0.83*	0.70	0.69	0.60
Zak16	0.86 (0.753-0.929)	0.87*	0.62	0.73*	0.69
Qian17	0.859 (0.753-0.929)	0.83*	0.60	0.70	0.56
Gjoen11	0.858 (0.753-0.929)	0.81*	0.66	0.69	0.59
Roe2	0.835 (0.72-0.908)	0.72*	0.68	0.53	0.44
Huang11	0.833 (0.72-0.908)	0.77*	0.64	0.57	0.53
Sweeney3	0.804 (0.687-0.886)	0.72*	0.57	0.44	0.30
Roe1	0.785 (0.671-0.875)	0.66	0.49	0.57	0.34
Roe5	0.781 (0.671-0.875)	0.64	0.45	0.57	0.40
Duffy10	0.776 (0.656-0.863)	0.66	0.47	0.53	0.41
Penn6	0.761 (0.64-0.852)	0.68	0.45	0.47	0.29
Araujo1	0.758 (0.64-0.852)	0.68	0.40	0.56	0.51
Maertzdorf4	0.757 (0.64-0.852)	0.62	0.47	0.47	0.27
Gliddon4	0.753 (0.64-0.852)	0.64	0.51	0.47	0.26
Gliddon3	0.741 (0.624-0.84)	0.64	0.32	0.53	0.46
Verhagen8	0.72 (0.594-0.816)	0.60	0.49	0.31	0.21
Suliman4	0.705 (0.579-0.804)	0.57	0.45	0.33	0.27
Suliman2	0.703 (0.579-0.804)	0.53	0.40	0.33	0.27
Roe4	0.675 (0.549-0.779)	0.60	0.34	0.27	0.24
Araujo3	0.608 (0.49-0.728)	0.43	0.23	0.23	0.17

* Meets the WHO minimum guidelines

The relationship between signature size and AUC, sensitivity, or specificity was visualised again. In agreement with the previous comparison, there was a positive trend towards higher AUC with increasing number of genes in the signature, which plateaued at around 20 – 30 genes (**Figure 3.14**). However, there were a greater number of signatures with fewer genes which outperformed signatures with more genes. In addition, all signatures with fewer than 12 genes had a lower AUC than larger signatures. An analogous trend was observed when comparing signature size with sensitivity and specificity after stratifying by the optimal and minimal specificity and sensitivity values, respectively.

The Anderson, Kaforou, and Walter signature pairs were in the top 10 signatures based on AUC in both the active TB vs LTBI comparison and the active TB vs OD comparisons. The Anderson signatures and the Walkter45 signature were in the top 5 signatures based on AUC in both comparisons.

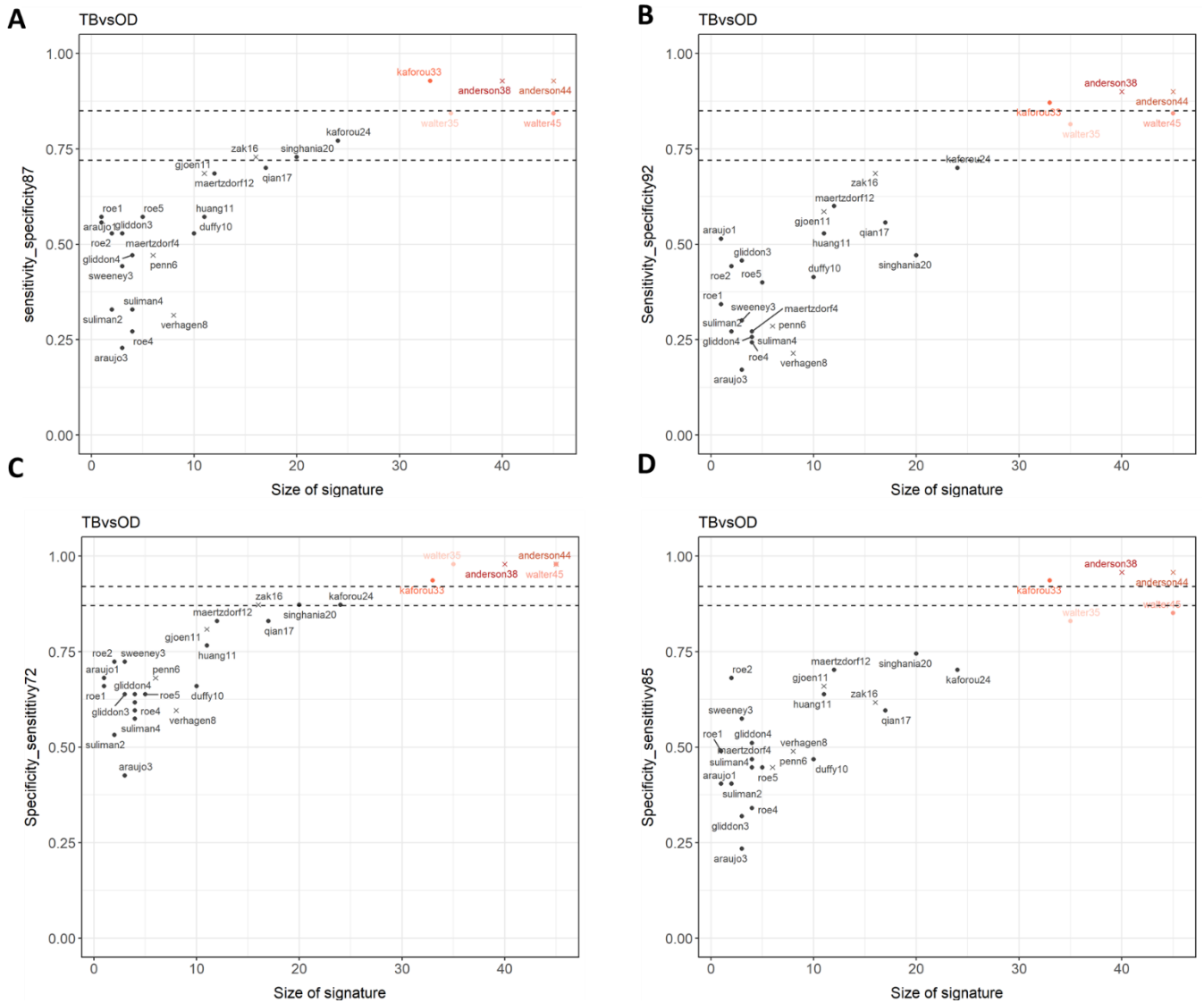
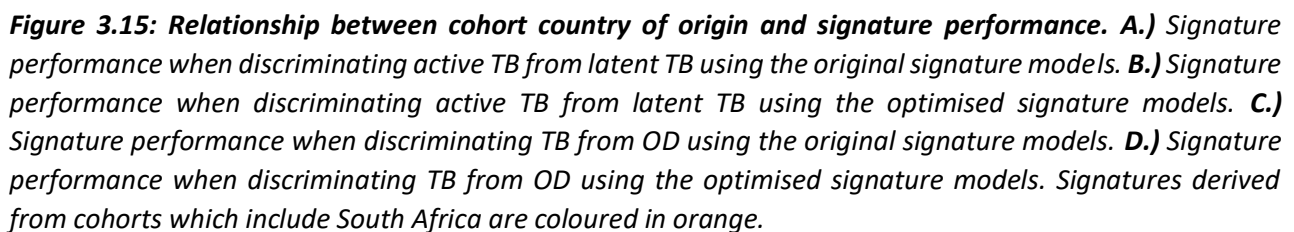


Figure 3.14: Relationship between signature size and sensitivity and specificity after stratification towards the minimum or maximum WHO guidelines to diagnose TB from OD. **A.)** Visualisation of signature sensitivity, after stratifying specificity to 87%, with increasing signature size. The top dotted line indicates the cut-off for the maximum WHO guidelines for the sensitivity and the bottom dotted line indicates the cut-off for the minimum WHO guidelines for the sensitivity. **B.)** Visualisation of signature sensitivity, after stratifying specificity to 92%, with increasing signature size. The top dotted line indicates the cut-off for the maximum WHO guidelines for the sensitivity and the bottom dotted line indicates the cut-off for the minimum WHO guidelines for the sensitivity. **C.)** Visualisation of signature specificity, after stratifying sensitivity to 72%, with increasing signature size. The top dotted line indicates the cut-off for the maximum WHO guidelines for the specificity and the bottom dotted line indicates the cut-off for the minimum WHO guidelines for the sensitivity. **D.)** Visualisation of signature specificity, after stratifying sensitivity to 85%, with increasing signature size. The top dotted line indicates the cut-off for the maximum WHO guidelines for the specificity and the bottom dotted line indicates the cut-off for the minimum WHO guidelines for the specificity. Signatures derived from adult cohorts are indicated with a solid dot while signatures derived from adolescent or paediatric cohorts are indicated with a cross. The top 5 signatures as measured by AUC are coloured throughout.

Finally, it was hypothesised that signatures derived from a cohort with the same country of origin as the validation dataset used here would be more likely to perform better than signatures derived from cohorts with different origins. This relationship between cohort country and signature performance was visualised in **Figure 3.15**. Signatures derived from South African cohorts did generally perform better when distinguishing active TB from LTBI and when distinguishing TB from OD using their original models (**Figure 3.15A, 3.15C**). However, this trend was not observed in either comparison when the optimised models of the signatures were used (**Figure 3.15B, 3.15D**).



3.3 Discussion

Paediatric TB is significantly understudied, particularly within the field of diagnostics. In these analyses, I aimed to further understand the changes in gene expression, and the implied differences in biological pathways, observed in a paediatric cohort of TB patients compared to LTBI and OD patients. In addition, this cohort was used for a validation study of the currently published biomarker signatures for TB to determine if any signature meets the WHO guidelines for novel diagnostic tests for paediatric TB.

Differential gene expression analysis and pathway analysis were performed to characterise this cohort. The expected genes and pathways were identified as significantly differentially expressed in the TB vs OD comparison, with *GBP5* ($\log_2$ fold change 1.5) being the most significantly differentially expressed gene and is induced by IFN γ signalling (Tretina et al., 2019). IFN γ signalling was also identified as the top pathway to be overexpressed. This was not the case for the TB vs LTBI comparison, where *SNORD17* ($\log_2$ fold change 2.6) was found as the top differentially expressed gene alongside other small nucleolar RNAs. This was also reflected in the pathway analysis, which identified many translation-related processes to be overexpressed. The TB vs OD comparison identified TB-specific genes and pathways; however, the TB vs LTBI comparison will instead identify non-specific pathways involved in infection responses in general as LTBI patients show similar blood transcriptional profiles to healthy controls (Broderick et al., 2021). Nevertheless, the observed pathways altered in TB vs OD patients confirm previous findings in both adult and paediatric cohorts (Anderson, et al., 2014; Sweeney et al., 2016), however those comparing TB with LTBI have not been observed in previous studies. This comparison found a large number of *SNOR* genes being significantly differentially expressed, which encode small nucleolar RNAs, a subset of small non-coding RNAs. These genes are involved in RNA regulation, reflected in the pathway analysis which consisted of predominantly translational pathways. Only one of the previous biomarker studies identified a similar result when comparing TB with LTBI in a paediatric cohort (Silva De Araujo et al., 2019). In this study the authors specifically investigate small non-coding RNAs as potential biomarker signatures for TB vs LTBI, as they have recently been implicated in TB disease. Other studies have not identified such a transcriptional signature, however all but one of these use microarray instead of RNA-seq which will not identify small non-coding RNAs unless they are specifically targeted (Anderson, et al., 2014; de Araujo et al., 2016; Silva De Araujo et al., 2019; Verhagen, et al., 2013).

The main aims of this Chapter were to evaluate the performance of previously published host transcriptomic TB signatures using a unique paediatric cohort from South Africa. Despite the number of adult TB biomarker signatures identified, and the validation studies on adult cohorts, no validation studies have been performed on current signatures in a paediatric cohort. Therefore, it is imperative that their performance on a paediatric cohort is validated as the development of a paediatric TB biomarker signature has been neglected to date. This analysis was performed using the original classification models for each signature and also through refitting an optimised model to measure the highest possible accuracy for each signature using their original genes. Only 1 signature, Araujo1, achieved the optimal WHO guidelines for sensitivity and specificity when using the original models to distinguish active from LTBI patients. In contrast, when using the original models to distinguish TB from OD patients, no signatures met the optimal WHO guidelines. Notably, signatures which were derived using SVM or RF methods showed the lowest performance. This may be due to the use of microarray datasets which produce different expression distributions for each gene compared to RNA-seq datasets, which was used as the validation cohort in this analysis. This suggests that signatures derived from microarray datasets using SVM or RF are not generalisable and therefore cannot be easily used to test patients from different cohorts. Overall, the currently published signatures perform poorly using their original models and cannot be reliably used as a diagnostic test for TB in independent cohorts.

Instead of using the original models to evaluate the performance of the signatures, the models were optimised to allow the signatures to reach their maximum potential in this dataset. This involved fitting logistic models to each signature using the South African paediatric dataset to train the models. The performance of almost every signature improved using the optimised models. The signatures with the highest accuracy, sensitivity, and specificities using the optimised models (top 5 signatures for TB vs LTBI being Walter45, Anderson38, Anderson44, Walter35 and Kaforou33 and the top 5 signatures for TB vs OD being Anderson38, Anderson44, Kaforou33, Walter45, and Walter35) differed from the signatures with the highest performances in these measurements using the original models (top 5 signatures for TB vs LTBI being Araujo1, Kaforou33, Kaforou24, Gliddon3, and Anderson44 and the top 5 signatures for TB vs OD being Araujo1, Kaforou33, Kaforou24, Gliddon3, and Anderson44). This was, at least in part, due to signatures consisting of a larger number of genes performing generally better than those consisting of a smaller number of genes. When distinguishing between active TB and LTBI, 17 optimised signatures met the minimum WHO guidelines for sensitivity and specificity, while 14 optimised signatures met the minimum WHO guidelines for these values when distinguishing TB from OD. Several signatures met the minimum WHO guidelines in both comparisons, including the

Anderson38, Anderson44, Gjoen11, Huang11, Kaforou24, Kaforou33, Singania20, Sweeney3, Walter35, Walter45, and Zak16 signatures. The Anderson44 and Kaforou33 signatures also met the optimal WHO guidelines for sensitivity and specificity using optimised models for both comparisons. Therefore, it is possible that certain signatures can achieve the WHO guidelines when optimised towards a dataset, and certain signatures show higher potential for development into a diagnostic test which is generalisable and can be used across different cohorts.

Several signatures, however, did not achieve the WHO guidelines when using either their original or optimised models, including Roe1, Roe2, Roe5, Maertz4, Maertz12, Suliman2, Suliman4, Gliddon4, and Verhagen8 when distinguishing active TB from latent TB patients, and Roe1, Roe4, Roe5, Maertz4, Gliddon3, Gliddon4, Suliman2, Suliman4, Araujo1, Araujo3, Penn6, Verhagen8, and Duffy10 when distinguishing TB from OD patients. It is important to note that the majority of these signatures consist of a small number of genes which may impact their performance. This relationship between signature performance and signature size was visualised and indeed indicated that signatures with a greater number of genes performed better overall. Our analysis observed a plateau of the positive relationship between signature performance and gene size at around 20 – 30 genes. This analysis suggests that it would be beneficial if the development of a TB diagnostic test could facilitate this number of genes instead of a limited gene set.

It is likely that other factors influence signature performance in this analysis other than gene number. The validation cohort used here consists of RNA-seq data from paediatric patients recruited in South Africa and, therefore, signatures derived from similar datasets may have a higher chance of performing better than those from dissimilar datasets. For example, the Anderson signatures were derived from South African and Malawian paediatric cohorts, and the Kaforou33 signature were also derived from South African and Malawian adult cohorts, and all performed well (AUCs of above 0.990, sensitivities above 90%, and specificities of 100% when distinguishing active TB from LTBI and AUCs above 0.952, sensitivities above 82%, and specificities above 90% when distinguishing TB from OD). However, this is not always the case as the Qian17 signature, which was derived from a cohort recruited in the UK using healthy controls and OD as the control variable, performed well when comparing active TB with latent TB patients, with an AUC of 0.986, a sensitivity of 96%, and a specificity of 100%. Furthermore, the optimised Walter signatures also performed well when distinguishing active TB patients from latent TB patients, despite being derived from adult cohorts recruited in the USA (with AUCs above 0.995, sensitivities above 98%, and specificities of 100%). These results suggest that while deriving signatures from multiple countries may improve the performance of that signature

overall, it is not necessarily true that signatures derived from individual countries cannot perform well on specific independent datasets.

Certain genes are consistently identified in the top performing biomarker signatures from the Turner, et. al., study (Turner et al., 2020) and from the analysis of this Chapter. Two of the top 5 signatures in the Turner, et. al., study included the gene *BATF2* (Roe1 and Roe3) and two of the signatures also contained *GBP5* (Roe3 and Sweeney3), while Kaforou25 contained *GBP6*. *BATF2* has been shown to be induced in uninfected IFN γ -stimulated BMDMs and is required for the increase in expression of *NOS2* and *TNF* in these cells (S. Roy et al., 2015), and has been shown to be increased in Mtb infected BMDM compared to uninfected stimulated macrophages (S. Roy et al., 2018). Similarly, while the GBP family of proteins have been shown to be induced by IFN γ and are involved in a range of immune responses including inflammasome activation (Kim et al., 2016) and is also involved in an antiviral response (Feng et al., 2017; Krapp et al., 2016). Both these genes have yet to be fully investigated and characterised for their roles in Mtb infection. Another study also identified common genes across several other signatures which were validated for their performance of diagnosing incipient TB, defined as the asymptomatic phase of early disease prior to the development of symptoms seen in active disease (Gupta et al., 2020). Again, these included *BATF2* and *GBP5* and also *ANKRD22* and *FCGR1A* which were all found in 3 out of 8 of the top performing signatures in that study. Despite the number of signatures evaluated in this Chapter being much greater than that previously published, the same genes were identified as most common across current signatures (**Appendix, Figure S3.4**).

Despite these promising findings, this analysis is not without limitation. Independent validation of these signatures in cohorts recruited from other countries would provide valuable insights on the conclusions drawn from these results. It should be noted that the optimised models were derived to show the highest potential accuracy each signature could achieve instead of being incorporated into specific signatures for future validation. It is possible that the optimised models used here may be overfitting the signatures to this dataset, and therefore these models may not perform well on additional datasets, although this would provide a measure of robustness of these models.

In addition, this cohort comprises of a limited number of different presentations of TB disease, including extrapulmonary TB, LTBI, and HIV infected TB patients. While their numbers are small, this is the first paediatric biomarker study to use such a diverse cohort. In addition, the control group used to generate these signatures differ between studies and may affect the performance of these

signatures on this cohort. Ultimately, the results here can be used to inform the development of different diagnostic tests for TB, depending on the requirements.

This is the first analysis to independently validate published host blood transcriptomic signatures for TB using a paediatric cohort from South Africa. In addition, I have shown that the models used to generate some signatures can be improved to enhance accuracy, sensitivity, and specificity, and can meet the WHO guidelines when using optimised models. Further validation of these signatures and their optimised models, in other paediatric cohorts from other countries, would further clarify their robustness and determine which, if any, signatures warrant further development into diagnostic tests for TB.

3.4 Appendix

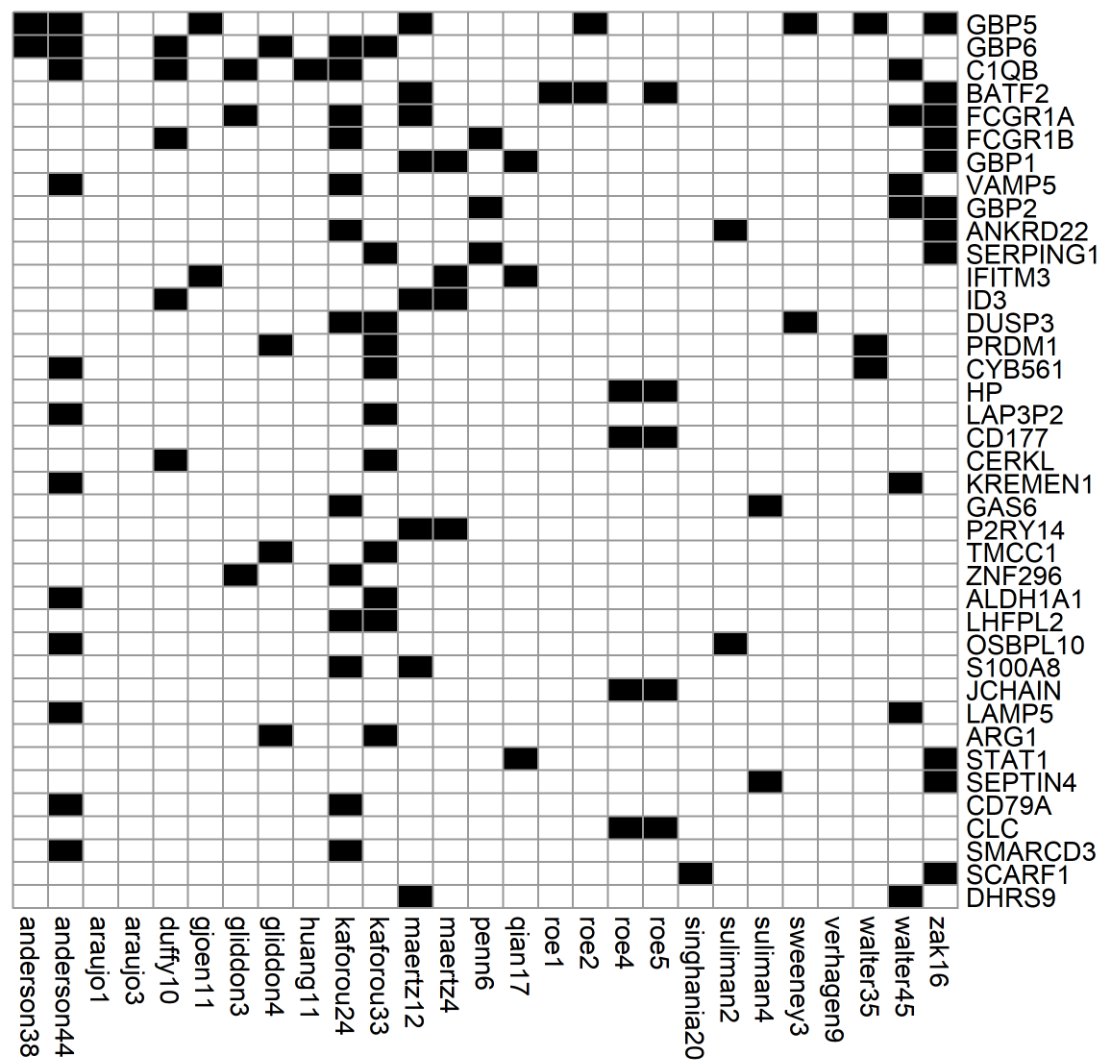


Figure S3.4: Heatmap of common biomarker signature genes. Signatures filtered for those with genes found in at least one other signature. Ordered by genes found in most signatures. Black boxes indicate gene is present in that signature and white boxes indicate gene is absent in that signature.

Chapter 4: The bacterium: Multi -omics integration of genomic and metabolomic level information in TB lineages

4.1 Background

The majority of TB research continues to focus on the use of reference Mtb strains such as H37Rv, despite the prevalence of 9 Mtb lineages across the world (Coscolla et al., 2021). These lineages represent genetically similar Mtb strains which have evolved through distinct conditions such as different geographical and social factors which affect the host (Coscolla et al., 2021). While TB research traditionally utilises the H37Rv strain which is part of lineage 4, studies have since shown significant differences between Mtb lineages, in both bacterial response to the host and the host immune response to infection (Krishnan et al., 2011). As a result, there has been a growing number of publications investigating these lineages to understand the evolutionary drivers which differentiate them from each other and their impact on host responses.

In addition to Mtb lineages, strains within each lineage can be classified by sub-lineage. Different sub-lineage nomenclatures exist, such as spoligotyping which is used in this study. Spoligotyping is a molecular classification technique which is frequently used to group Mtb strains into sub-lineage level categories. This technique amplifies the direct repeat (DR) locus found in Mtb strains through PCR (Jagielski et al., 2014). The DR locus consists of 36 bp repeats with unique 34 – 41 bp spacer sequences which are conserved across lineages. Strains are therefore identified by the absence of spacer sequences from the known total of spacer sequences. In the SITVIT2 spoligotyping database there are 9,658 spoligotype patterns which type 103,856 Mtb strains (Couvin et al., 2019).

Differences across Mtb spoligotypes may be associated with increased or decreased virulence and be associated with specific host responses (Krishnan et al., 2011). While the field of Mtb lineage research is small, even fewer studies investigate Mtb spoligotypes. The exception to this is the study of strains within a subset of Mtb lineage 2, specifically the Beijing strains (Merker et al., 2015). Significant research has been published to understand the differences between this highly virulence subset of lineage 2 strains and those less virulent, but this focus has left other Mtb strains understudied.

Furthermore, these analyses predominantly investigate the effect of genomic differences between virulent and avirulent strains through sequencing and functional validation studies, which would be challenging to incorporate into larger studies which would require high-throughput methodologies.

Since the sequencing of the Mtb genome in 1998 (Cole et al., 1998) and the advancements in next generation sequencing over the last few decades, Mtb research in the field of genomics has flourished. This field then expanded to include the genomic analysis of Mtb lineages, which identified several lineage-specific traits such as SNPs in genes suggested to be involved in Mtb virulence (Phelan et al., 2016) and genes involved in antibiotic resistance related pathways (Oppong et al., 2019). Functional interpretation of these genetic differences focussing on individual genetic differences through laboratory validation studies are time-consuming and limited in throughput. For example, genomics data such as that containing SNPs may implicate several hundred genes, all of which would require laboratory validation through separate Mtb mutants. Therefore, there is a need for validation analyses through use of other high-throughput technologies, such as -omics technologies.

The increase in accessibility of other omics technologies, in addition to genomics, enables further validation of genomic findings through high-throughput techniques. Technologies such as metabolomics or proteomics are a direct read-out of the real-time activities of a cell, or multiple cells, at a single point in time. Combining these technologies with genomics can aid in identifying which genomic variations between strains bring about downstream effects.

Genomic studies have identified that a significant proportion of the Mtb genome encodes for metabolites, or proteins and lipids involved in metabolic pathways (Cole et al., 1998), and more recently the Mtb metabolism been shown to be significantly influential in Mtb infection and virulence (Muñoz-Elías et al., 2006). This includes significant genomic differences in central carbon metabolism between human-adapted Mtb strains and animal-adapted strains such as BCG as well as differences in enzymes involved in metabolism, such as isocitrate lyase (ICL), which have been shown to be essential for pathogenesis (Rhee et al., 2011). These findings highlight the importance of Mtb metabolic processes during infection. Despite this, few groups have investigated the link between genomics and Mtb metabolism in Mtb lineages specifically.

Oyas, et. al., published the only analysis investigating the impact of Mtb genomics on downstream metabolic pathways and compared these differences across Mtb lineages (Øyås et al., 2020). The authors had a small sample size of 18 Mtb strains from lineages 1, 2, 3, 4, 5, and 6 (China, India, Philippines, Ethiopia, Afghanistan, USA, Ghana, Germany, Sierra Leone, and The Gambia) and which represented a diverse range of strains within each lineage, including strains from different spoligotypes where possible. SNP data from these strains was integrated with exometabolites (metabolites secreted from Mtb cells). Due to the small sample size, they derived a novel integration approach using a constraint-based model involving flux balance analysis which estimated metabolic fluxes through networks as a result of the SNPs. Using this analysis, they predicted the TCA cycle, pyruvate metabolism, and glycolysis pathways, as well as glycine, serine, and threonine metabolism included the greatest number of functional SNPs and the largest SNP effects, as calculated by their model. They further modelled deletions of genes caused by at least one predicted functional SNP and measured the resulting epistatic interactions. Using this approach, strains from Mtb lineages 1 and 3 were enriched in glycine, serine, and threonine metabolic pathway perturbations, and strains from lineages 2, 4, and 6 were enriched in glycolysis perturbations. Overall, they confirmed one of their SNP predictions through laboratory validation and also identified several strain-specific vulnerabilities as a result of these genomic variations, suggesting that functional strain-specific variations can be identified using both genomics and metabolomics data.

While this is the first time genomic and metabolomic data have been integrated in this way, the approach used focussed on predicting interactions (Øyås et al., 2020). As a result, the predictions made in this study would need further validation and functional analysis, as with any bioinformatic approach. In addition, the methodology and workflow are complex to understand and implement and would therefore limit its usage by others undertaking similar analyses.

While integration of data is still a new and evolving field, certain data types, such as transcriptomics, are the focus of most integration studies. Further studies, in addition to the one performed by Oyas, et. al. (Øyås et al., 2020), of the effect of Mtb genomics on metabolism should be carried out to identify new methods to integrate these data types. This would also help to increase the throughput of investigating the biological implications of genomic events. Using these analysis approaches, insights into the biological drivers of evolution of Mtb lineages may be identified. As a result, this chapter I am to:

- Develop a novel and simple workflow for integration of genomics and metabolomics data.
- Identify spoligotype-specific and lineage-specific genomic drivers of differences in metabolomic pathways observed using Mtb strains from Vietnam.
- Identify differences between clinical Mtb strains from different lineages and the experimental reference Mtb strain H37Rv.

4.2 Results

4.2.1 Creating a multi-omics integration approach for genomic and metabolomic data

To understand the causative, i.e. pathogenic, functions of Mtb lineage strains, we selected 4-5 Mtb strains from 3 major Mtb lineages to integrate SNP and metabolite data. The lineages used were lineage 1 (Indo/Oceanic), lineage 2 (Beijing/Beijing-like), and lineage 4 (Euro/American). Lineages 1 and 2 were chosen as they have restricted geographical boundaries, whereas lineage 4 can be found globally and also contains the reference Mtb strain used for tuberculosis research in the UK, H37Rv. The strains in these lineages also represent different spoligotype groups (**Table 2.1**) and may therefore represent a broad range of Mtb genomes even within their lineages. To understand the effect of SNPs on downstream metabolites, I created a novel data integration framework (**Figure 4.1**).

As our sample size was small (14 strains) but our datasets were large (5,258 SNPs and 1,061 targeted and untargeted metabolites), both SNPs and metabolites had to be filtered (**Chapter 2.2.6**) to select the data with highest significance prior to integration. The nsSNPs (SNPs which cause a change in the amino acid sequence) were selected as I chose to focus on SNPs which are most likely to cause phenotypic changes. nsSNPs which perfectly correlated with one another (correlation cut-off of absolute 0.99) were grouped into “SNP blocks” which were named according to the first nsSNP position found in that block (explained in detail below). SNP blocks which contained a single SNP, or a SNP found in only a single Mtb strain, were removed from the analysis as they were less likely to affect downstream metabolic pathways and as differences between individual strains, and not lineages or spoligotypes, was not the focus of this analysis. This resulted in a total of 26 nsSNP blocks (explained in detail below).

In addition, the metabolite data was filtered (**Chapter 2.2.6**) to decrease the metabolite pool used for integration. Differential abundance analysis was performed on untargeted metabolites to identify metabolites which were significantly different between Mtb strains and the Mtb H37Rv reference. Strains were compared against the Mtb H37Rv reference metabolome to maintain consistency with the SNP data where SNPs were called according to differences with the H37Rv reference genome. Differential abundance analysis identified 939 significant metabolites (adjusted p -value <0.01) which

were used in all further analyses. Differential abundance analysis was also performed on the targeted metabolites, which identified 35 significant metabolites (adjusted p -value <0.01).

nsSNP blocks and significantly differentially abundant metabolites were integrated using correlation and multi-omic integration packages in R. As the total number of metabolites analysed in this study was high, untargeted and targeted metabolites were integrated with nsSNP blocks separately for ease of visualisation. nsSNP blocks and targeted metabolites were integrated through correlation analysis to identify which blocks have a high positive, or negative, correlation with targeted metabolites and thereby which groups of nsSNPs may be the source of these metabolite changes. As the number of untargeted metabolites was larger than targeted metabolites (1,020 untargeted metabolites, 41 targeted metabolites), nsSNP blocks and untargeted metabolites were integrated using the DIABLO tool from the mixOmics package in R for ease of visualisation (Singh et al., 2019). This tool also correlates two variables, in this case nsSNP blocks and untargeted metabolites, and so was used to complement the correlation analysis performed on the nsSNP blocks and targeted metabolites. However, DIABLO produces a circoplot which can be more easily interpreted for datasets with higher numbers of features, as in this case with the nsSNP and untargeted metabolite comparison.

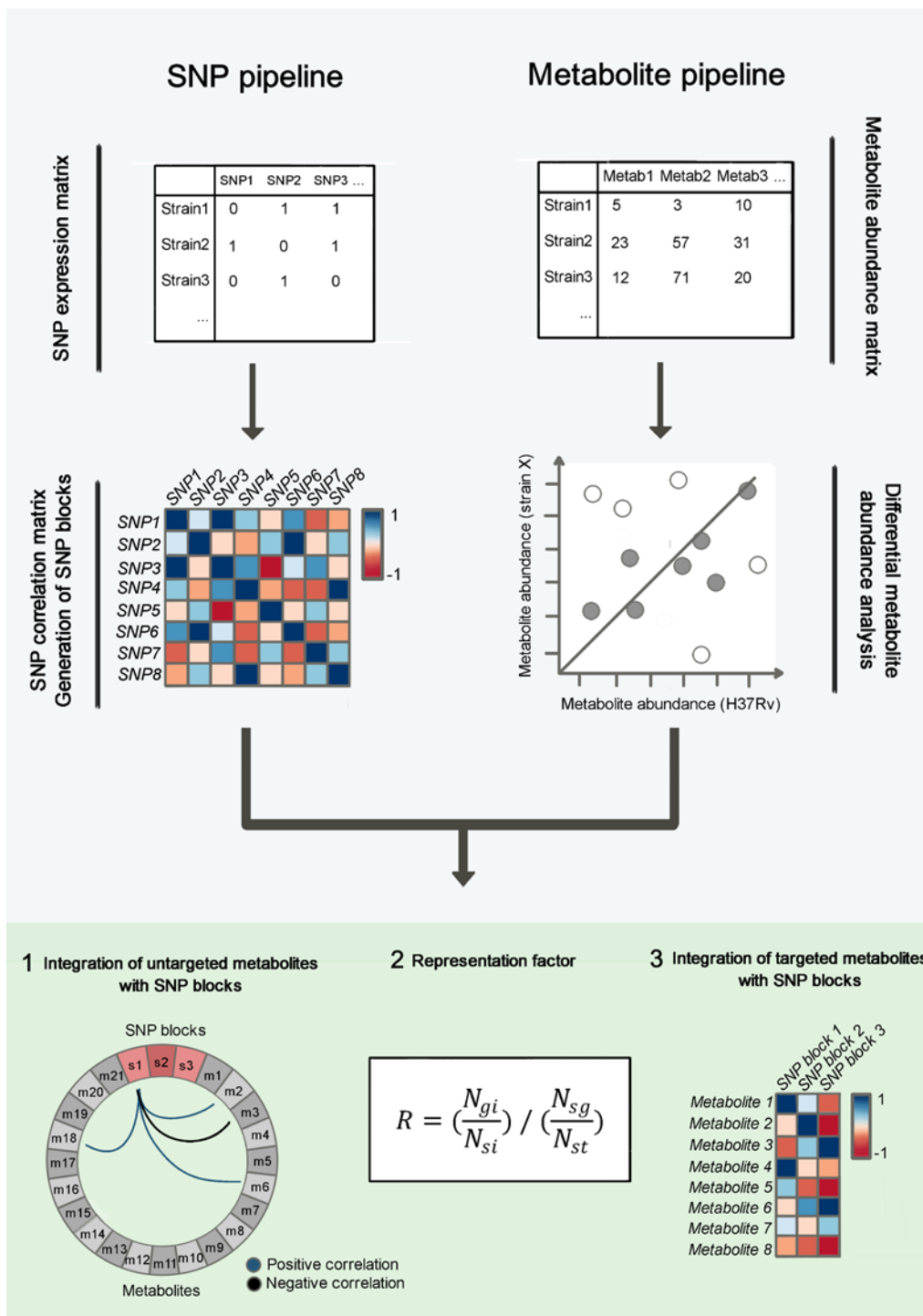


Figure 4.1: Genomic and metabolomic data integration workflow. Pre-processing pipelines for SNP (left) and metabolite (right) datasets are shown. SNPs were identified in each *Mtb* strain compared to H37Rv. nsSNPs were selected from all SNPs and were correlated to identify nsSNP blocks of perfectly correlated nsSNPs. SNP heatmap indicates nsSNP correlation scaled from absolute negative correlation (-1) to absolute positive correlation (1). These nsSNP blocks were used for downstream integration analysis with metabolites. Linear regression analysis was performed on normalised metabolite abundances comparing differential metabolite abundances in *Mtb* strains with H37Rv metabolite abundance. Differentially abundant metabolites were integrated with nsSNP blocks. The first integration stage involved the integration of differentially abundant untargeted metabolites with nsSNP blocks through correlation analysis using DIABLO (Singh et al., 2019b). The second integration stage involved the identification of over-represented genes within each nsSNP block using the representation score. The third integration stage involved the integration of differentially abundant targeted metabolites with nsSNP blocks using integration analysis.

4.2.2 SNP calling analysis and genomic-based clustering of Mtb strains

While the Mtb strains used in this study have been previously assigned to Mtb lineages, this assignment is carried out through spoligotyping. This process involves sequencing selected regions of the genome used as a baseline for lineage categorising. To confirm, through a global genome approach, that these strains do indeed belong to their assigned lineages, the clustering of Mtb strains based on their SNPs was interrogated. SNPs from all strains were called compared to the H37Rv reference genome, so that they could be later correlated to metabolite abundances for each strain compared to the same reference strain. Mtb strains grouped by their lineages when looking at the total number of SNPs and nsSNPs (**Figure 4.2A**). From a total of 5,258 SNPs across all Mtb strains, 3,501 of these were nsSNPs. Mtb strains in lineage 4 had the lowest number of SNPs, and nsSNPs, compared to the reference strain H37Rv, which is expected as the reference strain H37Rv originates from this lineage. This was followed by Mtb strains in lineage 2 then lineage 1. To visualise the placement of these strains within a more global dataset, all SNPs from these strains and an additional 95 strains from publicly available datasets representing lineages 1 – 7 were analysed using a phylogenetic tree (**Figure 4.2B**). The Mtb strains within the global dataset were aligned to the same reference genome and filtered similarly to the Vietnamese strains described in this analysis. The new Mtb strains in this analysis clustered according to their lineages within the global dataset. These strains were, however, separated within their lineages, likely due to genomic differences between spoligotypes. To identify if the presence and absence of specific SNPs, rather than the total number, would group the Mtb strains by lineage, unsupervised hierarchical clustering was performed (**Figure 4.2C**). Mtb strains did cluster into their lineages through clustering with all SNPs and nsSNPs, with both results indicating that strains in lineages 2 and 4 are more closely related to each other than they are with strains from lineage 1. This is in accordance with previously published results (Menardo et al., 2019) which also indicate that lineages 2 and 4 show higher genome similarities than they do with lineage 1.

To predict the downstream functional consequences of the nsSNPs, the nsSNPs were categorised by function categories using BioCyc annotation (Karp et al., 2019). There were an initial 5,406 functional categories identified, suggesting that certain nsSNPs were involved in multiple pathways. SNPs which were found in without a known function were excluded from further analysis. Furthermore, functional categories which consisted of fewer than 5 nsSNPs were also excluded. After filtering (**Chapter 2.2.6**), 2,366 categories were identified (**Figure 4.2**). Functional pathways identified in lineage 4 strains included cholesterol degradation, fatty acid synthesis and degradation, and nucleoside and nucleotide

biosynthesis. Strain in lineage 2 were predicted to have nsSNPs in the same pathways as lineage 4 strains, but with the addition of nsSNPs present in genes involved in the 2-methycitrate cycle, cobalamin biosynthesis, L-methionine biosynthesis, propanoyl coA degradation, L-serine and glycine biosynthesis, sulphate assimilation and L-cysteine biosynthesis, the TCA cycle, and trehalose biosynthesis and rRNA charging. The functional pathways predicted to be affected by the nsSNPs in strains from lineage 1 included all the previously mentioned pathways as well as cell wall biosynthesis, L-arginine biosynthesis, and L-glutamine degradation.

4.2.3 Metabolomics analysis and metabolomic-based clustering of Mtb strains

To identify the downstream consequences of the differences observed in the genomes of Mtb strains, Mtb metabolism was interrogated in these strains. One of the simplest indicators of a potential change in metabolism is the bacterial growth rate, as this is governed by central carbon metabolic processes. To this end, Mtb strains were grown in supplemented 7H9 broth and growth was determined every 24 hours by OD measurement. No significant difference in growth was observed between strains in any lineage, except when comparing the growth of strains in lineage1 with strains in lineage 2 (t-test p-value of 0.0021; **Figure 4.3A**). Therefore, the genomic differences between Mtb lineages may influence growth in strains in lineages 1 and 2 and is likely affects other biological pathways.

As the functional pathway analysis predicted significant remodelling of many metabolic pathways, and bacterial growth measurements only broadly reflect metabolism, further metabolomic analysis was performed using LC-MS. LC-MS analysis was performed on the intracellular metabolome for each strain through both untargeted and targeted analysis. Untargeted analysis extracts every detectable metabolite within a sample and, in this case, identified 1,020 metabolites across all Mtb strains. First, all untargeted metabolites were used to cluster Mtb strains using PCA (**Figure 4.3C, 4.3D**). Strains clustered by their lineage. The lineage 2 and 4 clusters overlapped as lineage 2 and 4 strains were more closely related than they were to lineage 1 strains, which separated distinctly. Strains also clustered by their spoligotyping group within each lineage.

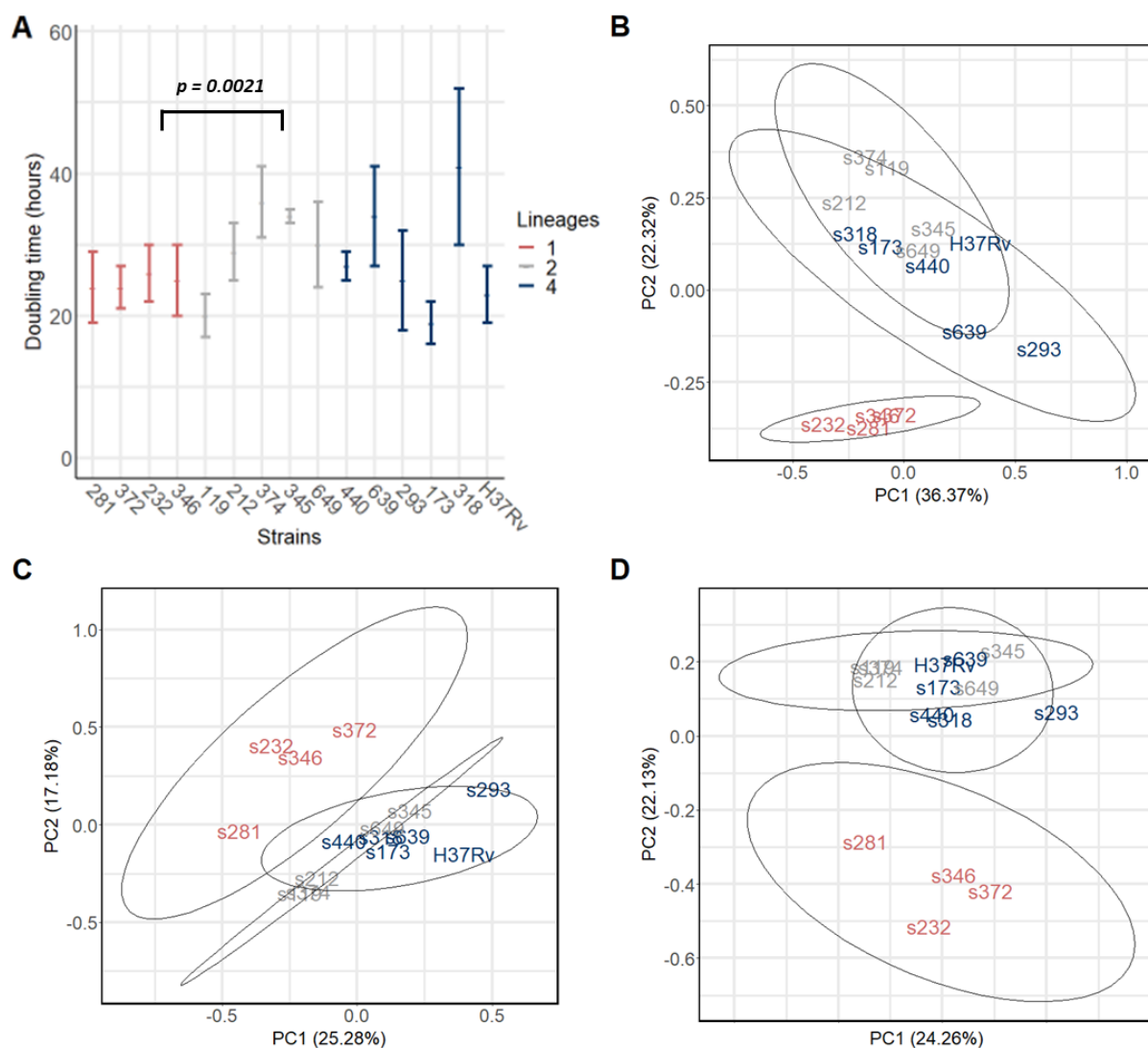


Figure 4.3: All metabolites cluster Mtb strains by their lineage. A.) Doubling times in hours of Mtb strains in lineage order. $N = 4$. **B.)** PCA clustering of Mtb strains based on all targeted metabolites. $N = 6$ technical replicates per strain. **C.)** PCA clustering of Mtb strains based on all untargeted metabolites from negative ion mode. $N = 6$ technical replicates per strain. **D.)** PCA clustering of Mtb strains based on all untargeted metabolites from positive ion mode. $N = 6$ technical replicates per strain. Pink = lineage 1 strains; grey = lineage 2 strains; blue = lineage 4 strains.

The majority of metabolites, 943/1,020, were differentially abundant compared to the reference H37Rv strain (false discovery rate < 0.05). The differentially abundant metabolites between each strain and H37Rv were used to repeat the clustering analysis through PCA (**Figure 4.4B and Figure 4.4C**). Strains clustered by their lineage groups. Strains from lineages 2 and 4 were less segregated but were distinct from strains found in lineage 1. In addition, the individual strains also clustered according to their spoligotype category within each lineage, which may reflect sub-lineage metabolic differences. As a similar pattern was observed in the nsSNPs clustering analysis, it is possible that these nsSNPs may be driving the differences in metabolites identified here.

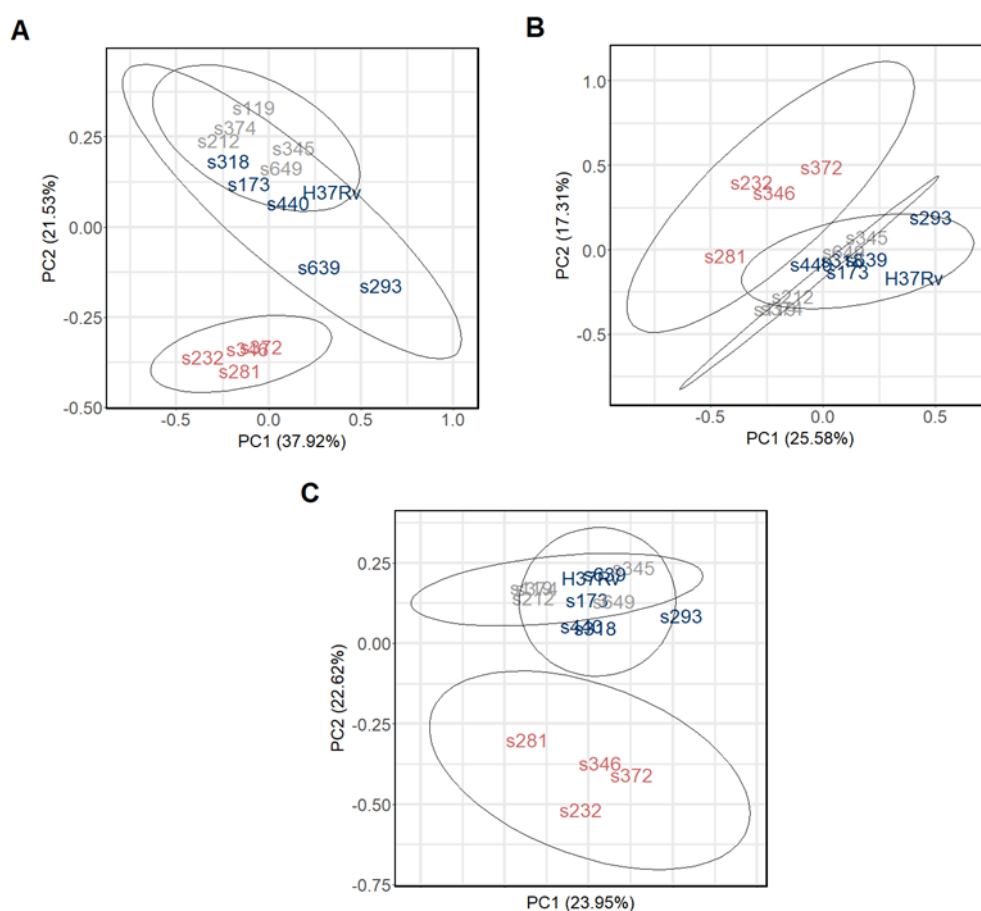


Figure 4.4: Differentially abundant metabolites cluster Mtb strains by their lineage. **A.)** PCA clustering of Mtb strains based on targeted differentially abundant metabolites. **B.)** PCA clustering of Mtb strains based on untargeted differentially abundant metabolites from negative ion mode. N = 6 technical replicates per strain. **C.)** PCA clustering of Mtb strains based on untargeted differentially abundant metabolites from positive ion mode. N = 6 technical replicates per strain. Pink = lineage 1 strains; grey = lineage 2 strains; blue = lineage 4 strains.

While untargeted metabolomics analysis can provide a global view of every metabolite in a sample, these are mainly uncharacterised or not fully assigned and therefore can be challenging to correlate with a particular metabolic pathway. As a result, and to supplement the untargeted analysis, targeted metabolomic analysis was also performed. Targeted metabolomics analysis identifies pre-defined metabolites within a sample based on their presence in major metabolic pathways of interest. As a broad number of metabolic pathways were predicted to be influenced by the genomic data, I selected a broad range of metabolites to be targeted using this analysis which are involved in nitrogen and central carbon metabolism. This included 41 metabolites involved in glycolysis, the TCA cycle, the urea cycle, peptidoglycan production, amino acids, and nucleosides. 38 of these metabolites were identified as differentially abundant between lineages (false discovery rate < 0.05; **Figure 4.5**). When comparing all strains with H37Rv, trans-aconitate and itaconic acid were identified as significantly differentially increased in almost all strains (in every strain with a median increase of 2.2 and an adjusted p-value of 1.2×10^{-7} and in every strain apart from 440 with a median increase of 2.3 and an adjusted p-value of 3.8×10^{-10} , respectively).

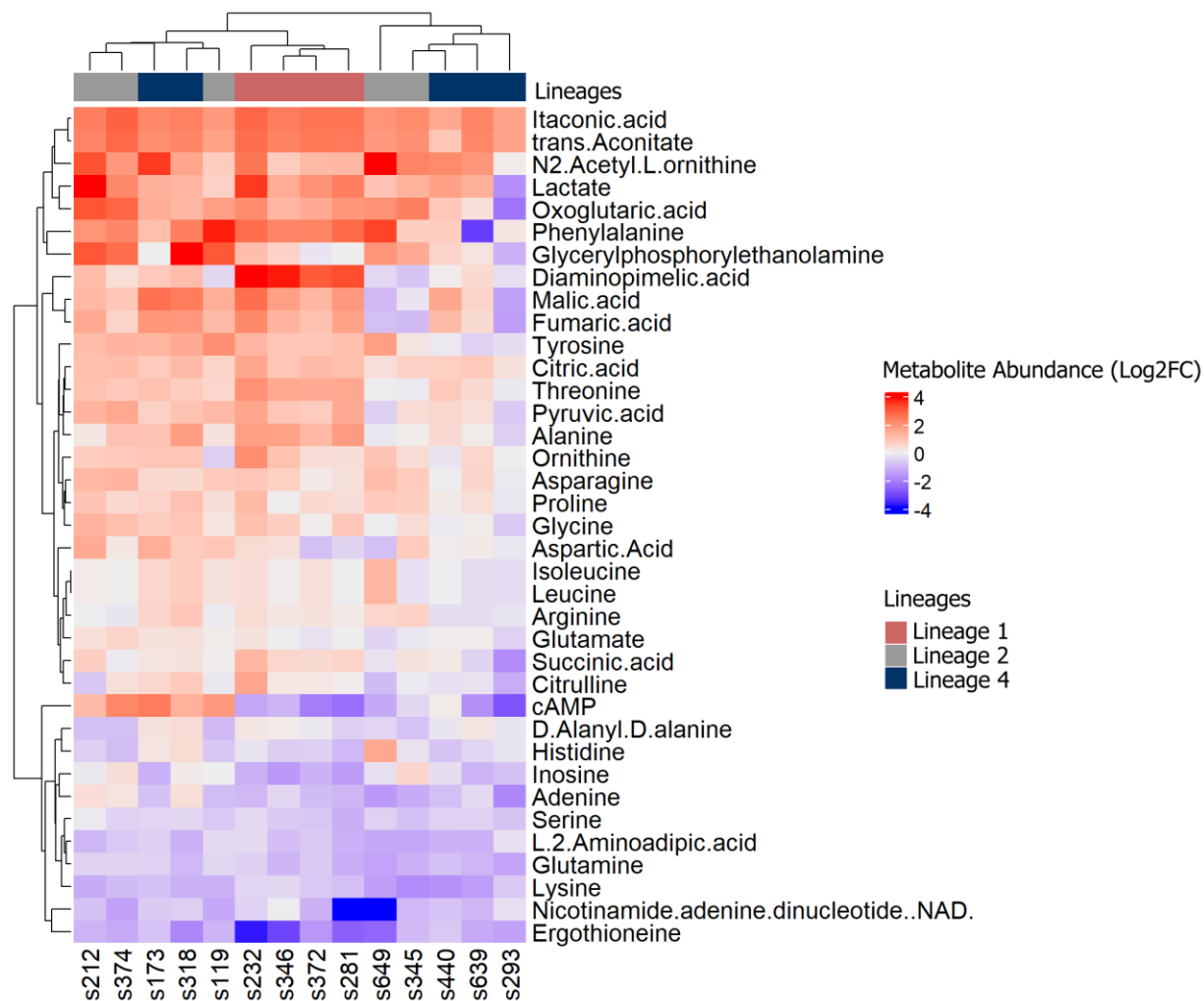


Figure 4.5: Heatmap showing log₂ fold changes for the differentially abundant targeted metabolites from at least one strain compared to the H37Rv reference strain. Pink = lineage 1 strains; grey = lineage 2 strains; blue = lineage 4 strains.

Among the differentially abundant metabolites, strains in lineage 1 showed a decrease in ergothioneine and increase in several metabolites including citric acid, alanine, threonine, and diaminopimelic acid. Within lineage 1 strains, the spoligotype EA14_NVM strains (232, 346, and 372) showed a decrease only in inosine but an increase in pyruvate and malic acid which are TCA cycle metabolites. Only glyceryl phosphoryl ethanolamine was increased in abundance in all lineage 2 strains, however pyruvate, oxoglutaric acid, asparagine, and tyrosine had higher abundance in the majority of lineage 2 strains compared to the reference strain. D-alanyl-D-alanine and lysine were found in decreased abundance compared to the reference strain in 4 lineage 2 strains, although these strains different for each metabolite. When investigating the metabolite abundances within spoligotype groups, 2-aminoadipic acid, adenine, and glutamine were found at decreased abundance in the Beijing-like strains (345 and 649). In contrast to the previous 2 lineages, no metabolites were consistently increased or decreased in all lineage 4 strains. However, both fumaric acid (173 and 293), and malic acid (173, 293, and 318) were found to be increased in abundance in a subset of lineage 4 strains compared to the reference. At spoligotype groups, ergothioneine and glutamine were decreased in all T1 (293 and 318) strains and lysine was decreased in strain 318 and H3 (440 and 639) strains. This analysis showed significant metabolic remodelling of major metabolic pathways in each lineage compared to the H37Rv reference strain, although few differences were shared between lineages.

To confirm that the differentially expressed metabolites would again cluster strains based on their lineage, unsupervised hierarchical clustering was performed using all the targeted metabolites and the significantly differentially abundant targeted metabolites. In both cases, strains clustered by their lineage and spoligotype groups as expected (**Figure 4.3B** and **Figure 4.4A**). However, strains in lineage 1 clustered closer together to strains in lineages 2 and 4 than that which was observed in the untargeted metabolomics analysis, which may suggest that the limited number of targeted metabolites does not capture fully the metabolic differences between these lineages as identified in the untargeted analysis. Together, the clustering of strains based on both metabolite analyses reflects what was observed in the genomics data as strains again cluster by lineage. This suggests that the genomics findings may be the cause of some of the differences observed in the metabolites across strains.

4.2.4 Integration of SNP and metabolite datasets

Many nsSNPs showed similar presence, or absence, in the same genes. This led to strong positive or negative correlations between several nsSNPs with the same metabolite, rendering visualisation of the results impossible and biological interpretation difficult. Therefore, instead of analysing the nsSNPs at the SNP level (i.e. individually), nsSNPs were grouped into nsSNP blocks to then be interrogated at the gene level. nsSNPs which were perfectly positively or negatively correlated were grouped into nsSNP blocks.

To identify lineage and spoligotype specific pathway perturbations caused by nsSNPs at the gene level, a representation factor was calculated for each gene in each SNP block. This was calculated based on over-representation of genes in each block. This analysis identified over-represented genes in each nsSNP block. This identified 102/116 nsSNP blocks which were found in more than one strain which were used for further analysis. Of these, there were 26 nsSNP blocks which contained nsSNPs affecting more than one gene (**Table 4.2**).

Table 4.2: nsSNP blocks containing nsSNPs which affect more than 1 gene. 1 indicates the presence of the nsSNP block in the strain, whereas 0 indicates the absence of that block. Pink = lineage 1 strains; grey = lineage 2 strains; blue = lineage 4 strains. Bold indicates nsSNP blocks which are lineage or spoligotype specific.

* indicates nsSNP blocks which are selected during downstream analyses.

nsSNP block/other	s232	s346	s372	s281	s119	s212	s374	s345	s649	s173	s293	s318	s440	s639
Spoligotype	EA14_VNM	EA14_NVM	EA14_NVM	ZERO	Beijing	Beijing	Beijing	Beijing-Like	Beijing-Like	H1	T1	T1	H3	H3
1849*	0	0	0	0	1	1	1	1	1	0	0	0	0	0
6112*	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11370	0	0	0	0	0	0	0	0	0	1	0	1	0	0
15117	1	1	1	1	1	1	1	1	1	0	0	0	0	0
21819	0	0	1	1	0	0	0	0	0	0	0	0	0	0
25137	0	0	0	0	0	0	0	0	0	0	0	0	1	1
57298	0	0	0	0	0	1	1	0	0	0	0	0	0	0
79480	0	0	0	0	1	1	1	1	0	0	0	0	0	0
103836	1	1	1	0	0	0	0	0	0	0	0	0	0	0
190646	1	1	0	0	0	0	0	0	0	0	0	0	0	0
336560	0	0	0	0	1	1	0	1	1	0	0	0	0	0
336590*	0	0	0	1	1	1	1	1	1	0	0	0	0	0
340132	0	0	0	0	1	1	1	1	1	0	0	0	1	1
340372	1	1	1	1	1	1	1	1	1	1	0	0	1	1
467526*	0	0	0	0	1	1	1	1	1	1	1	1	1	1
623472	0	0	0	1	1	1	1	1	1	1	1	0	1	1
836538	0	0	0	1	0	0	0	1	1	0	1	0	0	0
1096508	0	0	0	0	1	1	1	1	1	1	0	0	0	0
1722228	0	0	0	0	1	1	1	0	1	0	0	0	0	0
2413246*	0	0	1	1	1	1	1	1	1	0	0	0	0	0
2626600	0	0	0	0	0	0	0	0	1	0	0	1	0	0
2867458	0	0	0	0	1	1	1	1	1	0	0	1	0	0
2945042	0	0	0	0	0	0	0	0	0	0	1	0	1	0
3379726	0	0	1	1	1	1	0	1	1	0	1	1	0	0

3379742	1	1	1	1	1	1	1	1	1	0	1	1	1	0
3846866	1	1	1	0	1	1	1	1	0	1	0	0	1	0

nsSNP blocks rs1849 and rs6112 were found to be specific for strains within lineages 2 and 1, respectively (**Table 4.3** and **Table 4.4**). nsSNP block rs1849 contained 216 nsSNPs in unique genes and rs6112 contained 621 nsSNPs in unique genes. To clarify, nsSNPs found in the nsSNP block rs1849 are only present in Mtb strains from lineage 2 and not in strains from the other lineages. In both nsSNP blocks, many genes encoded for the PE and PPE families of proteins. The *PE* and *PPE* genes were also identified as the highest scoring, with representation scores significantly above average. There was no nsSNP block which contained nsSNPs present in strains of lineage 4 but not lineages 1 and 2, but the nsSNP block rs15117 did contain nsSNPs present in strains in lineages 1 and 2 but not lineage 4, resulting in a lineage 4 specific nsSNP block (**Table 4.5**). This nsSNP block cannot be interpreted as the nsSNPs in this block were identified against the reference H37Rv strain, which is also a member of lineage 4.

Understanding differences at the spoligotype level was more challenging than at the lineage level. The number of strains within each spoligotype was very low (maximum 3), and there were no nsSNP blocks specific to spoligotype groups containing at least 3 strains. It was therefore difficult to interpret nsSNP blocks at this level and the only nsSNP block which was specific to a spoligotype was rs25137 which was found only in strains from spoligotype H3 from lineage 4 (**Table 4.6**). rs25137 contained 155 nsSNPs.

Table 4.3: Representation score for genes with nsSNPs present in the rs1849 nsSNP block. Filtered for genes with known function (i.e. protein description) and representation score above 1. N strains = number of strains that gene is found in; N blocks = number of nsSNP blocks the gene is found in.

nsSNP	Gene	N strains	N blocks	Rep score	Protein
1160	PE17	1	1	70.62069	PE family protein PE17
1430	PE1	3	2	35.31034	PE family protein PE1
90	PE_PGRS4	6	3	6.62069	PE-PGRS family protein PE_PGRS4
1061	rodA	5	1	4.413793	noncanonical transglycosylase RodA
1064	leuS	5	1	4.413793	probable leucyl-tRNA synthetase
					probable nitrite reductase [NAD(P)H] large subunit [FAD
1080	nirB	5	1	4.413793	flavoprotein] NirB
1101	vapC4	5	1	4.413793	possible toxin
1104	recD	5	1	4.413793	probable exonuclease V (alpha chain)
1106	mmaA2	5	1	4.413793	methoxy mycolic acid synthase 2
					possible nitrate/nitrite response transcriptional regulatory
1115	narL	5	1	4.413793	protein
1124	lipU	5	1	4.413793	lipase LipU
1131	PPE17	5	1	4.413793	PPE family protein PPE17
1140	dinG	5	1	4.413793	probable ATP-dependent helicase
1173	ephB	5	1	4.413793	probable epoxide hydrolase
1190	cobC	5	1	4.413793	possible adenosylcobalamin/alpha-ribazole phosphatase
1191	mazF8	5	1	4.413793	possible toxin
1195	PE23	5	1	4.413793	PE family protein PE23
1197	mcr7	5	1	4.413793	Putative small regulatory RNA
1198	subI	5	1	4.413793	probable sulfate-binding lipoprotein
1199	PE24	5	1	4.413793	possible PE family-related protein PE24
1202	obg	5	1	4.413793	probable GTP1/Obg-family GTP-binding protein Obg
1204	lipQ	5	1	4.413793	probable carboxylesterase
1209	lppA	5	1	4.413793	probable conserved lipoprotein
					probable glutamine-transport ATP-binding protein ABC
1211	glnQ	5	1	4.413793	transporter
1221	fadE20	5	1	4.413793	probable acyl-CoA dehydrogenase
1224	echA16	5	1	4.413793	probable enoyl-CoA hydratase
1227	glnA4	5	1	4.413793	probable glutamine synthetase
1232	purU	5	1	4.413793	probable formyltetrahydrofolate deformylase
1255	spoU	5	1	4.413793	probable tRNA/rRNA methylase

1259	guaB2	5	1	4.413793	probable inosine-5'-monophosphate dehydrogenase
1269	ltp2	5	1	4.413793	3-oxo-23,24-bisnorchol-4-en-17-ol-22-oyl-CoA lyase
1270	fdxB	5	1	4.413793	possible electron transfer protein
1271	fadA6	5	1	4.413793	3-oxo-acyl CoA thiolase
1272	PE_PGRS58	5	1	4.413793	PE-PGRS family protein PE_PGRS58
1276	nth	5	1	4.413793	probable endonuclease III
1279	Rv3691	5	1	4.413793	Uncharacterized membrane protein Rv3691
1281	cut5b	5	1	4.413793	probable cutinase [second part]
1286	fadE35	5	1	4.413793	probable acyl-CoA dehydrogenase
163	galE3	6	1	2.206897	probable UDP-glucose 4-epimerase
860	hpx	6	1	2.206897	possible non-heme haloperoxidase
884	pcd	6	1	2.206897	probable piperidine-6-carboxylic acid dehydrogenase
885	lhr	7	2	2.206897	probable ATP-dependent helicase
1060	ppiA	7	2	2.206897	probable iron-regulated peptidyl-prolyl cis-trans isomerase A
1073	sigG	7	2	2.206897	probable alternative RNA polymerase sigma factor
1083	eccD3	6	1	2.206897	ESX conserved component EccD3. ESX-3 type VII secretion
1112	aldA	6	1	2.206897	system protein. Probable transmembrane protein.
1113	purF	6	1	2.206897	probable aldehyde dehydrogenase NAD dependent
1129	pimE	6	1	2.206897	amidophosphoribosyltransferase
1133	mmpl10	6	1	2.206897	phosphatidylinositol tetramannoside α-(1-2)-
1142	lprF	6	1	2.206897	mannosyltransferase
1143	carA	6	1	2.206897	probable conserved transmembrane transport protein
1146	lprG	6	1	2.206897	probable conserved lipoprotein
1148	bisC	6	1	2.206897	probable carbamoyl-phosphate synthase small chain
1155	Rv1506c	7	2	2.206897	conserved lipoprotein
1168	fadD31	6	1	2.206897	probable biotin sulfoxide reductase
1170	fadE17	7	2	2.206897	Putative methyltransferase Rv1506c
1187	pafB	6	1	2.206897	probable acyl-CoA ligase
1188	pyrD	6	1	2.206897	probable acyl-CoA dehydrogenase
1192	cyp121	6	1	2.206897	proteasome accessory factor B
1261	mycP4	6	1	2.206897	probable dihydroorotate dehydrogenase
1263	lipF	7	2	2.206897	mycocyclusin synthase
1273	lysS	7	2	2.206897	probable membrane-anchored mycosin
1274	esxW	6	1	2.206897	probable esterase/lipase
					lysyl-tRNA synthetase 1
					putative ESAT-6 like protein

1287	aftB	6	1	2.206897	G185E-8101-MONOMER
1289	glpQ1	6	1	2.206897	probable glycerophosphoryl diester phosphodiesterase
1294	gid	6	1	2.206897	probable glucose-inhibited division protein B Gid
1185	pknJ	8	3	1.655172	transmembrane serine/threonine-protein kinase J
707	lppB	10	9	1.241379	probable conserved lipoprotein
1069	PPE1	7	1	1.103448	PPE family protein PPE1
1153	lipL	7	1	1.103448	β-lactamase / esterase LipL

Table 4.4: Representation score for the top 50 genes with nsSNPs present in the rs6112 nsSNP block. Filtered for genes with known function (i.e. protein description) and representation score above 1. N strains = number of strains that gene is found in; N blocks = number of nsSNP blocks the gene is found in.

nsSNP	Gene	N strains	N blocks	Rep score	Protein
1164	PE19	1	2	141.2414	PE family protein PE19
1474	PE18	3	3	52.96552	PE family protein PE18
3	gyrB	4	1	8.827586	DNA gyrase (subunit B)
5	alaT	4	1	8.827586	tRNA-Ala
35	ctpA	4	1	8.827586	cation transporter P-type ATPase a
44	ldtA	4	1	8.827586	G185E-4233-MONOMER
49	ephF	4	1	8.827586	possible epoxide hydrolase
58	PE4	4	1	8.827586	PE family protein PE4
59	adhE1	4	1	8.827586	alcohol dehydrogenase
60	fadD5	4	1	8.827586	probable fatty-acid-CoA ligase
61	yrbE1A	4	1	8.827586	conserved integral membrane protein
68	ilvD	4	1	8.827586	probable dihydroxy-acid dehydratase
76	lipW	4	1	8.827586	possible esterase
78	lipC	4	1	8.827586	probable esterase
82	vapB24	4	1	8.827586	possible antitoxin
83	htdX	4	1	8.827586	probable 3-hydroxyacyl-thioester dehydratase
87	fadD2	4	1	8.827586	probable fatty-acid-CoA ligase
89	fadE6	4	1	8.827586	probable acyl-CoA dehydrogenase
96	vapB2	4	1	8.827586	possible antitoxin
111	dnaK	4	1	8.827586	probable chaperone protein
112	dnaJ1	4	1	8.827586	probable chaperone protein
138	Rv0433	4	1	8.827586	Putative glutamate--cysteine ligase 2
148	umaA	4	1	8.827586	possible mycolic acid synthase
155	proC	4	1	8.827586	probable pyrroline-5-carboxylate reductase
175	yrbE2A	4	1	8.827586	conserved hypothetical integral membrane protein
179	lpqO	4	1	8.827586	probable conserved lipoprotein

186	echA3	4	1	8.827586	probable enoyl-CoA hydratase
188	secE1	4	1	8.827586	probable preprotein translocase
191	mkl	4	1	8.827586	possible ribonucleotide-transport ATP-binding protein ABC transporter
196	fadE8	4	1	8.827586	probable acyl-CoA dehydrogenase
204	rplO	4	1	8.827586	50S ribosomal protein L15
209	xylB	4	1	8.827586	possible D-xylulose kinase
212	vapC31	4	1	8.827586	possible toxin VapC31. Contains PIN domain. possible zinc-containing alcohol dehydrogenase
218	adhB	4	1	8.827586	NAD dependent
222	purB	4	1	8.827586	probable adenylosuccinate lyase
237	pip	4	1	8.827586	probable proline iminopeptidase
256	Rv0909	4	1	8.827586	Antitoxin Rv0909
258	betP	4	1	8.827586	possible glycine betaine transport integral membrane protein
263	mntH	4	1	8.827586	divalent cation-transport integral membrane protein
265	pstS3	4	1	8.827586	periplasmic phosphate-binding lipoprotein
267	pstS2	4	1	8.827586	periplasmic phosphate-binding lipoprotein
272	sucD	4	1	8.827586	succinyl-CoA ligase [ADP-forming] α subunit mycobacterial persistence regulator MRPA (two component response transcriptional regulatory protein)
284	mprA	4	1	8.827586	
293	moeA1	4	1	8.827586	molybdopterin molybdenumtransferase 1
296	pabB	4	1	8.827586	probable para-aminobenzoate synthase component I PABD
301	mfd	4	1	8.827586	probable transcription-repair coupling factor
302	lpqU	4	1	8.827586	probable conserved lipoprotein
306	trcS	4	1	8.827586	two component sensor histidine kinase
310	fadD14	4	1	8.827586	probable medium chain fatty-acid-CoA ligase
312	PE_PGRS20	4	1	8.827586	PE-PGRS family protein PE_PGRS20

Table 4.5: Representation score for the genes with nsSNPs present in the rs15117 nsSNP block. N strains = number of strains that gene is found in; N blocks = number of nsSNP blocks the gene is found in.

nsSNP	Gene	N strains	N blocks	Rep score
1474	PE18	3	52.96552	PE family protein PE18
707	lppB	10	1.241379	probable conserved lipoprotein
40	Rv0104	9	0.551724	NA
73	Rv0210	9	0.551724	NA
147	Rv0465c	9	0.551724	NA
193	atsD	9	0.551724	possible arylsulfatase
287	Rv0987	9	0.551724	NA
342	PPE18	9	0.551724	PPE family protein PPE18
356	Rv1248c	9	0.551724	NA
380	glgB	9	0.551724	1,4-alpha-glucan branching enzyme
533	Rv1855c	9	0.551724	NA
621	fadD15	9	0.551724	long-chain-fatty-acid-CoA ligase
697	Rv2484c	9	0.551724	NA
748	Rv2714	9	0.551724	NA
750	Rv2719c	9	0.551724	NA
756	PPE44	9	0.551724	PPE family protein PPE44
802	Rv2971	9	0.551724	NA
858	Rv3161c	9	0.551724	NA
971	ftsH	9	0.551724	membrane-bound protease
983	dppA	9	0.551724	probable periplasmic dipeptide-binding lipoprotein
984	Rv3671c	9	0.551724	NA
1056	Rv3916c	9	0.551724	NA
466	Rv1634	10	0.275862	NA
625	adoK	10	0.275862	adenosine kinase
704	fas	10	0.275862	fatty acid synthase
303	eno	11	0.137931	probable enolase Eno

304	kdpF	11	0.137931	probable membrane protein
509	Rv1773c	11	0.137931	NA
996	Rv3729	11	0.137931	NA
538	lldD2	12	0.103448	possible L-lactate dehydrogenase (cytochrome)
842	lipY	12	0.068966	PE-PGRS family protein, triacylglycerol lipase
594	Rv2082	14	0.025862	NA
992	dnaQ	14	0.025862	probable DNA polymerase III (epsilon subunit)
197	mmpL5	14	0.017241	probable conserved transmembrane transport protein
900	PPE55	14	0.017241	PPE family protein PPE55
				ESX conserved component EccC2.
				ESX-2 type VII secretion system protein.
1049	eccC2	14	0.017241	Possible membrane protein.

Table 4.6: Representation score for genes with nsSNPs present in the rs25137 nsSNP block. Filtered for genes with known function (i.e. protein description) and representation score above 1. N strains = number of strains that gene is found in; N blocks = number of nsSNP blocks the gene is found in.

nsSNP	Gene	N strains	N blocks	Rep score	nsSNP
1474	PE18	3	3	52.96552	PE family protein PE18
1776	fhaA	2	1	35.31034	conserved protein with FHA domain,
1778	treS	2	1	35.31034	trehalose synthase
1779	fbpC	2	1	35.31034	antigen 85C
1780	lprK	2	1	35.31034	possible Mce-family lipoprotein
1783	gabD1	2	1	35.31034	NADP-dependent succinate-semialdehyde dehydrogenase
1789	lpdC	2	1	35.31034	dihydrolipoamide dehydrogenase subunit
1791	hemD	2	1	35.31034	probable uroporphyrin-III C-methyltransferase
1792	htpX	2	1	35.31034	probable protease transmembrane protein heat shock protein
1795	Rv0571c	2	1	35.31034	Putative phosphoribosyl transferase Rv0571c
1797	vapC27	2	1	35.31034	possible toxin VapC27. Contains PIN domain.
1799	rplD	2	1	35.31034	50S ribosomal protein L4
1800	ercc3	2	1	35.31034	DNA helicase
1801	citA	2	1	35.31034	probable citrate synthase II
1804	ksgA	2	1	35.31034	probable dimethyladenosine transferase
1808	mrp	2	1	35.31034	iron-sulfur cluster carrier protein
1812	lprA	2	1	35.31034	possible lipoprotein
1816	qor	2	1	35.31034	probable quinone reductase
1818	nadB	2	1	35.31034	probable L-aspartate oxidase
1819	hisF	2	1	35.31034	probable cyclase
1830	ureG	2	1	35.31034	urease accessory protein
1832	adhA	2	1	35.31034	probable alcohol dehydrogenase bifunctional enzyme MbtA: salicyl-AMP ligase (SAL-AMP ligase) + salicyl-S-ArCP
1845	mbtA	2	1	35.31034	synthetase
1853	ppsD	2	1	35.31034	phenolphthiocerol synthesis type-I polyketide synthase
1855	gltS	2	1	35.31034	glutamyl-tRNA synthetase
1858	prfB	2	1	35.31034	probable peptide chain release factor 2
1862	upp	2	1	35.31034	probable uracil phosphoribosyltransferase
1864	truA	2	1	35.31034	probable tRNA pseudouridine synthase a
1865	rpsK	2	1	35.31034	30S ribosomal protein S11
1866	fadE27	2	1	35.31034	3-oxocholest-4-en-26-oyl-CoA dehydrogenase β subunit
1869	ask	2	1	35.31034	aspartokinase

178	mce2C	3	1	17.65517	mce-family protein Mce2C
1235	fecB	3	1	17.65517	probable FEIII-dictrate-binding periplasmic lipoprotein
1247	ppk2	3	1	17.65517	polyphosphate kinase
					ESX conserved component EccB4. ESX-4 type VII secretion system protein. Probable
1420	eccB4	4	2	17.65517	membrane protein.
1491	scoB	3	1	17.65517	probable succinyl-CoA:3-ketoacid-coenzyme A transferase (beta subunit)
1580	nrdZ	3	1	17.65517	probable ribonucleoside-diphosphate reductase (large subunit)
1587	mca	3	1	17.65517	G185E-5244-MONOMER
1595	zwf2	3	1	17.65517	probable glucose-6-phosphate 1-dehydrogenase
1633	whiB4	3	1	17.65517	probable transcriptional regulatory protein WhiB-like
1677	lppC	3	1	17.65517	probable lipoprotein
1848	mrr	3	1	17.65517	probable restriction system protein
1299	mmpL11	4	1	8.827586	probable conserved transmembrane transport protein
1492	Rv2565	4	1	8.827586	Uncharacterized NTE family protein Rv2565
1125	celA2b	6	2	4.413793	probable cellulase
141	moeA2	6	2	4.413793	molybdopterin molybdenumtransferase 2
					probable acetyl-/propionyl-coenzyme A carboxylase alpha chain (alpha subunit)
280	accA2	6	2	4.413793	AccA2: biotin carboxylase + biotin carboxyl carrier protein (BCCP)
343	esxL	6	2	4.413793	putative ESAT-6 like protein
387	moeY	6	2	4.413793	possible molybdopterin biosynthesis protein
475	argH	6	2	4.413793	argininosuccinate lyase subunit
572	ctpF	6	2	4.413793	probable metal cation transporter P-type ATPase A
575	otsB1	6	2	4.413793	trehalose-6-phosphate phosphatase
					possible transmembrane phospholipid biosynthesis bifunctional enzyme PlsC:
					putative L-3-phosphoserine phosphatase (O-phosphoserine phosphohydrolase)
					(PSP) (pspase) + 1-acyl-SN-glycerol-3-phosphate acyltransferase (1-AGP
696	plsC	6	2	4.413793	acyltransferase) (1-AGPAT) (lysophosphatidic acid acyltransferase) (LPAAT)
733	PE_PGRS46	6	2	4.413793	PE-PGRS family protein PE_PGRS46
783	fdhF	6	2	4.413793	possible formate dehydrogenase H
969	clpC1	6	2	4.413793	probable ATP-dependent protease ATP-binding subunit
38	nrp	7	2	2.206897	probable peptide synthetase
74	mmpL3	7	2	2.206897	possible conserved transmembrane transport protein
384	mbtM	7	2	2.206897	probable fatty acyl-AMP ligase
1073	sigG	7	2	2.206897	probable alternative RNA polymerase sigma factor
1155	Rv1506c	7	2	2.206897	Putative methyltransferase Rv1506c
1170	fadE17	7	2	2.206897	probable acyl-CoA dehydrogenase

1263	lipF	7	2	2.206897	probable esterase/lipase
1273	lysS	7	2	2.206897	lysyl-tRNA synthetase 1
1185	pknJ	8	3	1.655172	transmembrane serine/threonine-protein kinase J
398	carB	8	2	1.103448	probable carbamoyl-phosphate synthase large chain
410	lipO	8	2	1.103448	probable esterase

To measure the effect of the differences in nsSNPs on the differences in metabolite abundances, the nsSNP blocks and metabolite datasets were integrated. Different correlation-based approaches were used to integrate the targeted and untargeted metabolites with the nsSNPs blocks to allow for effective visualisation of the integration results. First, the differentially abundant metabolites from the untargeted analysis were correlated with the nsSNP blocks using DIABLO, a multi-omics integration tool from the mixOmics package in R (Singh et al., 2019). Alongside correlation, DIABLO includes a feature selection step, which selects the features (in this case metabolites or nsSNP blocks) which it identifies as most likely to correlate with features from the other datasets being integrated. However, as the metabolites were previously filtered by significance, and therefore a selection step was manually conducted beforehand, all differentially abundant metabolites were included. To aid in visualisation, and select the most relevant nsSNP blocks, feature selection was performed on the nsSNP blocks to select 3 blocks to correlate with the untargeted metabolites. This number was chosen as it selected the nsSNP blocks which contained nsSNPs found in all strains in Mtb lineage 1 (rs6112) and lineage 2 (rs1849), as well one other nsSNP block. While there was no lineage 4 specific nsSNP block (containing nsSNPs found in all Mtb strains in lineage 4 but not the other 2 lineages), the nsSNP block rs15117 contained nsSNPs found in all strains in Mtb lineages 1 and 2 but not in lineage 4. This block was therefore specific for lineage 4, but interestingly it was not selected in the DIABLO feature selection step.

This technique produced a circoplot displaying the correlations between the nsSNP blocks and differentially abundant metabolites (**Figure 4.6**). Different correlation cut-offs were selected for plots correlating metabolites extracted from different ion modes. The cut-offs were determined as multiples of 5 which produced few enough correlations that individual correlations could be interpreted. This resulted in higher cut-offs for the metabolites identified using positive ion mode, as nsSNP blocks correlated with a greater number of these than metabolites identified using negative ion mode. When correlating negative ion mode metabolites, the nsSNP blocks rs467526, rs6112, and rs368092 were selected from the first component and nsSNP blocks rs2413246, rs336590, and rs1849 were selected from the second component. In the positive ion mode metabolite plots, the same nsSNP blocks were selected for both components, yet they correlated with different metabolites.

nsSNP blocks rs467526 and rs6112 positively and negatively correlated with several metabolites while nsSNP blocks rs2413246 and rs336590 only negatively correlated with several metabolites. The nsSNP block rs467526 is found in all strains in lineages 2 and 4 but not any strains in lineage 1. Therefore, the metabolites correlating with this block are specific to lineages 2 and 4. The nsSNP block rs336590 is

specific to all lineage 2 strains but also strain 281, found in lineage 1. The nsSNP block rs2413246 is also found in all lineage 2 strains but also strains 372 and 281, both found in lineage 1. Therefore, there is a strong correlation of lineage 2 strains with many metabolites, which has been selected as more highly correlated than correlations between metabolites and strains in lineages 1 and 4. The nsSNP block rs368092 was the only block to only contain nsSNPs affecting a single gene, *PPE5*. This block is found in strains 281, 232, and 346, which are 3/4 of the lineage 1 strains in this analysis.

Despite selecting 3 nsSNP blocks for correlation with the untargeted metabolites, not all nsSNP blocks showed high correlation with metabolites using the cut-offs here. Specifically, nsSNP block rs368092 showed very few high correlations across plots and nsSNP block rs1849 showed no high correlations in any of the analyses.

Untargeted metabolomics analysis allows the study of multiple metabolites in an unsupervised way. However, this can be limiting as the metabolites identified are not characterised. LC-MS/MS can be used to characterise these untargeted metabolites, and it was implemented here to identify the most highly correlated metabolites. Unfortunately, it was unable to accurately identify any metabolites in this analysis without further work using metabolite standards, which was outside of the scope of this Chapter.

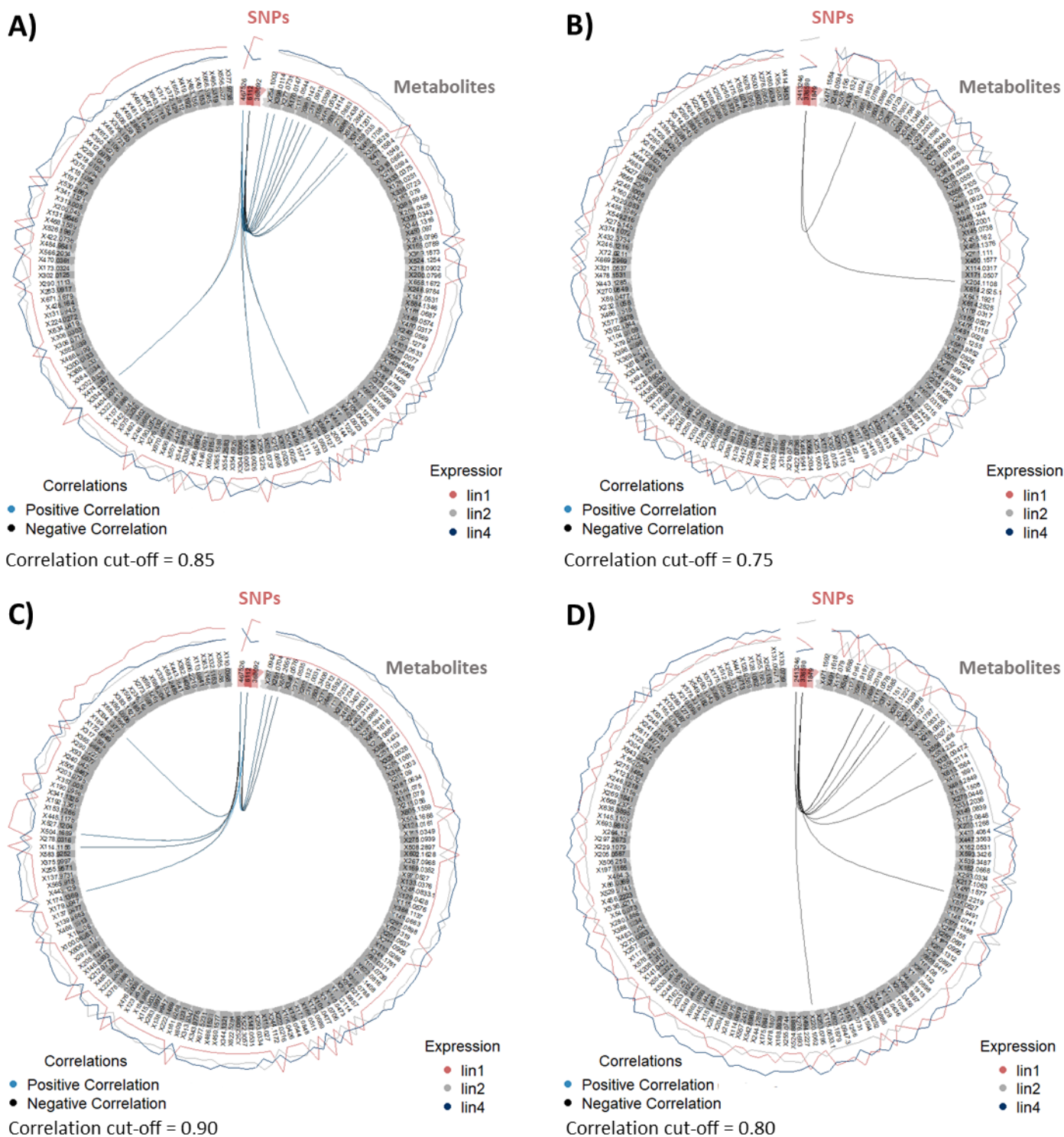


Figure 4.6: DIABLO integration analysis identifies nsSNP blocks which correlate with significantly differentially abundant untargeted metabolites. A) Correlations of the 3 selected nsSNP blocks with untargeted metabolites from negative ion mode in component 1. **B)** Correlations of the 3 selected nsSNP blocks with untargeted metabolites from negative ion mode in component 2. **C)** Correlations of the 3 selected nsSNP blocks with untargeted metabolites from positive ion mode in component 1. **D)** Correlations of the 3 selected nsSNP blocks with untargeted metabolites from positive ion mode in component 2.

Next, nsSNP blocks were integrated with targeted metabolite data. Integration through correlation analysis was performed and, as a result of the small number of targeted metabolites, was visualised through a correlation plot (**Figure 4.7**). High correlations (absolute 0.75) were observed between many targeted metabolites within the same metabolic pathway (e.g. pyruvic acid, citric acid, succinic acid, fumaric acid, malic acid and glutamic acid, which are in the TCA cycle). Strong correlations were also observed between certain nsSNP blocks, such as rs6112, rs103823, rs368092, rs1481185, rs3137237 and rs3735749 (those with similar presence or absence in some of the same strains). This indicates that there might be high redundancy between nsSNPs and between metabolites, thereby making conclusions of correlations between the two challenging.

Between nsSNP blocks and targeted metabolites, there were several strong correlations, suggesting that certain genomic changes are likely driving the differences observed in metabolites between strains. However, it is difficult to attribute biological meaning to these correlations as several nsSNP blocks were highly correlated to several metabolites. To investigate specific associations, correlations were filtered by metabolites (**Chapter 2.2.6**) which were differentially abundant in the same strains which included the nsSNP blocks. From this analysis, there was a strong negative correlation (-0.74) between ergothioneine (which was found to be negatively differentially abundant in all strains in lineage 1) and the nsSNP block rs6112 (which represented nsSNPs found in all strains in lineage 1) (**Figure 4.7 and Figure 4.8A**). While there were many significantly high correlations between nsSNP blocks and metabolites, this was the only significant and strong correlation with a metabolite which was also shown to be differentially abundant across the same strains (**Figure 4.8B**), and therefore is the most likely to be successfully validated experimentally. Genes affected by the nsSNPs in this nsSNP block can be seen in **Table 4.4**. None of the genes in this block were directly involved in the ergothioneine synthesis pathway (e.g. *egtE*, *egtD*, *egtC*, *egtB*, or *egtA*) (Cumming, Chinta, et al., 2018a).

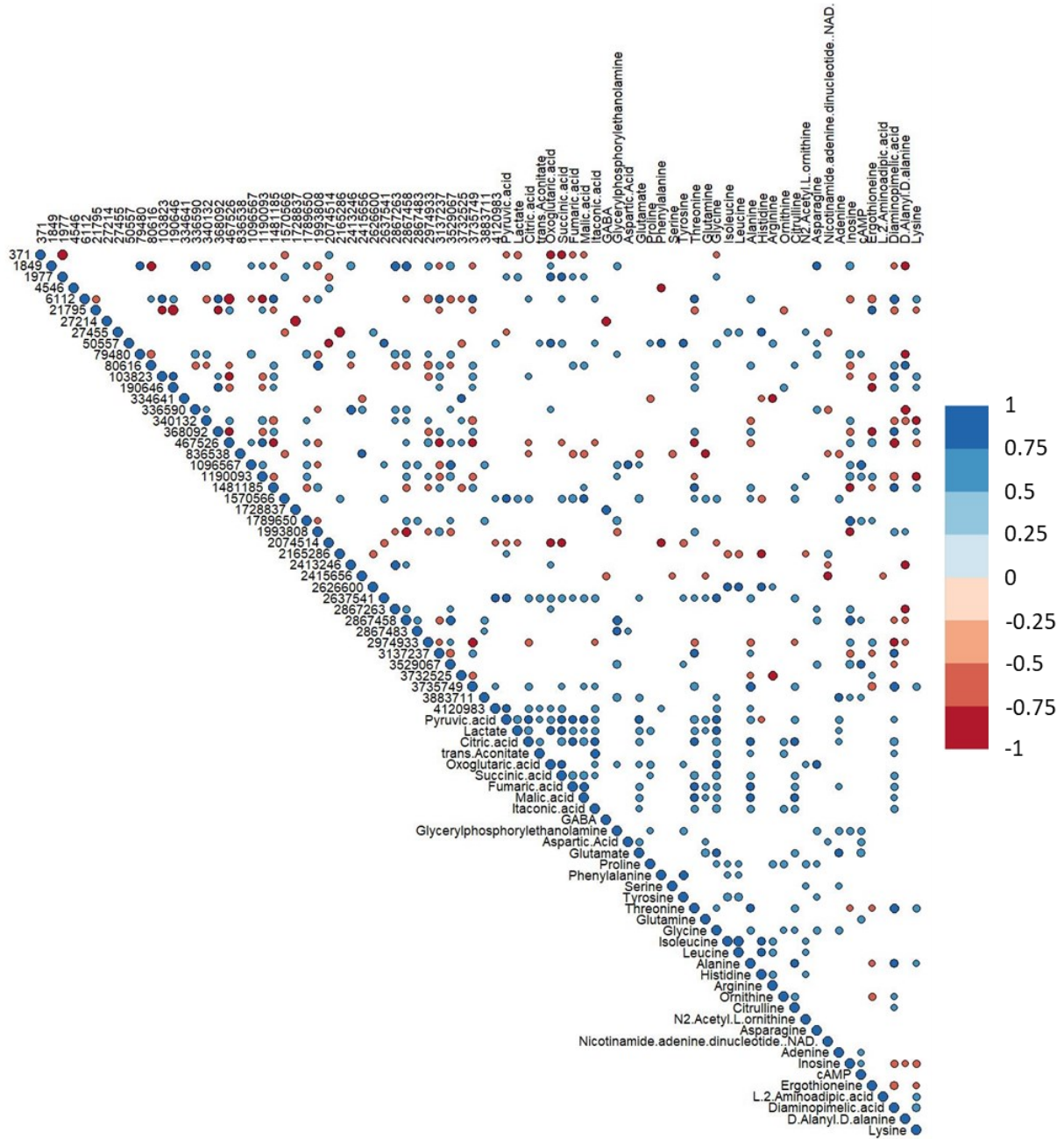
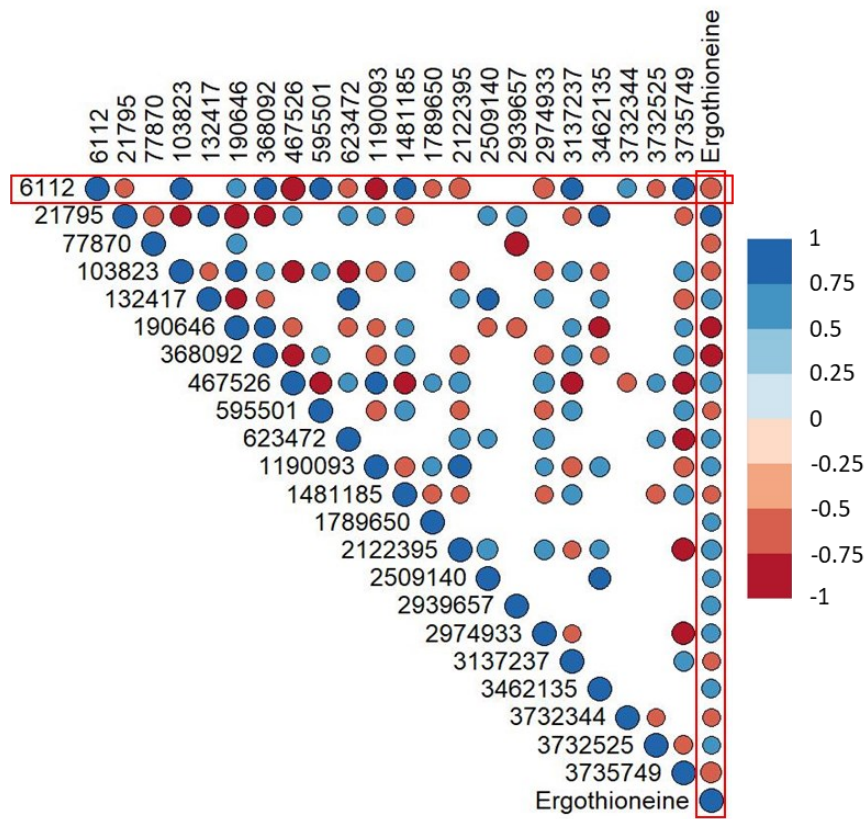


Figure 4.7: Correlation analysis identifies nsSNP block correlations with differentially abundant targeted metabolites. Blue dots indicate positive correlations and red dots indicate negative correlations. White space indicates correlations which did not reach statistical significance based on a p -value cut-off of <0.05 .

A



B

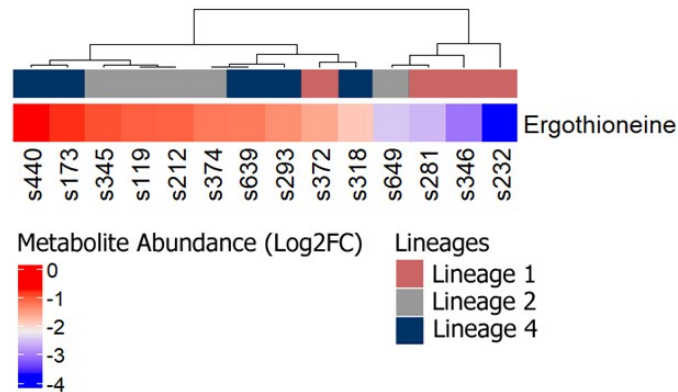


Figure 4.8: Correlation analysis identifies nsSNP block correlations with ergothioneine. A) Blue dots indicate positive correlations and red dots indicate negative correlations. White space indicates correlations which did not reach statistical significance based on a p -value cut-off of <0.05 . Red boxes indicate the correlation between nsSNP block rs6112 and ergothioneine. **B)** Clustered heatmap displaying log2 fold change values of ergothioneine across Mtb strains.

4.3 Discussion

The workflow I developed in these analyses allowed the identification of nsSNP-metabolite correlations using a small sample size ($n = 14$). The majority of previous studies investigating the impact of SNPs found in Mtb strains focus purely on genomics, using genome wide association studies (GWAS) or genome-scale models which require large sample sizes (usually in the thousands), with limited functional validation studies on specific findings as follow-up. While nsSNPs have previously been used to classify Mtb strains by lineage, the data presented here suggest that metabolites are also highly lineage specific. So far, only one study has attempted to integrate SNP and metabolite data which also used a limited sample size ($n = 18$). However, the analysis workflow in that study estimated their findings with the use of genome-scale models and flux balance analysis which are prediction-based methods and are complex to use for most researchers. Instead, the workflow designed in this thesis is without the use of complex models and, therefore, is more accessible, easily applied to genomic and metabolomic datasets, and transferable to any genomes and metabolomes datasets.

Through the workflow described here, several differences between Mtb lineages and spoligotypes were identified. The nsSNPs classified the Mtb strains by their lineage, reflecting the high conservation of genetics within Mtb lineages. Through functional pathway prediction, these differences were estimated to be due to nsSNPs within various functional pathways, although our analysis focussed on metabolic pathways generally. Strains within lineage 1 and 2 were estimated to have increased nsSNPs in the 2-methylcitrate cycle, which is required for Mtb to breakdown propionate and survival of the bacterium in mice (Eoh & Rhee, 2014; Muñoz-Elías & McKinney, 2005). Strains within lineage 4 presented with several differences compared to the reference H37Rv, despite being within the same lineage. These included cholesterol degradation and fatty acid synthesis and degradation pathways. Considering the importance of cholesterol and fatty acid metabolism in Mtb pathogenesis, these data suggest that key metabolic pathways may be significantly altered as a result of the genomic evolution of Mtb strains compared to the reference H37Rv, questioning how representative this strain is for future research.

In addition to functional pathway prediction, the biological interpretation of nsSNPs was also performed at the gene-level through the generation of nsSNP blocks and measuring the representation score of each gene within these blocks. nsSNP block rs1849 contains nsSNPs which are present in strains within lineage 1 and not in strains within lineages 2 and 4. At the gene-level, the

genes in this block with the highest representation score were PE17 (a representation score of 70.6) and PE1 (a representation score of 35.3). These genes encode for proteins which have been understudied and therefore have unknown function but are believed to be expressed at the cell membrane, suggesting membrane remodelling between lineages. Similarly, the nsSNP block rs6112 also contained PE genes as the genes with the highest representation score (PE19 with a score of 141.2 and PE18 with a score of 53.0). rs6112 contains nsSNPs which are present only in strains in lineage 2. Therefore, strains from both lineages 1 and 2 are predicted to have undergone significant membrane remodelling compared to strains in lineage 4. This may have implications for vaccine and drug development, which can be targeted to Mtb cell surface antigens, as lineage 4 contained the reference strain H37Rv which is typically used for vaccine and drug studies.

Significant differences were observed in the metabolite analysis. A total of 939 untargeted metabolites and 38 targeted metabolites were identified as significantly differentially abundant in Mtb strains compared to the reference H37Rv. Through correlation and integration analysis, previously unidentified correlations between SNP and metabolite datasets were also found. While several correlations between untargeted metabolites and nsSNP blocks were observed, their interpretation is limited due to the difficulty in characterising untargeted metabolites. However, the correlation of targeted metabolites with nsSNP blocks identified interesting results. Specifically, nsSNP block rs6112 was found to be highly negatively correlated with ergothioneine (-0.74). Ergothioneine is involved in many Mtb metabolic pathways, including cell wall metabolism, redox and bioenergetic homeostasis, virulence, and is essential for susceptibility towards anti-TB drugs including rifampicin and isoniazid (Saini et al., 2016).

Ergothioneine was also identified to be significantly negatively decreased in abundance in all Mtb strains within lineage 1, but not in strains within lineages 2 and 4, compared to the reference H37Rv (median log₂ fold change of -2.8, adjusted p-value of 4×10^{-8}). Upon investigation of ergothioneine-related genes which include nsSNPs within the nsSNP block rs6112, there were no genes directly involved in the ergothioneine biosynthesis pathway. Few studies have investigated the role of ergothioneine in Mtb. One study created Mtb mutants with disrupted, or abrogated, expression of *egtA*, *egtD*, and *mshA* genes (Saini et al., 2016). *EgtA* and *egtD* encode for enzymes involved in ergothioneine synthesis (Richard-Greenblatt et al., 2015). *MshA* encodes for mycothiol (MSH), which is a redox sensor produced by Mtb to protect against redox stress and TB drugs (Saini et al., 2016). Genes not directly involved in the ergothioneine biosynthesis pathway, but were altered in these

mutant Mtb strains, were identified, including *rpsJ*, *sucD*, *metE*, *pknH*, *vapC10*, *ppsE*, and *desA3* which are also found in the rs6112 nsSNP block and almost all have a representation score of above 1. These genes, or presumably others found in the same operon, may be influencing the change in ergothioneine abundance seen in the Mtb strains in this study. Further validation analysis targeting these genes in different Mtb strains should be performed and the abundance of ergothioneine measured.

Despite these findings, the approach used here does suffer from limitations and challenges. One of the main limitations of this analysis is the lack of standardised and curated databases for Mtb. This prevents many bioinformatic tools, which are often designed for use with human or mouse models, from being able to be applied to other less-studied organisms such as Mtb. For example, there are few functional pathway tools designed for Mtb specifically, with most publications using the classification pathways described in MycoBrowser (Kapopoulou et al., 2011) which are extremely broad and would not identify more specific differences across samples. These pathway groups include virulence, detoxification, adaption, information pathways, cell wall and cell processes, stable RNAs, insertion seqs and phages, PE/PPE, intermediary metabolism and respiration, unknown, regulatory proteins, conserved hypotheticals, lipid metabolism, and pseudogenes. These pathways were originally used for functional analysis but as they are so broad that no differences were detected between lineages or strains (data not shown). The only alternative database available was the BioCyc database which, although far more specific and can be tailored to the level of detail required, focusses on metabolic pathways. As a result, the functional pathway prediction results are highly biased towards metabolic pathways and may miss differences in other pathways of biological importance. In addition, interpretation of the genes affected by the nsSNPs found in this dataset is also impeded by the lack of characterised genes in the Mtb genome. A significant proportion of genes with high representation scores are completely uncharacterised (mainly those which are designated with Rv numbers) and so despite being highly likely to be driving differences across lineages, it is impossible to interpret these data in a meaningful way.

In addition, the limited sample size used and the large number of nsSNPs presented a challenge in interpreting the importance of these nsSNPs before and after integration analysis. As several nsSNPs were found to be present or absent in the same Mtb strains, it is difficult to attribute metabolic differences due to a single nsSNP with confidence that it was not caused by another. To provide greater weight to the associations made, the nsSNP blocks and representation score was derived. This

allows for the gene-level interpretation of nsSNPs and therefore genes with higher representation scores, and more nsSNPs present, are more likely to be causing a downstream effect. Nevertheless, despite calculation of representation scores, there was still a large number of genes with high scores found in certain nsSNP blocks. Overall, this approach can be used to reduce the number of nsSNPs and genes which can be chosen for further study.

Overall, the field of Mtb lineages and evolution often describes genomic differences between lineages without investigating downstream impact, particularly in metabolomic pathways which are significantly important for Mtb virulence and infection. Here, a novel and simple approach to understanding, integrating, and interpreting genomic and metabolomic data from 3 Mtb lineages has been developed. This approach has identified previously unknown differences between these lineages and has implications for the use of H37Rv as a reference strain for Mtb research. Specifically, ergothioneine and the role of redox metabolism has been shown to be a promising target from these data for further research into the genomic consequences of differences observed in Mtb lineages.

Chapter 5: Analysing host-pathogen interactions in tuberculosis infection through applying dual RNA-seq to the whole blood infection assay

5.1 Background

Determining the contributions of both the host and pathogen components of host-pathogen interactions, in the same experiment, is difficult. Most infection models focus on only one aspect of infection, either the host or the pathogen. With the advancement of new technologies, approaches, and computational analyses, complex infection models can now be studied. This has paved the way for multi-cellular models such as *in vitro* WBAs, presented in the Introduction, to be further mined for biological insight further than has previously been attempted.

Some infections, such as TB, are extremely complex, with multiple cells involved in protective responses against infection. In addition, there are still multiple diseases for which protection against infection is still not fully characterised. TB is one of these diseases, where research has identified several key players in infection (primarily the cellular immune response implicating multiple T cell subsets, but other cells such as macrophages are known to be critical for its control; (Sia & Rengarajan, 2019)). However, it is unclear why certain immune responses do not offer protection and why some individuals are able to contain Mtb and other cannot. These questions are difficult to investigate as they require analysis of global responses involving multiple cell types in a system which reflects *in vivo* infection. In addition, infection models which measure both pathogen and host activity in parallel may reveal which pathogen pathways stimulate these protective host immune responses and which host responses exacerbate infection.

Unlike cell infection models, *in vitro* WBAs offer the opportunity to study several cell types in one system. This WBA model is significantly more representative of true infection than cellular models. While the blood system may not be the primary site of infection for many infectious diseases, transcriptomics analyses have shown that whole blood can still provide useful information on gene

expression responses of immune cells, at a minimum (Gliddon et al., 2018). Moreover, the WBA models can also be used to measure changes over time. This can also provide an insight into both the initial innate immune response of the host to infection and the later adaptive immune response. However, it should be noted that several aspects of the WBA have not been characterised. For example, none of the previously published WBA studies have measured cell survival over longer time-courses (>24 hours), and this should be considered before undertaking a time-course experiment at later time-points.

Previous studies using differing WBA methodology have provided insight into different host immune responses towards BCG and Mtb infection (Baguma, et al., 2017; Bok et al., 2021; Newton et al., 2011; Silva et al., 2013). A recent WBA study has also investigated the global host transcriptomics response, over time, in both BCG *lux* and Mtb infected whole blood (von Both et al., 2018). The authors of this study identified a significant downregulation of early immune responses upon infection with Mtb, which they hypothesised may be an evolutionary response of the host against excessive inflammation during infection. The authors also compared the transcriptomic profiles of the WBA infected with BCG *lux* compared to Mtb *lux* and found that 88% of transcripts showed the same direction of expression. Whilst this is an interesting hypothesis, this study and others, have not investigated is the bacterial response towards these host immune cells throughout the WBA time-course. If information from both sides can be measured, then the instigators of some of these host immune responses may be identified.

Dual RNA-seq is a new approach which can capture gene expression changes from more than one entity within a sample and is typically used to investigate host and pathogen responses. Dual RNA-seq approaches often increase the proportion of bacterial RNA reads to allow for sufficiently powered differential gene expression analysis of both the bacteria and host genes in a single sample (Westermann et al., 2012). Dual RNA-seq is therefore well placed to investigate the bacterial and host responses in complex models in WBAs. This approach has been used most often in single cell line models and has most recently been used to model host-pathogen interactions in a mouse model of TB (Pisu et al., 2020) and on sputum samples from TB patients from South Africa (Lai et al., 2021).

The main aim of this chapter is to combine the WBA using BCG *lux* with dual RNA-seq to investigate host-pathogen interactions over time in TB. This approach will identify time-dependent changes in the

bacterial transcriptome which may reflect changes in the host transcriptome, and vice versa. As this dual approach has not previously been performed in whole blood to characterise the immune response to Mtb infection, novel complex interactions over time may be identified. However, the complexity of undertaking this experiment brought about several challenges, including developing and optimising an appropriate RNA extraction protocol for BCG *lux* infected whole blood, increasing the proportion of bacterial RNA reads compared to host RNA reads in each sample, and ensuring that host cell death due to *in vitro* culture was not the predominant transcriptomic signal.

5.2 Results

5.2.1 Optimisation of BCG *lux* growth

Defining the optimal growth media for BCG lux

Due to the safety requirements of a Biosafety Level (BSL) 3 laboratory, BCG *lux* (BSL-2 organism) was used as a surrogate strain for Mtb work. BCG *lux* can be grown in/on different media depending on the nature of the experiment. As such, different liquid and solid media were tested for BCG *lux* growth. Both 7H9 albumin dextrose catalase (ADC) media and 7H9 DGNT media were tested for their growth of BCG *lux* over 72 hours, or three doubling times (**Figure 5.1A**). 7H9 ADC media is traditionally used to grow mycobacteria but contains Tween 80 as a detergent to prevent mycobacterial clumping which can also be used as a carbon source in addition to glycerol. 7H9 DGNT media is used when growing mycobacteria for metabolomics experiments as it does not contain Tween 80, instead using tyloxapol as a non-ionic surfactant, and therefore limits the carbon sources used by the bacteria. BCG *lux* grew faster in 7H9 ADC media than 7H9 DGNT media after 24 hours, regardless of shaking or static conditions and previous stock. As a result, 7H9 ADC media was used to grow BCG *lux* in further experiments.

Growth ratio calculation of BCG lux growth across the WBA

To measure mycobacterial growth during the WBA, i.e. in RLU, the ratio of growth between RLU and CFU measurements must be calculated. To determine this, BCG *lux* was grown in 7H9 ADC media for 96 hours and RLU and CFU measurements were recorded every 24 hours (**Figure 5.1B**). RLU and CFU measurements demonstrated a doubling time of 24 hours, which is expected for slow-growing mycobacteria such as BCG *lux* (Beste et al., 2005). The RLU:CFU ratio was between 6:1 and 10:1 for each time-point after 0 hours (**Table 5.1**).

Table 5.1: RLU/CFU ratios of BCG lux growth curve. D1 = donor 1; D2 = donor 2.

Hrs	D1 RLU	D2 RLU	D1 CFU	D2 CFU	RLU:CFU ratio
0	4.25x10 ⁷	4.29x10 ⁷	1.84x10 ⁶	3.15x10 ⁵	39.6
24	4.21x10 ⁷	4.75x10 ⁷	8.15x10 ⁶	6.70x10 ⁶	6.0
48	1.04x10 ⁸	1.16x10 ⁸	1.25x10 ⁷	9.55x10 ⁶	10.0
72	2.22x10 ⁸	2.35x10 ⁸	2.60x10 ⁷	2.70x10 ⁷	8.6
96	5.28x10 ⁸	6.24x10 ⁸	1.02x10 ⁸	4.95x10 ⁷	7.6

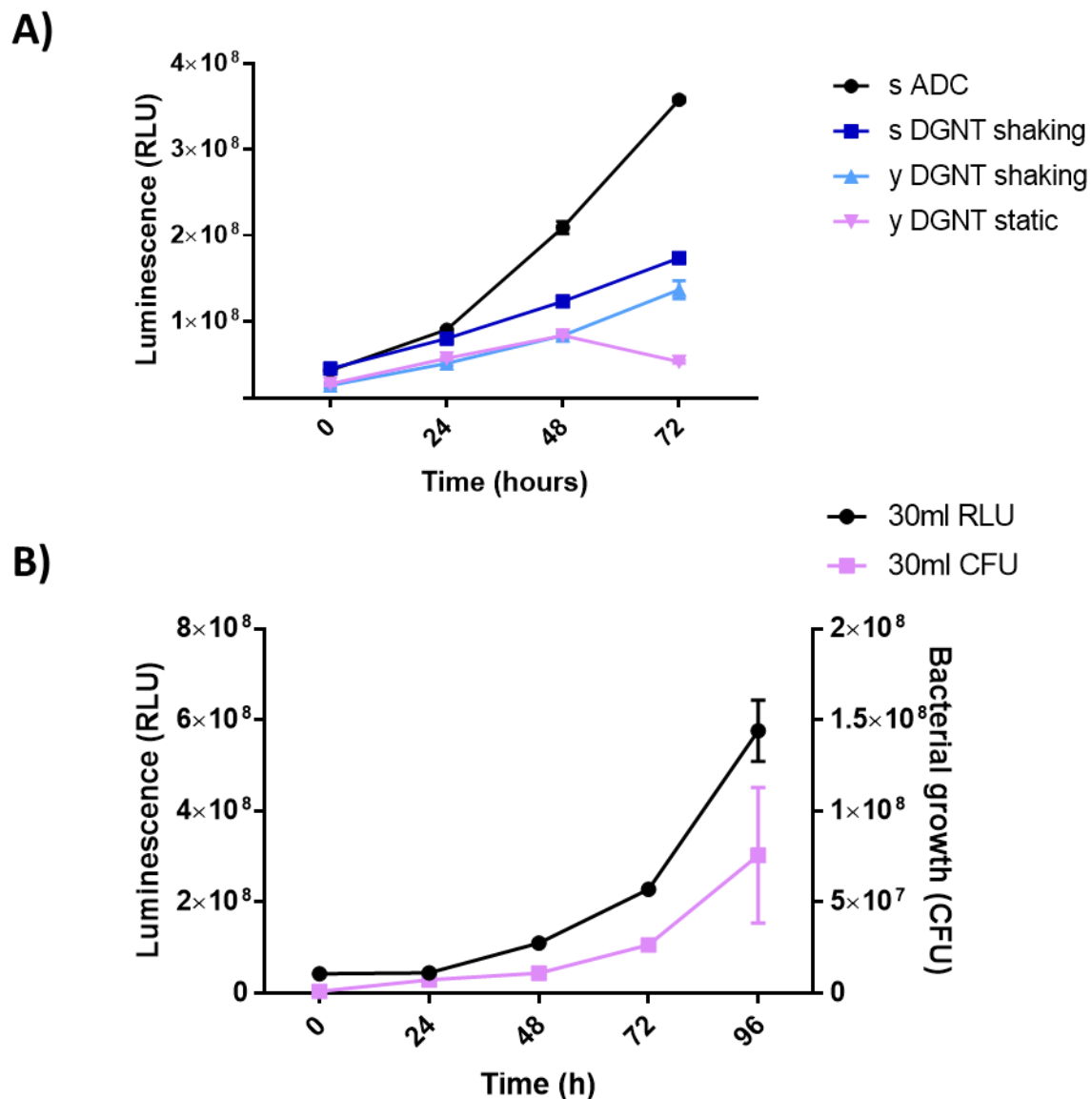


Figure 5.1: BCG lux growth in different growth media. A) BCG lux growth in 7H9 ADC media and 7H9 DGNT media in shaking and static conditions over 72 hours. S and Y indicate the initials of different BCG lux stocks used for growth. N = 1. **B)** BCG lux growth in 7H9 ADC media under shaking conditions measured through RLU (left-hand y axis) and CFU (right-hand y axis) over 96 hours. N = 2.

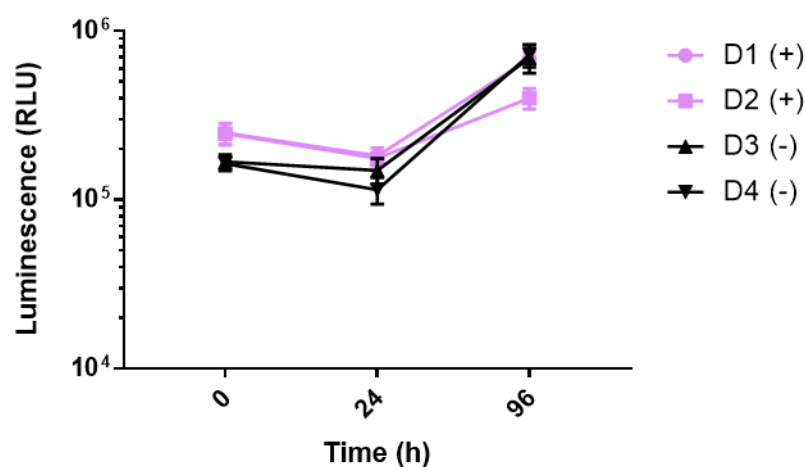
5.2.3 Validation of BCG *lux* growth in whole blood and cytokine analysis

Effect of BCG status on BCG lux infection in the WBA

While no significant IFN γ response was observed upon stimulation with PPD in BCG vaccinated donors, I hypothesised that BCG vaccination status could still affect other pathways, such as those involved in BCG *lux* growth, or different host immune responses to BCG *lux*. To test this, WBAs were performed using peripheral whole blood from BCG vaccinated and unvaccinated donors. One ml of whole blood from 8 donors was infected with BCG *lux* at an MOI of 1 and RLU was measured at 0, 24, and 96 hours (**Figure 5.3**). These time-points were chosen to reflect both early and late immune responses, with 96 hours being the limit of the WBA before host cell death from cell culture occurs. No difference was observed in BCG *lux* growth in BCG vaccinated and unvaccinated donors at any time-point. BCG *lux* growth decreased at 24 hours compared to 0 hours which is consistent with previous assays in our laboratory (data not shown) and is likely to reflect the bacteria establishing infection before replication. BCG *lux* growth increased after 24 hours, and above the initial RLU observed at 0 hours, indicating replication of the bacteria.

Luminescence ratios describe the quality of BCG *lux* growth at specific time-points compared to at 0 hours. These ratios were calculated for 24 hours and 96 hours (**Figure 5.3B**). Luminescence ratios did not differ between BCG vaccinated and unvaccinated donors at either time-point. However, the ratio remained consistently small at 24 hours. This was in contrast to the ratios observed at 96 hours which were scattered between 0.78 and 9.42 but on average increased compared to 24 hrs, indicating an increase in growth rate at late time-points.

A)



B)

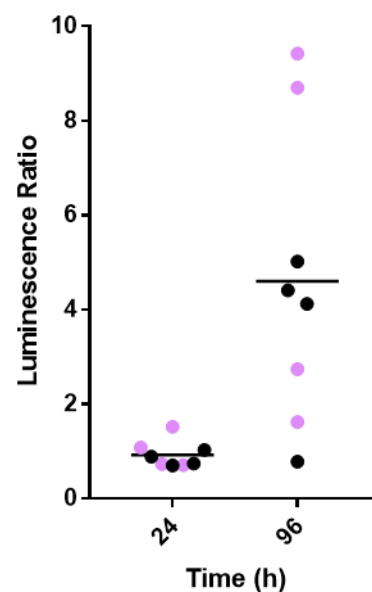


Figure 5.3: WBA using BCG vaccinated and unvaccinated donors. A) RLU measurements of BCG lux growth in BCG vaccinated (+) and unvaccinated (-) donor whole blood over 96 hours. **B)** Luminescence ratios for WBA measurements at 0 and 24 hrs and at 0 and 96 hrs. BCG vaccinated donors are indicated in pink and with a (+), while BCG unvaccinated donors are indicated in black and with a (-). D1 – 4 = donors 1 – 4. Error bars indicate the standard deviation from 3 technical replicates of luminescence readings. N = 2 independent experiments.

Optimisation of volume and MOI used in WBA for infection with BCG lux

The WBA was optimised in several ways to prepare samples for dual RNA-seq experiments. To obtain sufficient host and bacterial RNA quantities, the volumes of blood used in the WBA were increased. The WBA was performed with 1 ml blood volumes (which is the standard volume previously used (S. Newton et al., 2011)) and was compared to 2 ml blood volumes with the standard MOI of 1:1 (**Figure 5.4**). BCG *lux* growth over 96 hours was compared in the two different blood volumes. No difference in BCG *lux* growth was observed at any time-point between these two volumes.

To perform the dual RNA-seq experiment using the BCG *lux* WBA, the amount of bacterial RNA needed to be optimised and increased so as not to be diluted by host RNA and therefore not sequenced (Westermann et al., 2017). One method to increase the levels of bacterial RNA is to increase the MOI used during infection. The MOI used in the WBA was therefore increased from 1:1 to 10:1 (mycobacteria to monocyte cells per ml of whole blood) and BCG *lux* growth was measured over 96 hours in 2 ml blood volumes (**Figure 5.4**). The RLU of BCG *lux* infected in whole blood at 10:1 MOI was 10 times higher than the RLU measurements of BCG *lux* infected in whole blood at 1:1 MOI, as expected. In addition, a similar reduction in BCG *lux* RLU at 24 hours compared to 96 hrs was observed in all donors in all conditions (1 ml and 2 ml blood volumes and 1:1 and 10:1 MOI).

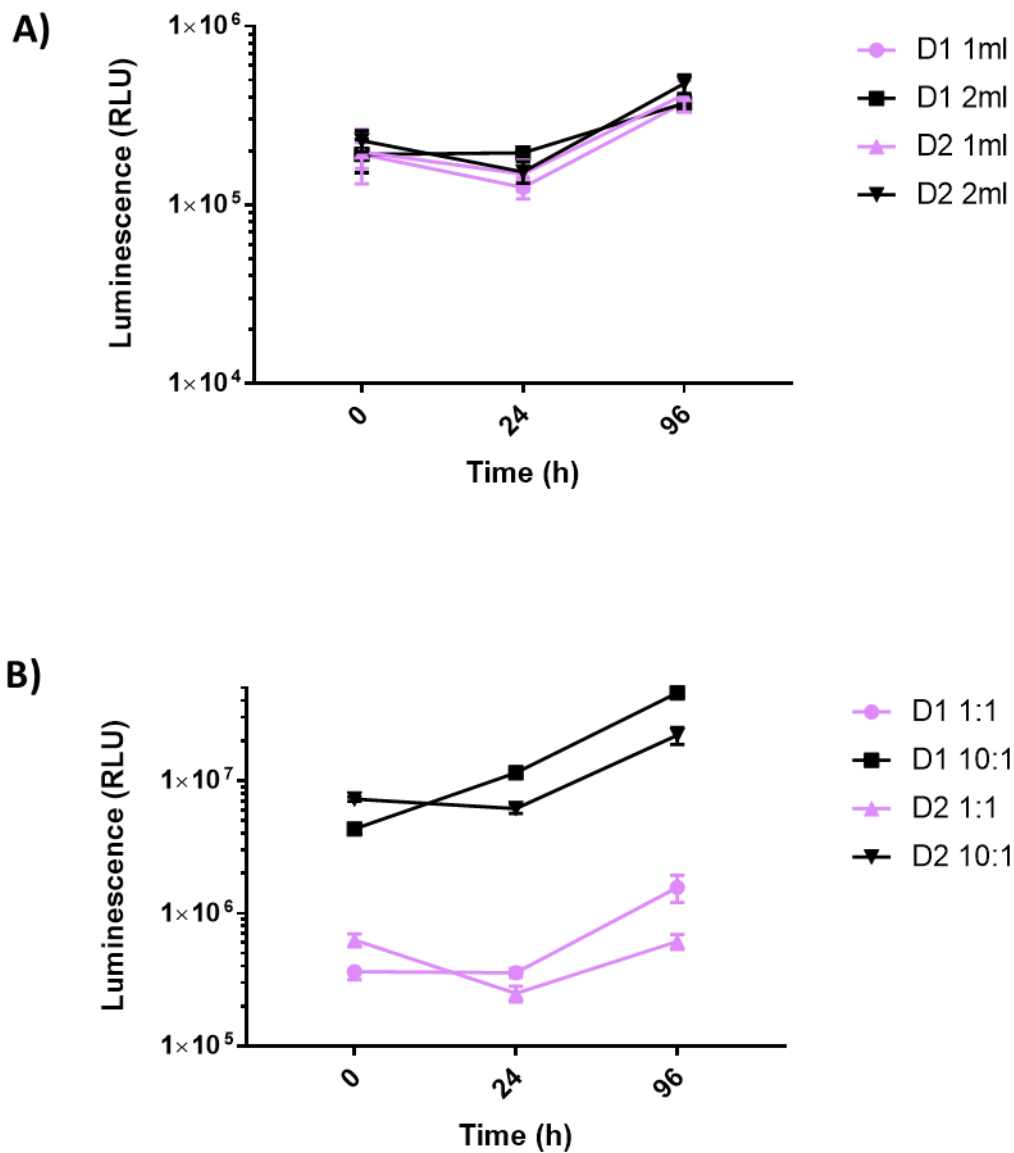


Figure 5.4: The effect of different whole blood volumes and MOI on BCG lux growth in the WBA over 96 hours of infection. A) The effect of changing the sample volume from 1 ml to 2 ml of whole blood on BCG lux growth with a MOI of 1:1 in the WBA over 96 hours. 1 ml samples are indicated in pink; 2 ml samples are indicated in black. **B)** The effect of changing the sample MOI from 1:1 to 10:1 on BCG lux growth in 2 ml samples in the WBA over 96 hours. Samples infected with a MOI of 1:1 are indicated in pink; samples infected with a MOI of 10:1 are indicated in black. D1 – 2 = donors 1 – 2. Error bars indicate the standard deviation from 3 technical replicates of luminescence readings. Figures shown are representative for 2 biological replicates.

Luminescence ratios were again calculated for donor blood samples using different blood volumes and MOIs to ensure similar luminescence trends were observed despite differences in initial RLU values. No difference in luminescence ratios was observed in 1 ml or 2 ml blood volumes (**Figure 5.6A**) or in those infected with an MOI of 1:1 or 10:1 (**Figure 5.6B**). This suggests that changing the blood volume to 2 ml, and the MOI to 10:1, did not change the growth trend of *BCG lux*. Therefore, further experiments could be conducted using these conditions in the WBA. The luminescence ratios were lower in all samples at 24 hours but increased at 96 hours as previously observed.

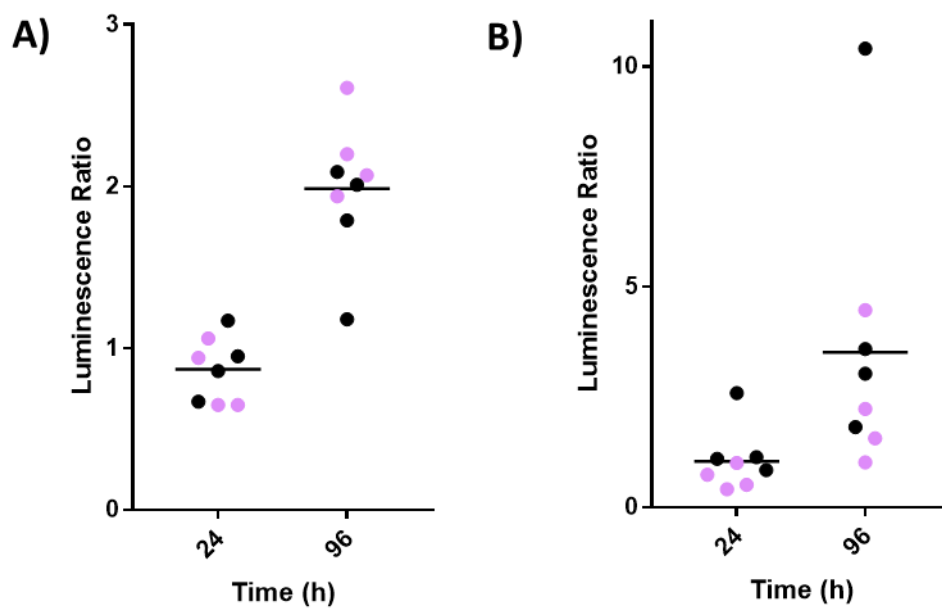


Figure 5.6: The effect of different blood volume and MOI conditions on luminescence ratios from the *BCG lux* WBA over 96 hours. A) The effect of changing the sample volume from 1 ml to 2 ml of whole blood on *BCG lux* luminescence ratios in the WBA over 96 hours. 1 ml samples are indicated in pink; 2 ml samples are indicated in black. *N* = 4. **B)** The effect of changing the sample MOI from 1:1 to 10:1 on luminescence ratios in the WBA over 96 hours. Samples infected with a MOI of 1:1 are indicated in pink; samples infected with a MOI of 10:1 are indicated in black. Error bars indicate the mean value. *N* = 4 independent experiments.

IFN γ and TNF α response to BCG lux infection of whole blood

The WBA can be used to measure features of infection, other than bacterial growth rate, such as host immune responses. To ensure host immune cells were responding to BCG *lux* infection, and as a proxy for cellular activity and therefore a measurement of live cells in the assay, IFN γ and TNF α cytokine levels were measured over 96 hours. This was performed using both 1 ml and 2 ml blood volumes and MOIs of 1:1 and 10:1 to determine if changing the blood volumes or MOIs of samples would affect the host cytokine response to infection.

IFN γ levels increased in some donors over time, but not all (**Figure 5.7A**). Very few 1 ml blood volume samples resulted in an increased IFN γ over time in response to BCG *lux* infection, however 3 samples, which were 2 ml in volume, did demonstrate an increase in this cytokine at 24 and 96 hours, although 1 was a modest increase. In contrast, almost all samples regardless of blood volume, showed a transient increase in TNF α levels at 24 hours before a reduction at 96 hours (**Figure 5.7B**). Samples with 2 ml blood volumes showed slight higher TNF α levels compared to those with 1 ml volumes.

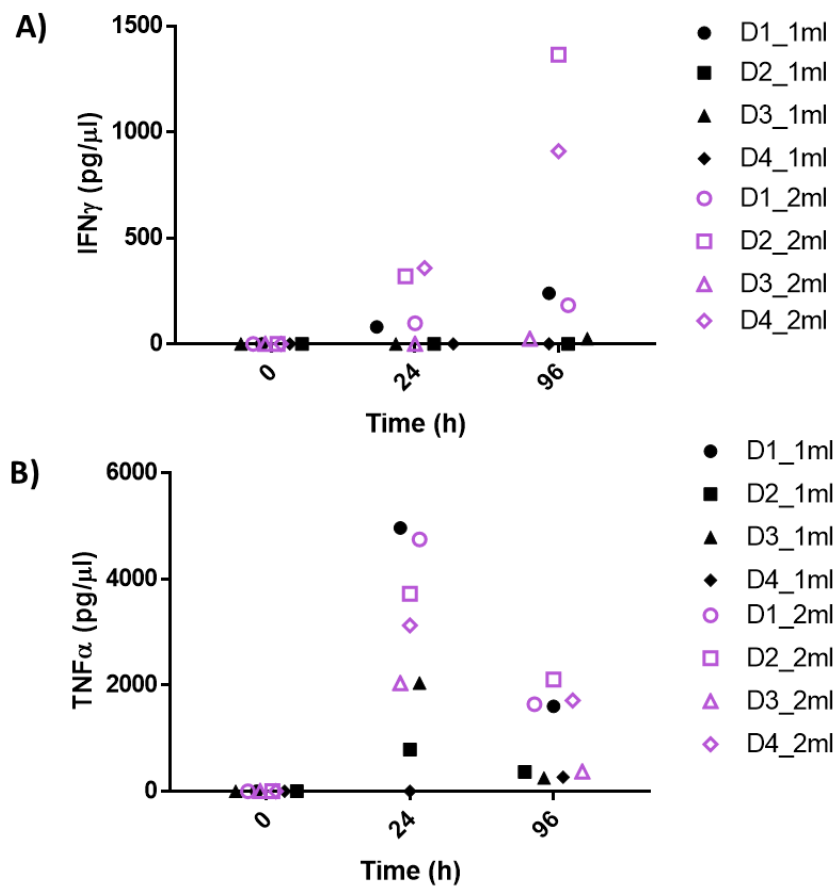


Figure 5.7: The effect of different sample volumes on the cytokine profile of whole blood infected with BCG lux at a MOI of 1:1 over 96 hours. A) IFN γ production from 1 ml or 2 ml samples of whole blood infected with BCG lux at a MOI of 1:1 over 96 hours. 1 ml samples are indicated in black; 2 ml samples are indicated in pink; different donors are indicated by differently shaped points. **B)** TNF α production from 1 ml or 2 ml samples of whole blood infected with BCG lux at a MOI of 1:1 over 96 hours. 1 ml samples are indicated in black; 2 ml samples are indicated in pink; different donors are indicated by differently shaped points. Plots represent 2 independent experiments consisting of 4 donors each. D1 – 4 = donors 1 – 4.

There was little difference in IFN γ production in samples infected with BCG lux at an MOI of 1:1 or 10:1, except for one donor where increased IFN γ levels were measured at both 24 and 96-hour time-points with both MOIs (**Figure 5.8A**). The sample which was infected with an MOI of 10:1 did show slightly higher IFN γ levels than the sample from the same donor with an MOI of 1:1. TNF α levels transiently increased in almost all samples regardless of MOI infection at 24 hours before decreasing at 96 hours (**Figure 5.8B**). Samples infected with an MOI of 10:1 showed a higher TNF α level at 24 hours and 96 hours than those infected with an MOI of 1:1; with the exception of one sample infected with a 1:1 MOI consistent with a 10:1 MOI.

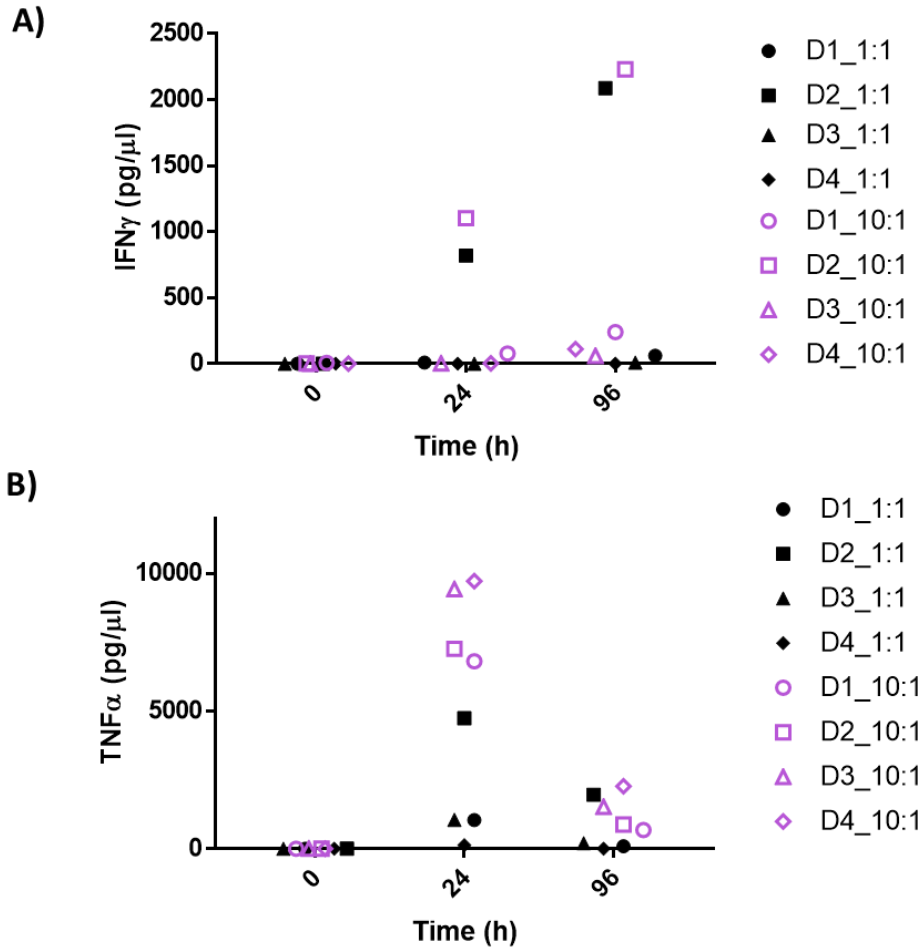


Figure 5.8: The effect of different MOIs on the cytokine profile of 2 ml of whole blood infected with BCG lux over 96 hours. A) IFN γ production from whole blood samples infected with a MOI of 1:1 and 10:1 of BCG lux over 96 hours. Samples infected with a MOI of 1:1 are indicated in black; Samples infected with a MOI of 10:1 are indicated in pink; different donors are indicated by differently shaped points. B) TNF α production from whole blood samples infected with a MOI of 1:1 and 10:1 of BCG lux over 96 hours. Samples infected with a MOI of 1:1 are indicated in black; Samples infected with a MOI of 10:1 are indicated in pink; different donors are indicated by differently shaped points. Plots represent 2 independent experiments consisting of 4 donors each. D1 – 4 = donors 1 – 4.

5.2.4 Optimisation of flow cytometry analysis of whole blood infected with BCG *lux*

Antibody titration optimisation of WBA samples

Measuring host cell populations and cell viability in the WBA is essential to deconvolute the transcriptomic signals from the dual RNA-seq experiment and to ensure appropriate time-points are chosen for the WBA to prevent measuring excessive cell death at later time-points. Therefore, flow cytometry was performed to measure cell populations and viability. Prior to performing flow cytometry experiments on WBA samples, an appropriate fluorochrome antibody panel had to be established and concentrations of each antibody optimised. This optimisation approach also reduced the reagent usage and cost of each experiment.

Antibody titration experiments were performed to identify the lowest possible antibody concentration which maintained sufficient fluorophore signal to measure the target cell population. Different cell populations were gated using the procedure described in **Chapter 2.4.6**. To identify the appropriate antibody concentration required for each cell surface marker the separation index (SI) for each antibody was calculated (**Table 5.2**). Generally, the concentrations with the highest SI are the optimal concentrations, however for some antibodies with a large SI range (e.g. CD15), a lower but still sufficient SI can be used to select an optimal antibody concentration for future experiments. These were evaluated individually based on cell population separation observed in the gating strategies (**Figure 5.9**).

Table 5.2: Separation indices for flow cytometry antibodies. Bold indicates the antibody concentration used in future experiments based on the SI value and gating plots.

Antibody target	Antibody dilution	SI	Dilution to be used
CD3	1 in 28	17.7	1 in 50
CD3	1 in 57	27.2	
CD3	1 in 114	25.3	
CD3	1 in 228	22.1	
CD15	1 in 28	151	1 in 50
CD15	1 in 57	137	
CD15	1 in 114	111	
CD15	1 in 228	72.9	
CD14	1 in 28	54.6	1 in 100
CD14	1 in 57	43.5	
CD14	1 in 114	38.4	
CD14	1 in 228	8.48	
CD56	1 in 28	2.11	1 in 100
CD56	1 in 57	5.51	
CD56	1 in 114	6.36	
CD56	1 in 228	8.08	
CD11b	1 in 28	9.36	1 in 20
CD11b	1 in 57	8.78	
CD11b	1 in 114	8.22	
CD11b	1 in 228	4.98	
CD19	1 in 28	32.5	1 in 100
CD19	1 in 57	28.2	
CD19	1 in 114	25.7	
CD19	1 in 228	19.4	
HLA-DR	1 in 28	4.24	1 in 20
HLA-DR	1 in 57	2.24	
HLA-DR	1 in 114	4.35	
HLA-DR	1 in 228	5.84	
LD	1 in 100	149	1 in 100
LD	1 in 500	5.4	
LD	1 in 1000	3.22	
LD	1 in 2000	3	

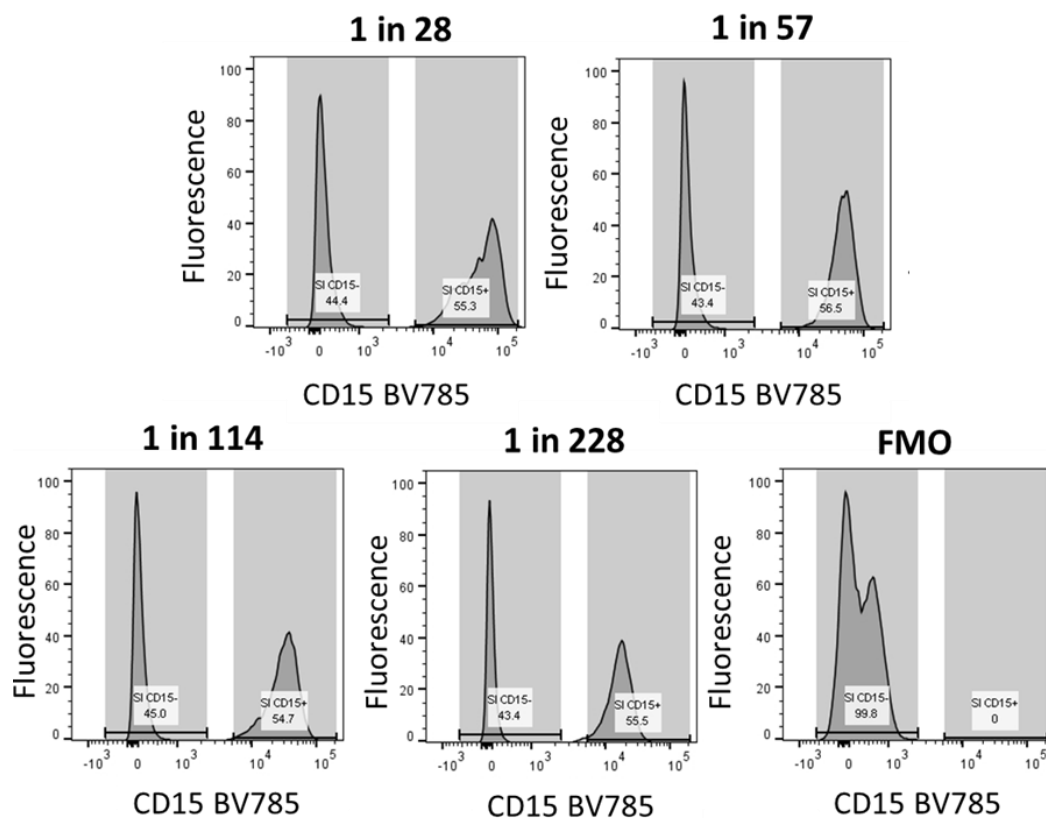


Figure 5.9: Example histograms of the antibody titration of the CD15 fluorophore in whole blood. Different titrations are indicated for each plot. The CD15+ cell populations are indicated on the left of each plot. FMO = fluorescence minus one.

Host cell survival in whole blood samples with BCG lux infection at different MOIs

To investigate whether using a MOI of 10:1 instead of 1:1 contributed to higher host immune cell death during the WBA, flow cytometry analysis was carried out. Survival of three major cell populations, namely granulocytes, lymphocytes, and monocytes, was measured during a BCG *lux* WBA over the course of 0, 24, and 48 hours. These time-points were chosen to reflect early (0 – 24 hours) and late (24 – 48 hours) infection. It was suggested that culturing whole blood for 96 hours may considerably change the cellular environment of the whole blood compared to the *in vivo* environment, and therefore 96 hours was not included in this, or future, analyses.

Granulocytes comprise of basophils, eosinophils, neutrophils, and mast cells which generally have short life cycles *in vivo*, e.g. neutrophils often survive for around 24 hours ((Akgul et al., 2001), **Figure 1.6**). In addition, as neutrophils have been shown to become infected with Mtb during infection (Eum et al., 2010) they may therefore undergo apoptosis or necrosis in response to infection. As expected, the percentage of live granulocytes decreased over time, regardless of infection, although there was a significantly larger decrease in infected samples (**Figure 5.10A**). This resulted in a decrease from 97.8% live cells to an average of 61.7% of cells being live in blood samples infected with a BCG *lux* MOI of 1:1. In blood samples infected with the MOI of 10:1, there was a greater decrease, with an average 23.6% live cells at 48 hours. In addition, the percentage of total granulocytes also decreased over time, regardless of infection (**Figure 5.10B**). In infected samples, this decrease in live cells resulted in extremely few granulocytes being present at the 24- and 48-hour time-points, particularly for samples infected with a MOI of 10:1 where there was almost no granulocyte population present (**Figure 5.10B, Figure 5.10C**).

Lymphocytes comprise of NK cells, T cells, and B cells, of which T and B cells survive *in vivo* for several days, if not weeks ((Baliu-Piqué et al., 2018; Sun et al., 2011), **Figure 1.6**). The percentage of live lymphocytes remained high (average 77.3% at 0 hours and 94.2% and 95.0% at 24- and 48-hour time-points, respectively) throughout the time-course and in both uninfected and infected samples (**Figure 5.11A**). However, the percentage of total lymphocytes increased during the WBA time-course, with a higher increase in samples infected with a MOI of 1:1 and again a higher increase still in blood samples infected with a MOI of 10:1 (**Figure 5.11B, 5.11C**).

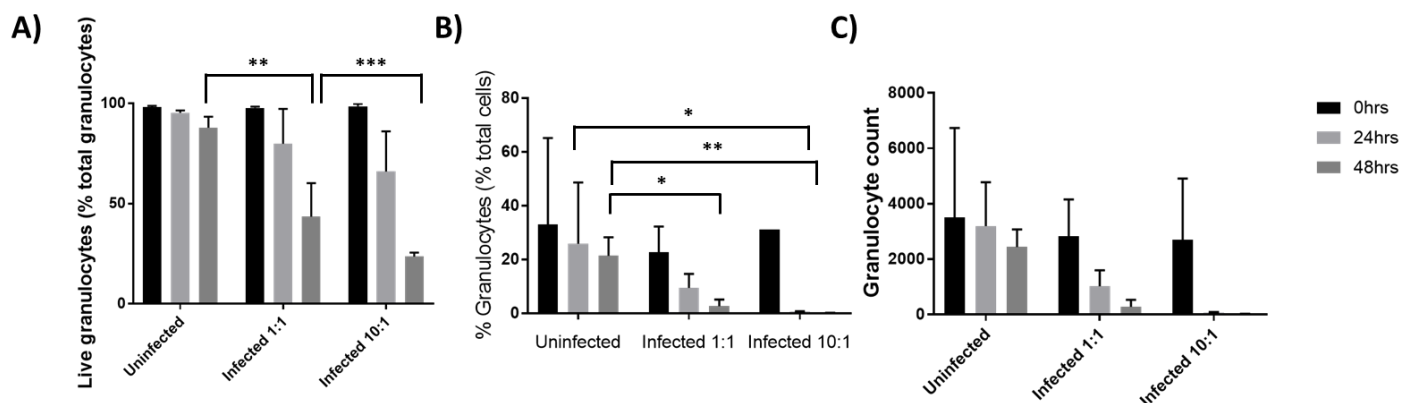


Figure 5.10: Granulocyte cell population analysis of whole blood upon infection of BCG lux at different MOIs over 48 hours. **A)** Live granulocytes as a percentage of total granulocytes in uninfected samples, samples infected at a MOI of 1:1, and samples infected at a MOI of 10:1 whole blood. **B)** Percentage of granulocytes from total cells in uninfected samples, samples infected at a MOI of 1:1, and samples infected at a MOI of 10:1 whole blood. **C)** Granulocyte count in uninfected samples, samples infected at a MOI of 1:1, and samples infected at a MOI of 10:1 whole blood. $N = 2$. * = p -value < 0.05; ** = p -value < 0.01; *** = p -value < 0.001.

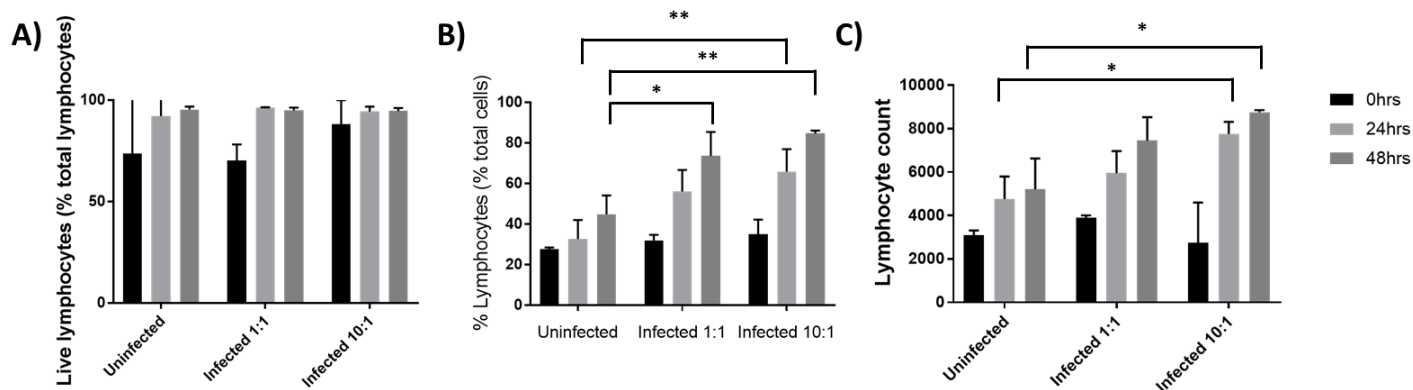


Figure 5.11: Lymphocyte cell population analysis of whole blood upon infection of BCG lux at different MOIs over 48 hours. **A)** Live lymphocytes as a percentage of total granulocytes in uninfected samples, samples infected at a MOI of 1:1, and samples infected at a MOI of 10:1 whole blood. **B)** Percentage of lymphocytes from total cells in uninfected samples, samples infected at a MOI of 1:1, and samples infected at a MOI of 10:1 whole blood. **C)** Lymphocyte count in uninfected samples, samples infected at a MOI of 1:1, and samples infected at a MOI of 10:1 whole blood. $N = 2$. * = p -value < 0.05; ** = p -value < 0.01.

Monocytes have been implicated in both control of Mtb infection through production of reactive nitrogen species but may also be detrimental to disease through harbouring the bacteria (Sia & Rengarajan, 2019). In uninfected samples and samples infected with a MOI of 1:1, the percentage of live monocytes increased slightly between 0 hours and 48 hours, but this population remained consistent at an average of 61% in samples infected with a MOI of 10:1 (**Figure 5.12A**). However, at 24 hours post-infection, the live monocyte population was on average smaller than at the other time-points in infected samples at an average of 31.8%. This is difficult to interpret at this time-point as there is a large deviation in live monocyte percentage, as observed by the large error bars. The total number of monocytes differed for each sample and did not follow the same trend as for live monocytes, although their numbers were much lower (<7.4% of all cells) than for granulocytes and lymphocytes, as expected from the literature ((Autissier et al., 2010), **Figure 5.12B**). The uninfected monocyte population showed a small transient increase at 24 hours but remained consistent across other time-points. Samples infected with a MOI of 1:1 showed a significant expansion of monocytes at 24 hours compared to 0 hours (doubling from an average 3.2% to 6.3%), but this decreased to low levels at 48 hours (an average of 1.6%). Samples infected with a MOI of 10:1 showed a gradual decrease over all time-points, starting with an average of 4.6% at 0 hours which was higher than the other samples, lowering to an average of 1.6% at 24 hours, to less than 0.9% of cells at 48 hours. This decrease likely reflects an increase in monocyte cell death due to infection in both samples infected with a MOI of 1:1 and 10:1, although this was more extreme for samples infected with a MOI of 10:1. The trend in total monocyte percentage was also observed in total monocyte counts, with cell counts in samples infected with a MOI of 10:1 being an average of 193 at 24 hours and an average of 89 at 48 hours (**Figure 5.12C**).

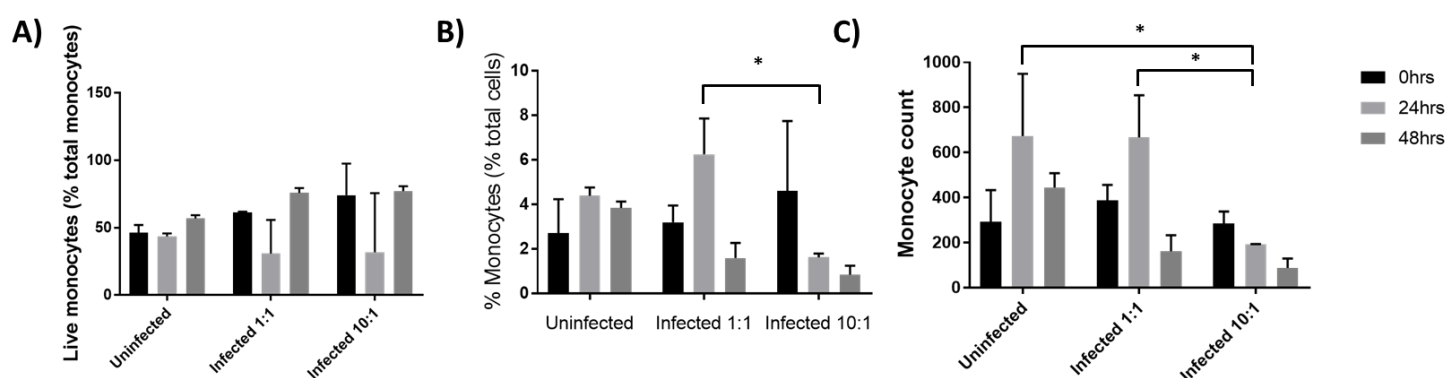


Figure 5.12: Monocyte cell population analysis of whole blood upon infection of BCG lux at different MOIs over 48 hours. A) Live monocytes as a percentage of total granulocytes in uninfected samples, samples infected at a MOI of 1:1, and samples infected at a MOI of 10:1 whole blood. **B)** Percentage of monocytes from total cells in uninfected samples, samples infected at a MOI of 1:1, and samples infected at a MOI of 10:1 whole blood. **C)** Monocyte count in uninfected samples, samples infected at a MOI of 1:1, and samples infected at a MOI of 10:1 whole blood. $N = 2$. * = p -value < 0.05.

Due to the significant increase in cell death in granulocytes and monocyte populations, to the point where the number of cells was extremely low, it was determined that a MOI of 10:1 caused a level of cell death which is unlikely to be representative of *in vivo* infection. Furthermore, in an *in vivo* model, these cells would be replaced by new cells within the system. It was therefore hypothesised that transcriptomics analysis of samples infected with a MOI of 10:1 would be over-diluted with a cell death transcriptional signal. Therefore, a MOI of 1:1 would be required for future experiments to obtain results similar to what is observed in *in vivo* infection.

5.2.5 Optimisation of RNA extraction procedures from whole blood infected with BCG *lux*

Prior to dual RNA-seq, the quality and quantity of RNA in samples must be evaluated so that they meet minimum requirements described by the sequencing facility chosen. For sequencing experiments at Oxford Genomics Centre (Oxford Genomics Centre, 2022) the requirements include a minimum RNA concentration of 2 µg, a 260/280 and 260/230 Nanodrop ratio of 2, and a RIN of 7 or above. The ratios provide an estimate of RNA purity in a sample, while the RIN provides an integrity score for the RNA within the sample, with 7 being conventionally used as a standard cut-off. These values can be adjusted depending on the experiment where certain concessions in quality or quantity can still provide adequate sequencing results.

Baseline RNA extraction measurements from WBA

For RNA extraction, a modified TRIzol/miRNA extraction kit was used on uninfected WBA samples. Due to restrictions in blood volumes used in the protocol, each 2 ml WBA sample was aliquoted into 4 separate smaller samples before extraction. The method to extract RNA from these WBA samples was then optimised through several steps. First, the baseline RNA concentration was measured as seen in **Table 5.3**.

Table 5.3: An example of baseline RNA concentrations from 3 uninfected WBA samples at 0 hrs.
Samples A – D are aliquotes of the original sample (e.g. 1A – 1D are from sample 1).

Sample Code	Yield (ng/µl)	Yield (ng)	Calculated combined yield (µg)
1A	33.72	1045	4.22
1B	48.19	1494	
1C	28.03	869	
1D	26.07	808	
2A	22.28	691	2.91
2B	27.49	852	
2C	20.05	622	
2D	24.19	750	
3A	30.54	947	3.90
3B	29.33	909	
3C	34.12	1058	
3D	31.84	987	

The effect of DNase treatment on RNA quantity and quality

In addition to RNA quantity and quality, it is standard procedure to identify the quantity of DNA contamination in samples before undertaking RNA sequencing. The concentration of DNA was therefore measured in samples before and after DNase treatment to determine if the treatment reduced the DNA concentration in the sample without significantly lowering the RNA yield (**Table 5.4**). DNA concentration was relatively high in all samples (a minimum of 22% and a maximum of 46%) prior to DNase treatment but was significantly reduced after treatment (a minimum of 7% and a maximum of 28%). Additionally, RNA quality measurements were performed to identify if the additional DNase treatment steps affected RNA quality. The 260/280 ratio dropped in all samples, and the 260/230 ratio dropped in the majority of samples. It should be noted that the RNA quantity increased in certain samples, which may have been partially due to the way that RNA quantity was measured using the Nanodrop. This technology uses wavelengths of molecules to calculate the concentration of RNA, DNA, or proteins, and often misclassifies DNA concentration as RNA concentration if high levels of DNA contamination are present.

The effect of clean-up treatment on RNA quantity and quality

Additional clean-up kits can be used to improve RNA quality, however adding additional steps to any extraction protocol is likely to decrease RNA quality and quantity as the RNA molecules degrade quickly with time. A clean-up kit was used to determine if improvement of RNA quality and quantity from the kit itself would be greater than potential quality and quantity lost through further handling of samples (**Table 5.4**). Whilst improving the RNA quality, the RNA clean-up kit significantly reduced the RNA concentration and was therefore not included in further RNA extraction experiments.

Table 5.4: The effect of DNase and clean-up treatment steps on RNA quantity and quality in uninfected WBA samples. Samples A – D are aliquotes of the original sample.

	Sample ID	RNA yield (ng/μl)	260/280	260/230	RNA yield (ng)	DNA yield (ng)
Before DNase treatment	4A	5.26	1.89	0.25	147	68
	4B	8.62	2.54	0.38	241	72
	4C	13.47	1.72	0.25	377	84
	4D	14.69	1.35	0.34	411	94
After DNase treatment	4A	20.02	1.33	0.38	400	29
	4B	18.04	1.41	0.35	361	51
	4C	11.45	1.14	0.21	229	63
	4D	16.37	1.25	0.29	327	72
After clean-up treatment	4A	10.73	1.89	0.57	118	
	4B	11.29	1.7	0.53	124	
	4C	11.18	1.7	0.47	123	
	4D	10.79	2.37	0.55	119	

Measuring RNA integrity

Furthermore, RNA quality must be measured using the RIN value calculated using a Bioanalyzer. This provides an integrity score and a trace plot which can be used to identify problems with a sample such as RNA degradation. For Bioanalyzer measurements, samples must be pooled and concentrated to a suitable volume and concentration for the Bioanalyzer kit. Therefore, the 4 subsamples from the previous extraction steps were pooled into their original samples. RIN values were measured before and after DNase treatment (**Table 5.5**). RIN levels decreased in 3/5 samples, although this difference was marginal and values were still at acceptable levels (>7) with the exception of samples 6 and 9. RNA concentration decreased for all samples after DNase treatment as described previously, which was observed using both the Nanodrop and Bioanalyzer methods to measure concentration. Sample 5 was an exception, where the RNA concentration increased after DNase treatment when measured with the Bioanalyzer. Standard procedure for sequencing quality control checks prioritises Nanodrop measurement of RNA concentrations over Bioanalyzer measurement of RNA concentrations, therefore I used the Nanodrop values as the most accurate measurements.

Table 5.5: RIN measurements of uninfected WBA samples before and after DNase treatment.

		Sample ID	RNA (ng/μl)	yield 260/280	260/230	RNA (ng)	yield RIN	RNA yield (Bioanalyzer; ng)
Before treatment	DNase	5	61.25	2.1	0.61	3369	8.1	1045
		6	35.08	1.48	0.33	1929	7.5	1430
		7	25.05	1.99	0.48	1378	7.8	1540
		8	47.92	2.04	0.75	2636	8.8	2145
		9	39.1	1.47	0.23	2151	5.5	1100
After DNase treatment		5	55.73	1.84	0.57	2229	8.6	1720
		6	30.56	1.35	0.32	1070	5.9	595
		7	21.26	1.53	0.42	1084	7.8	561
		8	39.24	1.78	0.61	1491	8.7	950
		9	25.43	1.22	0.18	1017	5.3	720

The effect of bead-beating on RNA quantity and quality

Finally, an additional bead-beating step was introduced to the extraction protocol. This was required to mechanically break the thick mycobacterial cell envelope which is rich in lipids such as mycolic acids and therefore require mechanical breakage to release intracellular RNA (Oh et al., 2016). Therefore, a bead-beating step was introduced after adding TRIzol to all WBA samples, both infected and uninfected, and the RNA concentration and quality was measured (**Table 5.6**). RNA concentration and quality decreased with the addition of the bead-beating step, but the combined concentration was still sufficient for sequencing. While the decrease in quality was undesirable, this step is essential to enable sequencing of the mycobacterial RNA and RNA quality likely needs to be compromised to allow for adequate sequencing of host and bacterial RNA within a sample. As a result, this step was included in further experiments.

Table 5.6: The effect of bead-beating on the concentration and quality of RNA from infected WBA samples. Samples A – D are aliquotes of the original sample.

Sample ID	RNA yield (ng/μl)	260/280	260/230	RNA yield (ng)	Calculated combined yield (μg)
10A	8.26	1.63	0.15	264	1.54
10B	13.63	1.79	0.23	436	
10C	11.21	1.94	0.28	359	
10D	15.14	1.89	0.24	484	
11A	19.89	1.87	0.23	636	1.81
11B	10.55	2.06	0.24	338	
11C	13.44	1.52	0.4	430	
11D	12.71	1.71	0.32	407	
12A	19.06	1.59	0.22	610	2.54
12B	19.55	1.4	0.22	626	
12C	18.44	1.22	0.28	590	
12D	22.31	1.32	0.14	714	

After optimising the extraction procedure, the finalised methods can be found in **Chapter 2.4.7**. To summarise, samples were frozen at -80°C until extraction. Samples were then thawed and aliquoted into 8 subsamples for bead-beating (if applicable). Subsamples were then pooled into 4 subsamples and RNA was extracted. After extraction, subsamples were pooled into original samples and DNase treatment was performed. RNA was concentrated, and the q

uality and quantity was measured using both the Nanodrop and Bioanalyzer before sending to Oxford Genomics Centre for RNA sequencing.

5.2.6 Pilot dual RNA-seq analysis of whole blood infected with BCG *lux*

Pilot dual RNA-seq experiment design

Before performing a full dual RNA-seq experiment, and as concerns around mycobacterial RNA concentration in each sample were yet to be answered, a pilot dual RNA-seq experiment was performed and corresponding flow cytometry analysis carried out. Whole blood from 2 donors was infected with BCG *lux*, and uninfected whole blood was used as controls, and samples were incubated for 0-, 11-, and 24-hours. These time-points were chosen as they reflected the times chosen for the main dual RNA-seq experiment. Furthermore, later time-points would allow for increased BCG *lux* replication and hence increased mycobacterial RNA reads to be sequenced. Due to financial restraints, a limited number of samples were sent for sequencing (N = 6), representing duplicates samples from each time-point from the same donor. This was to answer several questions:

1. Was bead-beating a necessary step in the sample preparation protocol to extract sufficient mycobacterial RNA for bioinformatic analyses?
2. Would bead-beating damage host RNA which may be broken up during this step of the extraction protocol?
3. Would the low RIN values obtained for our samples affect sequencing quality?
4. Is the sequencing depth chosen sufficient to obtain enough sequencing reads for bioinformatic analyses?

While some of these questions could be addressed through RT-PCR, this method was attempted but due to time restrictions and troubleshooting issues it was decided to continue with a pilot sequencing experiment instead. To answer these questions, the pilot dual RNA-seq experiment was set up to include 6 samples, 3 of which included the bead-beating step in the extraction protocol, and 3 samples did not include a bead-beating step in the extraction protocol (**Figure 5.13**).

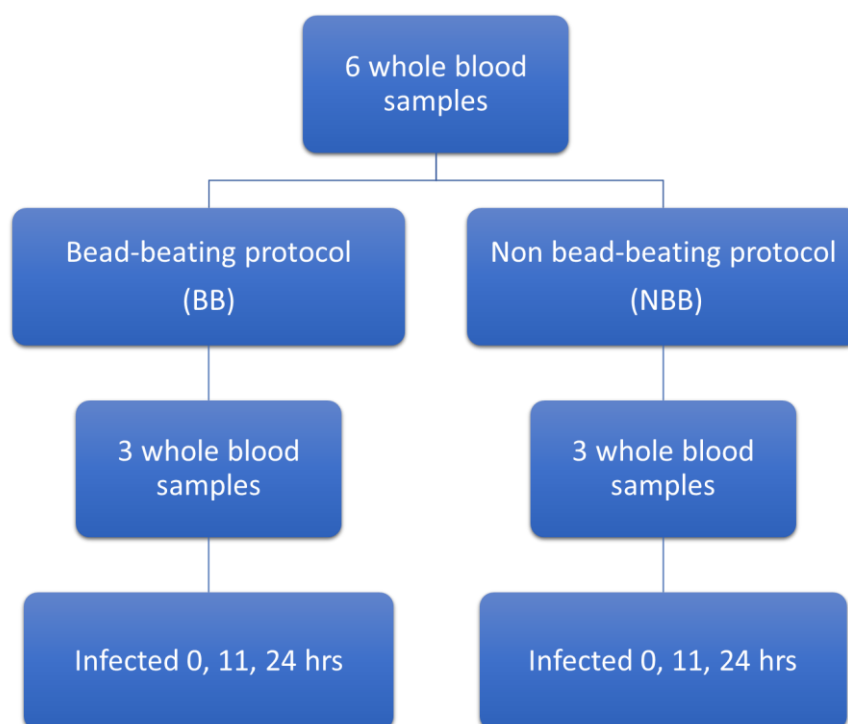


Figure 5.13: Pilot dual RNA-seq experimental design.

RNA quality and quantity check for pilot dual RNA-seq experiment of BCG lux infected WBA

Prior to sequencing, RNA quality and quantity was again measured (**Table 5.7**). Although all samples showed lower 260/280 and 260/230 measurements (ideally these should be 2) and some samples showed low RIN values (ideally this should be above a RIN of 7), sequencing was still possible. Furthermore, sample concentrations and volumes were overall low for bead-beat samples compared to non-bead-beat samples. These results were in alignment with the extraction optimisation steps, which suggested additional steps such as bead-beating would decrease sample concentrations. Despite this, sequencing was continued to determine if the data produced would be able to be analysed with a bioinformatic pipeline.

Table 5.7: RNA quality and quantity of BCG lux WBA samples prior to sequencing. Conc = concentration. D = donor; NBB = non bead-beat; BB = bead-beat.

Sample ID	Sample conc (ng/uμ)	Volume (μl)	260/280	260/230	RIN
D1_I_NBB_0hrs	30.0	18	1.81	0.97	9.3
D1_I_NBB_11hrs	30.0	18	1.72	0.74	8.8
D1_I_NBB_24hrs	30.0	18	1.72	0.71	9
D1_I_BB_0hrs	12.5	18	1.86	0.54	7.8
D1_I_BB_11hrs	8.5	18	1.83	0.39	7.1
D1_I_BB_24hrs	20.0	18	1.55	0.65	6.6

Flow cytometry analysis of pilot WBA samples

To account for variations in cell populations in different samples, flow cytometry analysis was performed to measure these populations during the WBA for the pilot dual RNA-seq experiment. Separate samples, from the same donors, were used for this analysis and treated identically to the samples sent for sequencing, prior to flow cytometry analysis. It was infeasible to carry out flow cytometry analysis on the samples from the 11-hour time-point as the sample preparation and analysis required 6 hours of work. However, flow cytometry analysis was performed on all samples from 0 and 24 hours despite not all samples being sent to sequencing. Cell population values were estimated from the other time-points during bioinformatics analysis. As a pilot analysis, the major cell populations of granulocytes, lymphocytes, and monocytes were selected to give an overview of cell change over time and during infection.

Live granulocytes remained at almost 100% in all samples and time-points, except infected samples at 24 hours where a small decrease in live granulocytes was observed (an average decrease of 8%, **Figure 5.14A**). The percentage of total granulocytes was observed at approximately 40% of the total cell population in all samples except, infected samples at 24 hours where the percentage of granulocytes decreased to an average of 23.4% (**Figure 5.14B**). This was also reflected in the granulocyte count (**Figure 5.14C**). As there was no reduction in cell counts in the uninfected samples over time, this decrease in granulocyte cell numbers is a consequence of infection and likely reflects infected cell death.

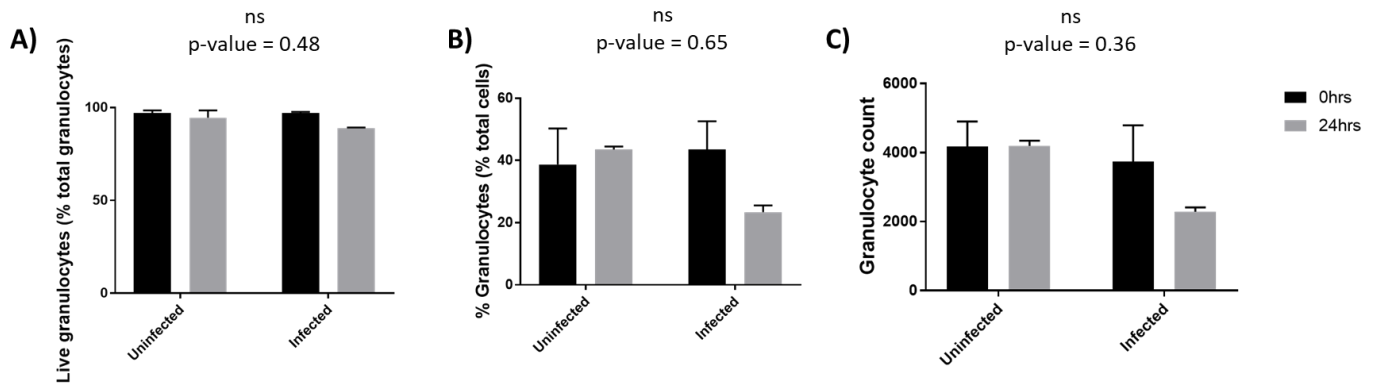


Figure 5.14: Granulocyte cell population analysis upon infection of whole blood with BCG lux for the pilot dual RNA-seq experiment. A) Live granulocytes as a percentage of total granulocytes in uninfected and infected samples at 0- and 24-hours post-infection. **B)** Percentage of granulocytes from total cells in infected and uninfected samples at 0- and 24-hours post-infection. **C)** Granulocyte count in infected and uninfected samples at 0- and 24-hours post-infection. $N = 2$. ns = non-significant.

In contrast, no statistical difference was observed in live lymphocyte levels due to the time-course or infection, as the percentage of live lymphocytes remained at around 100% over the time-course of the experiment (Figure 5.15A). Lymphocyte percentage, as a proportion of total cells, was between 30.8 – 40.2% in uninfected cells which did not change over the time-course (Figure 5.15B). During infection, however, the lymphocyte population increased to an average of 51.1%, likely as a response to infection. This trend was also observed in the lymphocyte counts (Figure 5.15C).

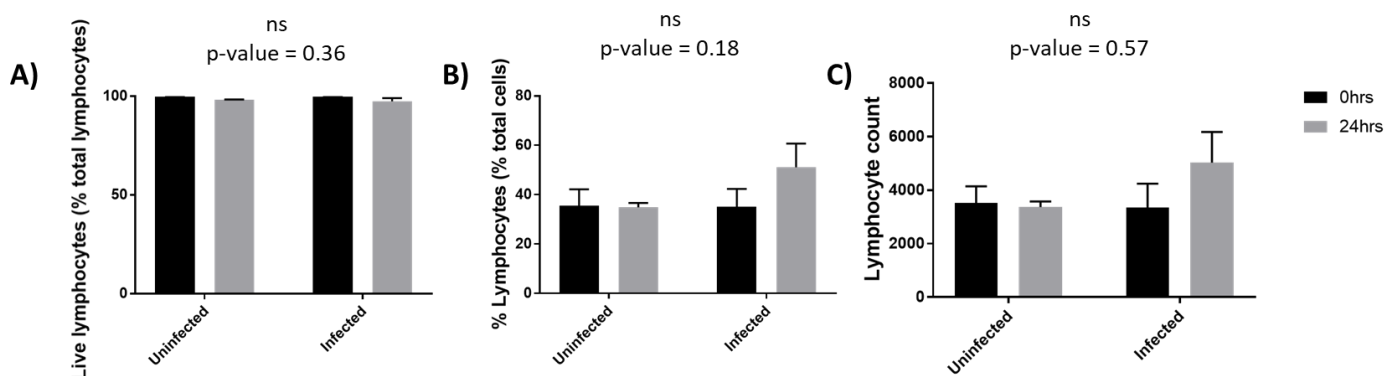


Figure 5.15: Lymphocyte cell population analysis upon infection of whole blood with BCG lux for the pilot dual RNA-seq experiment. A) Live lymphocytes as a percentage of total lymphocytes in uninfected and infected samples at 0- and 24-hours post-infection. **B)** Percentage of lymphocytes from total cells in infected and uninfected samples at 0- and 24-hours post-infection. **C)** Lymphocyte count in infected and uninfected samples at 0- and 24-hours post-infection. $N = 2$. ns = non-significant.

As was observed in previous flow cytometry experiments, the monocyte population followed a different trend to the granulocytes and lymphocytes. The percentage of live monocytes was low (at an average of 23.1%) in both uninfected and infected cells at 0 hours, however this population significantly increased to an average of 97.2% at 24 hours for both infected and uninfected samples (**Figure 5.16A**). As a percentage of total cells, the monocyte population remained between 2.1 and 5.4% in all samples and across time-points, with the only difference observed as an average increase of 0.9% in monocytes in infected blood samples from 0 hours to 24 hours (**Figure 5.16B**). This was again reflected in the monocyte count, where monocyte numbers remained overall low (**Figure 5.16C**).

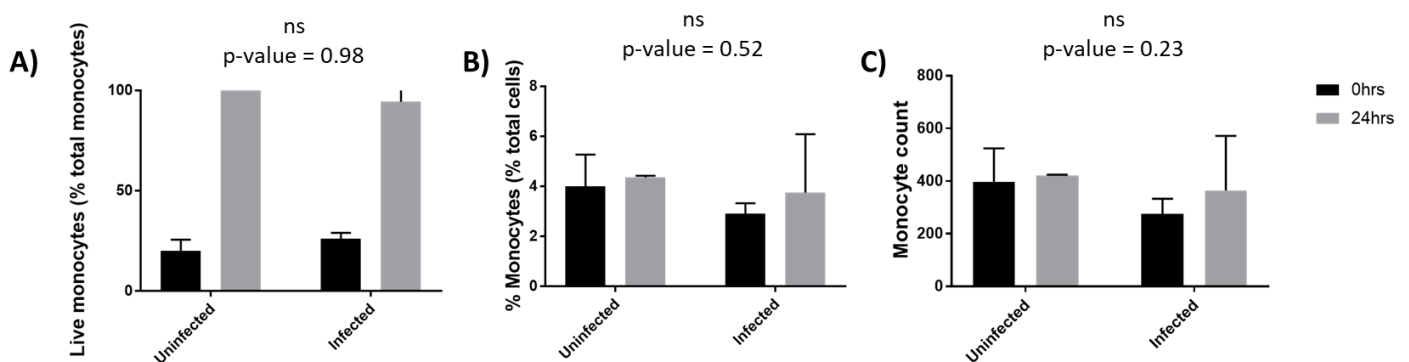


Figure 5.16: Monocyte cell population analysis upon infection of whole blood with BCG lux for the pilot dual RNA-seq experiment. A) Live monocytes as a percentage of total lymphocytes in uninfected and infected samples at 0- and 24-hours post-infection. **B)** Percentage of monocytes from total cells in infected and uninfected samples at 0- and 24-hours post-infection. **C)** Monocyte count in infected and uninfected samples at 0- and 24-hours post-infection. N = 2. Ns = non-significant.

Pilot dual RNA-seq experiment bioinformatic pipeline

A bioinformatic pipeline was set up to enable sequence mapping of both host and bacterial reads (**Figure 2.6**). Briefly, host reads were aligned and mapped using a standard workflow including a splice-aware aligner, STAR, before unmapped reads were used for bacterial alignment and mapping using a standard workflow including a splice-unaware aligner called bwa.

Dual RNA-seq quality control

Prior to gene expression analysis, it is important to perform several quality control steps to prevent contamination, and variation between samples, from affecting downstream bioinformatic analyses. It is also important to understand the composition of each sample in terms of different RNA species which may affect downstream results. Several steps are taken both in the sample preparation stage, such as DNase treatment, and in the library preparation stage, such as rRNA depletion, to ensure the

samples contain as few contaminants as possible and only specific mRNA are sequenced. In this case, the RNA species of most interest was mRNA. Furthermore, as red blood cells make up a greater proportion of whole blood than the white blood cells of interest in this experiment, red blood cell RNA, namely globin RNA, was removed during library preparation.

The first quality control step was running FASTQC on the FASTQ files which are obtained from sequencing. These files contained the read sequences of each sample and FASTQC provides information on several factors including basic statistics, sequence quality, sequence content, sequence lengths, sequence duplication, and overrepresented sequences. **Figure S5.2** shows an example FASTQC report for one of the samples. All of the FASTQC reports generated from this experiment suggested there was high GC content across the sequences. The GC content of RNA describes the proportion of G and C bases across different sequences, and is characteristic of the sample species, and RNA species, being sequenced. The expected GC content for human DNA is 40.4% (<https://www.ncbi.nlm.nih.gov/genome/?term=Homo+sapiens>), while the expected GC content for BCG DNA is 65.6% (<https://www.ncbi.nlm.nih.gov/genome/161>). The FASTQC reports here identified a GC content of between 55 and 71%, which is too high for most human mRNA reads but can correspond to either bacterial or human globin reads (Bohlin & H-O Pettersson, 2019).

To detect the level of globin in each sample, the proportion of the globin subunit beta RNA gene, *HBB*, was measured (**Figure 5.17**). The levels of the globin subunit alpha RNA gene, *HBA*, was also measured but these were relatively low across all samples and were therefore not removed (**Table 5.8**). Across all samples, the proportion of *HBB* was very high. The proportion of *HBB* was on average higher in samples without the bead-beating step than those including that step, and the lowest levels of *HBB* were detected in samples at 24 hours post-infection. This level of *HBB* would significantly impact downstream transcriptomics analysis as the *HBB* gene has been sequenced instead of the mRNA of interest in this analysis, host and bacterial mRNA. Reads from this gene were removed before differential gene expression analysis.

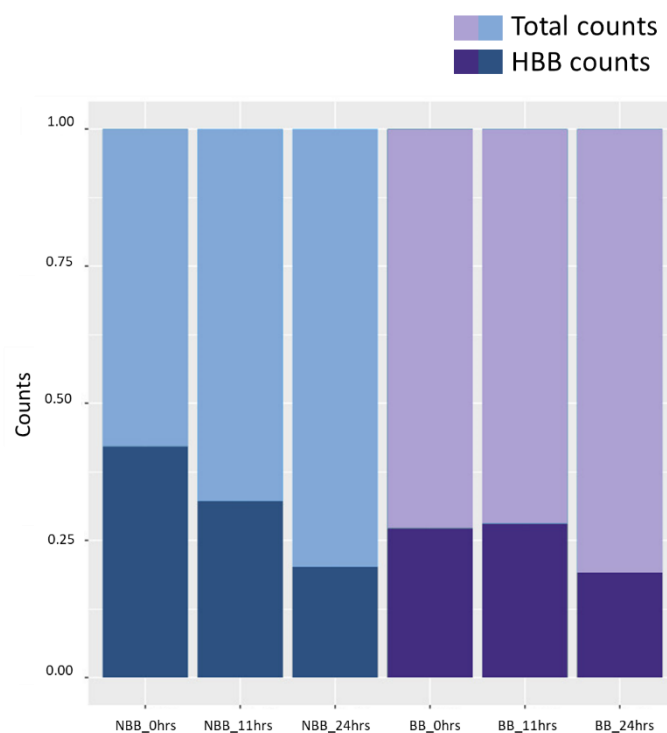


Figure 5.17: Percentage of HBB gene counts in pilot dual RNA-seq samples. NBB = non-bead-beat samples; BB = bead-beat samples. Darker colours indicate the proportion of total RNA reads corresponding to HBB gene counts.

Table 5.8: HBB and HBA counts and percentages in host samples. NBB = non bead-beat samples; BB = bead-beat samples.

Sample	HBB	HBA1	HBA2	Total counts	HBB %	HBA1 %	HBA2 %
NBB 0 hrs	1.37E+07	6.69E+04	1.06E+06	3.25E+07	42.17	0.21	3.26
NBB 11 hrs	1.09E+07	8.07E+04	1.01E+06	3.37E+07	32.18	0.24	3.01
NBB 24 hrs	6.43E+06	5.90E+04	8.08E+05	3.18E+07	20.19	0.19	2.54
BB 0 hrs	5.61E+06	2.87E+04	3.64E+05	2.06E+07	27.21	0.14	1.77
BB 11 hrs	6.52E+06	2.40E+04	3.44E+05	2.33E+07	28.01	0.10	1.48
BB 24 hrs	4.10E+06	1.76E+04	2.16E+05	2.14E+07	19.13	0.08	1.01

Furthermore, the concentration of rRNA, which should also have been depleted in previous library preparation steps, was measured. In contrast to globin RNA, rRNAs were found at very low levels (<5%) across all samples (**Table 5.9**). This would therefore not cause any significant effect on sequencing of the target RNAs in the samples.

Table 5.9: Percentage of mapped rRNA reads in pilot dual RNA-seq samples. NBB = non bead-beat samples; BB = bead-beat samples.

Sample	Total counts	rRNA gene counts	Percentage of total RNA
NBB 0 hrs	6629	172	2.6
NBB 11 hrs	11243	173	1.5
NBB 24 hrs	11297	143	1.3
BB 0 hrs	33081	1411	4.3
BB 11 hrs	29304	894	3.1
BB 24 hrs	27622	623	2.3

To identify if the bead-beating step resulted in a different level of host or bacterial reads being sequenced from samples, the percentage of reads which mapped and number of genes detected was calculated. Furthermore, it has been suggested that 2 to 5×10^6 bacterial reads after rRNA depletion are required in a dual RNA-seq experiment to be sufficient for statistical analysis (Westermann et al., 2012).

Table 5.10: Summary of mapped reads to host and bacteria. NBB = non bead-beat samples; BB = bead-beat samples.

Sample	Organism	Input reads	Uniquely mapped reads	Uniquely mapped reads (%)	Unmapped reads	Unmapped reads (%)
NBB 0 hrs	Host	7.E+07	4.4E+07	62.79	1.7E+06	2.4
NBB 11 hrs	Host	8.E+07	5.1E+07	67.42	2.3E+06	3.1
NBB 24 hrs	Host	7.E+07	5.4E+07	73.57	2.5E+06	3.4
BB 0 hrs	Host	6.E+07	5.4E+07	84.67	1.9E+06	3.0
BB 11 hrs	Host	7.E+07	6.3E+07	87.5	2.1E+06	2.9
BB 24 hrs	Host	7.E+07	6.5E+07	90.23	2.1E+06	3.0
NBB 0 hrs	Bacteria	7.E+07	7.2E+03	1.0E-02	3.4E+06	99.5
NBB 11 hrs	Bacteria	8.E+07	1.2E+04	1.6E-02	4.5E+06	96.7
NBB 24 hrs	Bacteria	7.E+07	1.2E+04	1.6E-02	5.0E+06	97.7
BB 0 hrs	Bacteria	6.E+07	3.4E+04	5.4E-02	3.7E+06	97.7
BB 11 hrs	Bacteria	7.E+07	3.1E+04	4.2E-02	4.2E+06	97.6
BB 24 hrs	Bacteria	7.E+07	2.9E+04	4.1E-02	4.3E+06	97.2

The percentage of genes which mapped to the host transcriptome was higher in bead-beat samples (minimum of 85%) than those without the bead-beating step (maximum of 74%, **Figure 5.18A**). This was also observed for the total number of reads mapping to the host transcriptome (e.g. minimum of 5.4×10^7 for bead-beat samples and a maximum of 5.4×10^7 for non bead-beat samples, **Figure 5.18B**). However, the total number of uniquely mapped host genes detected from these reads were similar in both bead-beat and non-bead-beat samples (an average of 2.5×10^4 for non bead-beat samples and 2.6×10^4 for bead-beat samples, **Figure 5.18C**). This suggests that a bead-beating step in the extraction procedure may have improved the number of reads mapped to the host transcriptome. Overall, the percentage of mapped reads, total number of mapped reads, and number of uniquely mapped genes from the host increased over time.

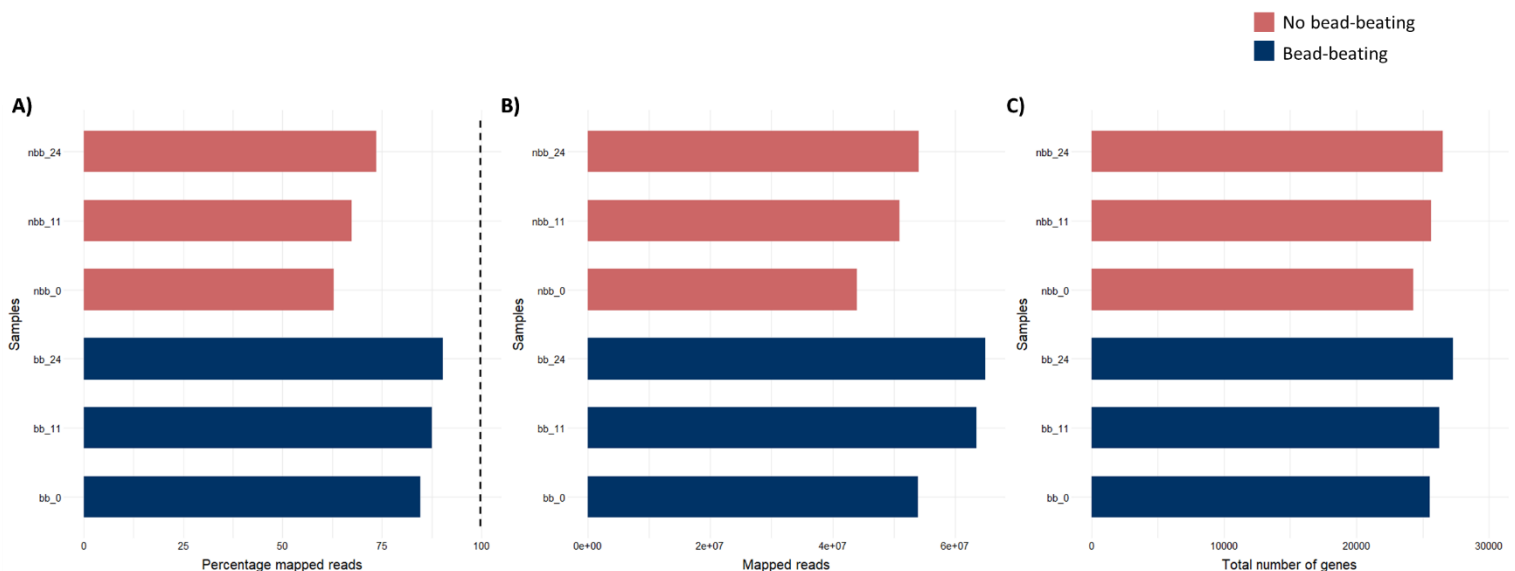


Figure 5.18: Proportion of reads mapping to host transcripts in bead-beat and non-bead-beat samples. A) Percentage of reads which mapped to host transcripts in bead-beat and non-bead-beat samples. **B)** Number of reads which mapped to host transcripts in bead-beat and non-bead-beat samples. **C)** Number of host genes in bead-beat and non-bead-beat samples. $N = 6$.

The percentage of reads which mapped to the bacterial transcriptome (from total reads) was also higher in samples which included a bead-beating step (a minimum of 0.041%) than those without (a maximum of 0.016%), although this difference was significantly larger for the bacterial mapping than the host (**Figure 5.19A**). This was also reflected in the number of reads mapped to the bacterial transcriptome (minimum of 2.9×10^4 for bead-beat samples and maximum of 1.2×10^4 for non bead-

beat samples, **Figure 5.19B**) and in the number of uniquely mapped bacterial genes detected across samples (minimum of 2,736 for bead-beat samples and maximum of 783 for non bead-beat samples, **Figure 5.19C**). Contrary to previous studies which observed an increase in bacterial reads with time hypothesised to be due to increased bacterial replication, the percentage of mapped reads did not increase with time and instead decreased.

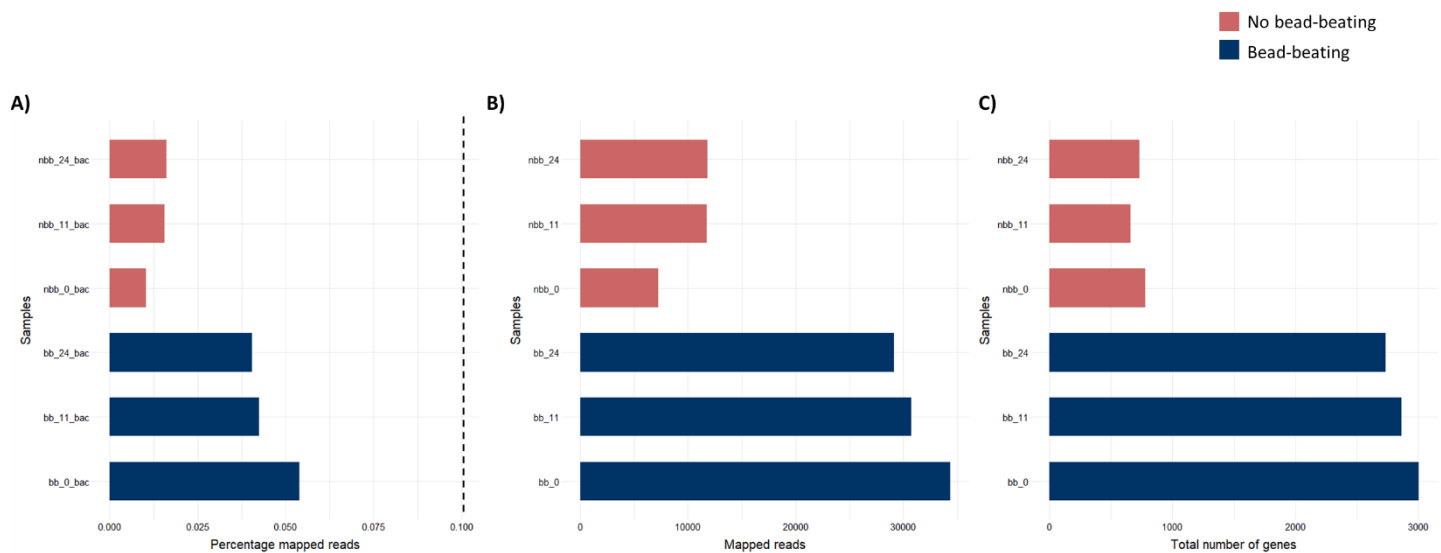


Figure 5.19: Proportion of reads mapping to bacterial transcripts in bead-beat and non-bead-beat samples. A) Percentage of total reads which mapped to bacterial transcripts in bead-beat and non-bead-beat samples. B) Number of reads which mapped to bacterial transcripts in bead-beat and non-bead-beat samples. C) Number of bacterial genes in bead-beat and non-bead-beat samples. N = 6.

Dual RNA-seq pre-processing analysis

The first step of differential gene expression analysis of RNA-seq data is to perform quality checks across all the available sequence data through visualisation analyses such as PCA. PCA can identify samples which are outliers due to both technical differences, such as read depth, or biological differences. PCA was performed using the read counts for all transcripts, for both host and bacterial transcripts separately. For the host, the greatest source of variation was due to the time-course, as shown across PC1 (**Figure 5.20A**). Specifically, the greatest variation was observed between 0 hours and late time-points (11 and 24 hours), in both bead-beat and non-bead-beat samples, although the later time-points also showed separation. The second greatest source of variation across samples was the different bead-beating conditions, as denoted by the large gap observed in PC2.

In contrast, the largest source of variation in bacterial reads was due to the different bead-beating conditions (**Figure 5.20B**). This is likely due to the significant difference in number of bacterial reads observed between these two conditions from the mapping analysis. The second greatest source of variation across bacterial samples was the time-course, again with the greatest spread resulting from the 0 hours time-point compared to the later time-points (11 and 24 hours). The 11- and 24-hour time-points also separated from each other, but not as vastly. This was seen for both bead-beat and non-bead-beat samples. This suggests that for both the host and bacterial reads, the greatest difference in transcriptomics will be observed between the 0 hour and later time-points.

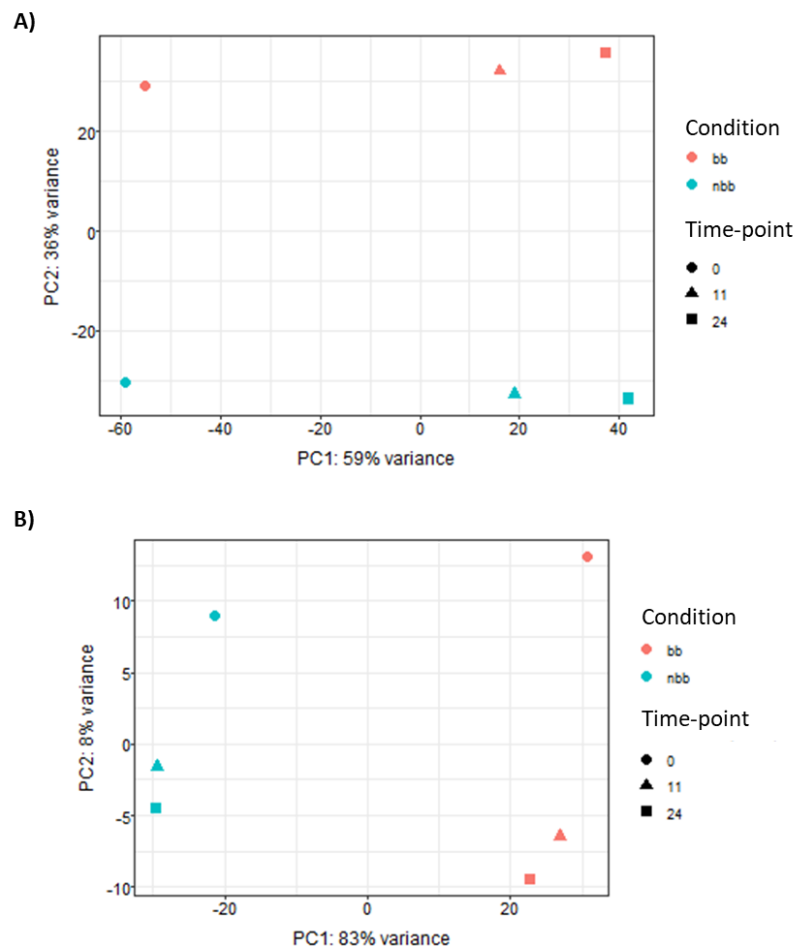


Figure 5.20: PCA plots of host and bacterial transcripts from the pilot dual RNA-seq experiment after *r*-log normalisation. A) PCA plot of host transcripts for all samples in the pilot dual RNA-seq experiment. B) PCA plot of bacterial transcripts for all samples in the pilot dual RNA-seq experiment. BB = bead-beat; NBB = non-bead-beat. N = 6.

Differential gene expression analysis of pilot dual RNA-seq samples

Differential gene expression analysis was carried out as described in **Chapter 2.4.8**. Due to limitations in performing differential gene expression analysis on only 3 samples, both non-bead-beat and bead-beat samples were used for this analysis despite non-bead-beat samples producing few bacterial reads. It is likely that major changes in host gene expression will still be captured using all samples, however bacterial gene expression changes may be compromised. While there are specific approaches to analyse time-course data, such as using likelihood ratio tests, the lack of biological replicates in this experiment meant that these were unusable. Therefore, standard differential gene expression analysis was performed for each time-point comparison, however no significance values are described and results should be considered with caution as there is a lack of appropriate biological replicates. The total number of differentially expressed genes for the host can be seen in **Table 5.10**.

Table 5.11: Summary of host differential gene expression results of the pilot dual RNA-seq samples. Non-0 = genes with non-zero counts.

Comparison	Total	Non-0	Log ₂ FC > 2	Log ₂ FC < -2
0 vs 11 hrs	60623	31082	435	137
11 vs 24 hrs	60623	31082	20	32
0 vs 24 hrs	60623	31082	750	542

The top 50 host genes which were differentially expressed in the 0 vs 11 hrs time-point can be seen in **Table 5.11**, those which were differentially expressed in the 11 vs 24 hrs time-point can be seen in **Table 5.12**, and those which were differentially expressed in the 0 vs 24 hrs time-point can be seen in **Table 5.13**. There was overall a greater proportion of up-regulated genes at the early time-point and a greater proportion of down-regulated genes at the later time-point. However, there was a large reduction in differentially expressed genes in the 11 vs 24 hrs comparison than the 0 vs 11 hrs comparison. This could be due to several reasons; it is expected that large changes in gene expression are observed at early time-points and further gene expression changes may only be seen after 24 hours.

Nevertheless, several C-C motif chemokine ligand (*CCL*) genes were up-regulated at the early time-point, alongside other immune and inflammatory genes *IL1A* (involved in inflammation (Dinarelli, 2018)) and *ACOD1* (involved in immunometabolism in response to pathogens (Wu et al., 2020)).

At the later time-point, many of the most up-regulated genes were several long noncoding RNAs and genes involved in cell development such as *TBX3* (Khan et al., 2020) and a T cell receptor gene *TRGJ2* (Sachdeva et al., 1997). Down-regulated genes at the later time-point also included long noncoding RNAs but also some immune-related genes such as *TNFRSF10C* (reduced in active TB patient PBMCs compared to latent TB patient PBMCs after stimulation with PPD (Wang et al., 2019)) and *CXCR2* (involved in the inflammasome response to Mtb (Boro & Balaji, 2017)). At the final time-point comparison of 0 vs 24 hours, the same genes are identified as in the previous 2 comparisons.

Table 5.12: Top 50 differentially expressed host genes for 0 vs 11 hrs time-points in the pilot dual RNA-seq experiment. Top 25 host genes which are up-regulated and top 25 host genes which are down-regulated are shown below. Ordered by log₂ fold change and filtered for an absolute log₂ fold change of 2.

Ensembl ID	Base mean	log2FoldChange	Gene ID
ENSG00000108691	9729.5	16	<i>CCL2</i>
ENSG00000115008	6353.3	13	<i>IL1A</i>
ENSG00000102794	784.3	12.7	<i>ACOD1</i>
ENSG00000106178	313.1	11.7	<i>CCL24</i>
ENSG00000108688	695.5	11.6	<i>CCL7</i>
ENSG00000166920	1768.1	11.1	<i>C15orf48</i>
ENSG00000259354	858.6	10.7	
ENSG00000103855	293.6	10.6	<i>CD276</i>
ENSG00000122861	4874.9	10.3	<i>PLAU</i>
ENSG00000115009	477.7	10.1	<i>CCL20</i>
ENSG00000108700	2486.4	10	<i>CCL8</i>
ENSG00000260135	150.8	10	<i>MMP2-AS1</i>
ENSG00000276085	3133	9.8	<i>CCL3L3</i>
ENSG00000008118	262.8	9.8	<i>CAMK1G</i>
ENSG00000173391	1543.8	9.7	<i>OLR1</i>
ENSG00000179362	152.2	9.7	<i>HMGN2P46</i>
ENSG00000137491	150.7	9.7	<i>SLCO2B1</i>
ENSG00000102962	910.5	9.4	<i>CCL22</i>
ENSG00000267034	93.5	9.3	<i>NA</i>
ENSG00000242048	118	9.2	
ENSG00000277632	16871.7	9.1	<i>CCL3</i>
ENSG00000122641	2632.9	9.1	<i>INHBA</i>
ENSG00000222047	114.5	9.1	<i>C10orf55</i>
ENSG00000203364	383.3	9	
ENSG00000138135	48.5	9	<i>CH25H</i>
ENSG00000283378	63.2	-10.1	<i>CNTNAP3C</i>
ENSG00000157445	30.0	-9.0	<i>CACNA2D3</i>
ENSG00000213876	15.7	-8.0	<i>RPL7AP64</i>
ENSG00000124302	15.0	-8.0	<i>CHST8</i>
ENSG00000184389	14.0	-7.8	<i>A3GALT2</i>

ENSG00000275278	11.0	-7.4	
ENSG00000267262	9.7	-7.4	
ENSG00000100079	9.4	-7.3	<i>LGALS2</i>
ENSG00000267589	6.9	-6.9	
ENSG00000231680	6.1	-6.7	<i>LINC02723</i>
ENSG00000182272	6.7	-6.4	<i>B4GALNT4</i>
ENSG00000034239	4.5	-6.3	<i>EFCAB1</i>
ENSG00000259496	4.4	-6.2	<i>GEMIN8P1</i>
ENSG00000145708	4.8	-6.2	<i>CRHBP</i>
ENSG00000106714	938.1	-6.2	<i>CNTNAP3</i>
ENSG00000119457	39.2	-6.2	<i>SLC46A2</i>
ENSG00000161944	25.7	-6.2	<i>ASGR2</i>
ENSG00000141505	12.0	-6.2	<i>ASGR1</i>
ENSG00000163485	3.9	-6.1	<i>ADORA1</i>
ENSG00000233392	3.7	-6.0	<i>UICLM</i>
ENSG00000272434	3.7	-5.9	
ENSG00000092051	3.9	-5.9	<i>JPH4</i>
ENSG00000171385	18.1	-5.9	<i>KCND3</i>
ENSG00000231508	3.3	-5.8	<i>RPL34P20</i>
ENSG00000142494	11.4	-5.8	<i>SLC47A1</i>

Table 5.13: Top 50 differentially expressed host genes for 11 vs 24 hrs time-points in the pilot dual RNA-seq experiment. Top 25 host genes which are up-regulated and top 25 host genes which are down-regulated are shown below. Ordered by log₂ fold change and filtered for an absolute log₂ fold change of 2.

Ensembl ID	Base mean	log2FoldChange	Gene ID
ENSG00000232679	9.5	7.3	<i>LINC01705</i>
ENSG00000196932	28.6	6.5	<i>TMEM26</i>
ENSG00000135111	5.2	6.5	<i>TBX3</i>
ENSG00000154864	4.9	6.3	<i>PIEZO2</i>
ENSG00000229664	4.5	6.2	
ENSG00000226289	4.7	6.0	<i>CDCA4P3</i>
ENSG00000258826	10.2	5.9	<i>LINC01147</i>
ENSG00000267729	2.9	5.6	
ENSG00000231784	3.5	5.6	<i>DBIL5P</i>
ENSG00000213277	2.6	5.5	<i>MARCKSL1P1</i>
ENSG00000235480	2.9	5.4	
ENSG00000169908	7.6	5.4	<i>TM4SF1</i>
ENSG00000124251	2.6	5.4	<i>TP53TG5</i>
ENSG00000238042	4.5	5.4	<i>LINC02257</i>
ENSG00000211687	2.8	5.4	<i>TRGJ2</i>
ENSG00000197149	2.8	5.3	
ENSG00000166396	2.4	5.3	<i>SERPINB7</i>
ENSG00000282849	2.3	5.3	
ENSG00000227495	2.3	5.3	<i>KIF1C-AS1</i>
ENSG00000212409	2.3	5.3	<i>RNY4P18</i>
ENSG00000218857	3.6	5.2	
ENSG00000197046	10.4	5.2	<i>SIGLEC15</i>
ENSG00000220201	3.1	5.2	<i>ZGLP1</i>
ENSG00000228957	2.1	5.1	
ENSG00000263683	2.0	5.1	
ENSG00000249662	77.9	-7.9	<i>LINC02218</i>
ENSG00000249001	9.5	-7.6	
ENSG00000173535	1963.3	-7.0	<i>TNFRSF10C</i>
ENSG00000275371	25.2	-6.9	
ENSG00000184113	33.8	-6.8	<i>CLDN5</i>

ENSG00000127831	51.2	-6.7	<i>VIL1</i>
ENSG00000263629	5.9	-6.2	<i>MIR5586</i>
ENSG00000287131	9.1	-6.1	
ENSG00000258831	89.6	-6.1	
ENSG00000212283	7.5	-6.0	<i>SNORD89</i>
ENSG00000258476	449.1	-6.0	<i>LINC02207</i>
ENSG00000286480	27.8	-6.0	
ENSG00000236345	14.7	-5.9	<i>SCAT8</i>
ENSG00000251073	7.0	-5.8	<i>NUDT19P5</i>
ENSG00000188107	14.1	-5.8	<i>EYS</i>
ENSG00000130881	29.3	-5.7	<i>LRP3</i>
ENSG00000136999	103.3	-5.5	<i>CCN3</i>
ENSG00000171798	9.6	-5.4	<i>KNDC1</i>
ENSG00000157168	70.6	-5.4	<i>NRG1</i>
ENSG00000254134	2.1	-5.4	<i>IGHVII-74-1</i>
ENSG00000178722	8.9	-5.4	<i>C5orf64</i>
ENSG00000134042	2.4	-5.3	<i>MRO</i>
ENSG00000180871	14328.5	-5.3	<i>CXCR2</i>
ENSG00000234383	41.5	-5.2	<i>CTBP2P8</i>
ENSG00000165731	6.7	-5.2	<i>RET</i>

Table 5.14: Top 50 differentially expressed host genes for 0 vs 24 hrs time-points in the pilot dual RNA-seq experiment. Top 25 host genes which are up-regulated and top 25 host genes which are down-regulated are shown below. Ordered by log₂ fold change and filtered for an absolute log₂ fold change of 2.

Ensembl ID	Base mean	log2FoldChange	Gene ID
ENSG00000108691	9729.5	16.0	<i>CCL2</i>
ENSG00000108700	2486.4	13.5	<i>CCL8</i>
ENSG00000115008	6353.3	12.9	<i>IL1A</i>
ENSG00000108688	695.5	12.6	<i>CCL7</i>
ENSG00000203364	383.3	12.2	
ENSG00000102794	784.3	12.0	<i>ACOD1</i>
ENSG00000108342	238.8	11.6	<i>CSF3</i>
ENSG00000008118	262.8	11.4	<i>CAMK1G</i>
ENSG00000259354	858.6	11.4	
ENSG00000103855	293.6	11.2	<i>CD276</i>
ENSG00000166920	1768.1	11.1	<i>C15orf48</i>
ENSG00000122641	2632.9	11.0	<i>INHBA</i>
ENSG00000235385	170.2	10.9	<i>LINC02154</i>
ENSG00000122861	4874.9	10.8	<i>PLAU</i>
ENSG00000272512	151.4	10.8	
ENSG00000136244	2919.2	10.7	<i>IL6</i>
ENSG00000102962	910.5	10.7	<i>CCL22</i>
ENSG00000276085	3133.0	10.4	<i>CCL3L3</i>
ENSG00000196917	313.9	10.3	<i>HCAR1</i>
ENSG00000115009	477.7	10.3	<i>CCL20</i>
ENSG00000165474	119.8	10.3	<i>GJB2</i>
ENSG00000137491	150.7	10.2	<i>SLCO2B1</i>
ENSG00000179362	152.2	10.2	<i>HMG2P46</i>
ENSG00000277632	16871.7	10.2	<i>CCL3</i>
ENSG00000260135	150.8	10.0	<i>MMP2-AS1</i>
ENSG00000173535	1963.3	-10.5	<i>TNFRSF10C</i>
ENSG00000157168	70.6	-10.4	<i>NRG1</i>
ENSG00000249662	77.9	-10.3	<i>LINC02218</i>
ENSG00000283378	63.2	-10.3	<i>CNTNAP3C</i>
ENSG00000127831	51.2	-9.8	<i>VIL1</i>

ENSG00000106714	938.1	-9.8	<i>CNTNAP3</i>
ENSG00000234383	41.5	-9.6	<i>CTBP2P8</i>
ENSG00000157445	30.0	-9.2	<i>CACNA2D3</i>
ENSG00000184113	33.8	-9.1	<i>CLDN5</i>
ENSG00000130881	29.3	-9.0	<i>LRP3</i>
ENSG00000146122	72.7	-9.0	<i>DAAM2</i>
ENSG00000286480	27.8	-8.9	
ENSG00000180871	14328.5	-8.7	<i>CXCR2</i>
ENSG00000275371	25.2	-8.6	
ENSG00000171385	18.1	-8.5	<i>KCND3</i>
ENSG00000136999	103.3	-8.4	<i>CCN3</i>
ENSG00000287586	18.5	-8.4	
ENSG00000124302	15.0	-8.2	<i>CHST8</i>
ENSG00000105507	42.3	-8.1	<i>CABP5</i>
ENSG00000163464	7363.1	-8.1	<i>CXCR1</i>
ENSG00000238078	13.4	-8.0	<i>LINC01352</i>
ENSG00000188013	14.0	-8.0	<i>MEIS3P2</i>
ENSG00000284948	75.8	-7.9	
ENSG00000236279	13.0	-7.9	<i>CLEC2L</i>
ENSG00000139835	12.8	-7.9	<i>GRTP1</i>

DESeq2 normalised bead-beat samples were used to calculate log₂ fold change values. Genes were filtered for high up or down-regulation (log₂ fold change of absolute 2). The total number of genes meeting this filter selection can be seen in **Table 5.14**. A total of 3411 genes had at least one count, covering approximately 84.6% of the BCG genome. Similar to the host results, a greater number of genes were found to be up and down-regulated at the early time-points compared to the later time-points.

Table 5.15: Summary of bacterial differential gene expression results of the pilot dual RNA-seq bead-beat samples. Non-0 = genes with non-zero counts.

Comparison	Total	Non-0	Log ₂ FC > 2	Log ₂ FC < -2
0 vs 11 hrs	4034	3411	124	194
11 vs 24 hrs	4034	3411	99	122
0 vs 24 hrs	4034	3411	123	262

The top 50 bacterial genes which were differentially expressed in the 0 vs 11 hrs time-point can be seen in **Table 5.15**, those which were differentially expressed in the 11 vs 24 hrs time-point can be seen in **Table 5.16**, and those which were differentially expressed in the 0 vs 24 hrs time-point can be seen in **Table 5.17**. Unlike the host transcriptome analysis, there was a greater proportion of down-regulated genes at both time-point comparisons. However, similar to the host transcriptome analysis but not as pronounced, there was a reduction in both up-regulated and down-regulated differentially expressed genes at the later time-point comparison.

When comparing 0 hrs to 11 hrs, the most up-regulated gene identified was *hsp*, a heat shock protein which has shown to be involved in TLR signalling and inflammation (Bulut et al., 2005), followed by *mbtB* and *aprB* which are required for growth and phagolysosome adaptation within macrophage cells, respectively (Abramovitch et al., 2011; de Voss et al., 2000). The most down-regulated genes at this time-point comparison were *bfrB* which encodes a ferritin, a major iron storage protein in Mtb (Kurthkoti et al., 2015), *hupB* which is involved in chromosome structuring (Hołówka et al., 2017), as well as genes involved in metabolism such as *leuC* (leucine metabolism (Amorim Franco & Blanchard, 2017)) and *inhA* (fatty acid metabolism (Marrakchi et al., 2000)), suggesting a down-regulation of pathways not relevant for initial infection. Of note, several uncharacterised genes were also identified as highly up or down-regulated in this comparison.

When comparing changes in gene expression at 11 hrs with 24 hrs, many of the top genes identified are uncharacterised. The top up-regulated genes were *irtA*, involved in iron transport (Ryndak et al., 2010), *cydD*, involved in growth (Lin Aung et al., 2014), and *malQ*, involved in cell envelope biosynthesis (Birhanu et al., 2019), possibly reflecting a down-regulation of pathways involved in initial infection and up-regulation of pathways involved in survival and replication. The top down-regulated genes were *cysE*, involved in serine biosynthesis and self-defence (Qiu et al., 2014), *pra*, a proline-rich antigen with limited characterisation (Bateman et al., 2021), and *omt*, a catechol o-methyltransferase involved in a range of functions in other bacteria but is yet uncharacterised in Mtb (Bateman et al., 2021; Lee et al., 2019). Again, at the final time-point comparison of 0 vs 24 hours, the same genes are identified as in the previous 2 comparisons.

Table 5.16: Top 50 up and down-regulated BCG *lux* genes at 0 vs 11 hrs according to log₂ fold change.

Up-regulated		Down-regulated	
Gene ID	Log ₂ fold change	Gene ID	Log ₂ fold change
<i>hsp</i>	5.08	<i>bfrB</i>	-5.26
<i>mbtB</i>	3.96	<i>BQ2027_MB1561</i>	-3.96
<i>aprB</i>	3.74	<i>hupB</i>	-3.72
<i>lat</i>	3.49	<i>BQ2027_MB2976</i>	-3.71
<i>mbtD</i>	3.48	<i>lpqE</i>	-3.71
<i>BQ2027_MB3532C</i>	3.48	<i>BQ2027_MB1494C</i>	-3.57
<i>pks2</i>	3.42	<i>BQ2027_MB3122</i>	-3.57
<i>BQ2027_MB3617</i>	3.39	<i>BQ2027_MB1395</i>	-3.50
<i>sigL</i>	3.28	<i>vapc10</i>	-3.46
<i>kstd</i>	3.22	<i>leuC</i>	-3.43
<i>BQ2027_MB1939</i>	3.17	<i>inhA</i>	-3.42
<i>BQ2027_MB3260C</i>	3.14	<i>BQ2027_MB1803</i>	-3.28
<i>sodA</i>	3.10	<i>cyp144</i>	-3.24
<i>ephE</i>	3.05	<i>BQ2027_MB0907C</i>	-3.15
<i>papA1</i>	3.05	<i>fadD14</i>	-3.15
<i>recG</i>	2.98	<i>BQ2027_MB3512C</i>	-3.15
<i>ctpE</i>	2.92	<i>gid</i>	-3.15
<i>lppM</i>	2.92	<i>leuD</i>	-3.05
<i>BQ2027_MB2591</i>	2.92	<i>BQ2027_MB2804C</i>	-3.05
<i>BQ2027_MB3125C</i>	2.92	<i>BQ2027_MB3469C</i>	-3.05
<i>BQ2027_MB3125A</i>	2.92	<i>rplR</i>	-2.99
<i>dnaE2</i>	2.92	<i>BQ2027_MB1560</i>	-2.99
<i>cyp142b</i>	2.92	<i>BQ2027_MB0498</i>	-2.94
<i>BQ2027_MB3763</i>	2.92	<i>tpi</i>	-2.94
<i>cysE</i>	2.89	<i>prfA</i>	-2.92
<i>embA</i>	2.89	<i>dop</i>	-2.82
<i>gmk</i>	2.82	<i>ethR</i>	-2.82
<i>BQ2027_MB0154</i>	2.80	<i>cspB</i>	-2.82
<i>moaB2</i>	2.80	<i>BQ2027_MB0950C</i>	-2.82
<i>BQ2027_MB3320</i>	2.80	<i>BQ2027_MB0975C</i>	-2.82
<i>thiG</i>	2.77	<i>fabG1</i>	-2.82
<i>espk</i>	2.77	<i>subI</i>	-2.82
<i>35kd_ag</i>	2.71	<i>BQ2027_MB3121C</i>	-2.82
<i>BQ2027_MB1457</i>	2.70	<i>rpoZ</i>	-2.76
<i>BQ2027_MB2713C</i>	2.67	<i>BQ2027_MB2801C</i>	-2.72
<i>BQ2027_MB0032</i>	2.61	<i>BQ2027_MB3400C</i>	-2.71
<i>BQ2027_MB1054</i>	2.61	<i>moaD2</i>	-2.69
<i>acs</i>	2.61	<i>malQ</i>	-2.69
<i>sseC2</i>	2.60	<i>BQ2027_MB1957</i>	-2.69
<i>aprA</i>	2.60	<i>BQ2027_MB2342C</i>	-2.69
<i>mbtE</i>	2.54	<i>ethA</i>	-2.63
<i>BQ2027_MB3220</i>	2.53	<i>BQ2027_MB3119C</i>	-2.63
<i>mprB</i>	2.50	<i>ctaE</i>	-2.61
<i>BQ2027_MB0072</i>	2.49	<i>rplY</i>	-2.60
<i>BQ2027_MB0836C</i>	2.49	<i>BQ2027_MB0846C</i>	-2.56
<i>sigI</i>	2.49	<i>vapb2</i>	-2.55

<i>BQ2027_MB2748C</i>	2.49	<i>BQ2027_MB0434C</i>	-2.55
<i>fadD19</i>	2.49	<i>BQ2027_MB0901</i>	-2.55
<i>accA1</i>	2.43	<i>BQ2027_MB0934</i>	-2.55
<i>esxG</i>	2.42	<i>prsA</i>	-2.55

Table 5.17: Top 50 up and down-regulated BCG *lux* genes at 11 vs 24 hrs according to log₂ fold change.

Up-regulated		Down-regulated	
Gene ID	Log ₂ fold change	Gene ID	Log ₂ fold change
<i>irta</i>	2.93	<i>cysE</i>	-3.58
<i>BQ2027_MB0112C</i>	2.90	<i>pra</i>	-3.39
<i>BQ2027_MB0363C</i>	2.90	<i>omt</i>	-3.39
<i>cydD</i>	2.90	<i>aroG</i>	-3.35
<i>malQ</i>	2.90	<i>BQ2027_MB1394</i>	-3.28
<i>adhA</i>	2.90	<i>snzp</i>	-3.17
<i>BQ2027_MB3212A</i>	2.90	<i>ephE</i>	-3.05
<i>acrA1</i>	2.90	<i>sdaA</i>	-2.92
<i>BQ2027_MB0650C</i>	2.74	<i>BQ2027_MB1567</i>	-2.92
<i>cyp144</i>	2.74	<i>BQ2027_MB2085</i>	-2.92
<i>guaB1</i>	2.74	<i>rmlC</i>	-2.92
<i>BQ2027_MB2110</i>	2.74	<i>celA1</i>	-2.77
<i>BQ2027_MB2172C</i>	2.74	<i>BQ2027_MB0162A</i>	-2.77
<i>BQ2027_MB0069</i>	2.55	<i>bpoC</i>	-2.77
<i>BQ2027_MB0734</i>	2.55	<i>fum</i>	-2.77
<i>BQ2027_MB0901</i>	2.55	<i>BQ2027_MB2144C</i>	-2.77
<i>BQ2027_MB0995</i>	2.55	<i>BQ2027_MB3186C</i>	-2.77
<i>vapB16</i>	2.55	<i>mce4D</i>	-2.77
<i>ptka</i>	2.55	<i>proV</i>	-2.77
<i>yajC</i>	2.55	<i>BQ2027_MB3809</i>	-2.77
<i>BQ2027_MB2689C</i>	2.55	<i>espk</i>	-2.77
<i>nrdr</i>	2.55	<i>ppiA</i>	-2.72
<i>BQ2027_MB2825C</i>	2.55	<i>gabD2</i>	-2.72
<i>BQ2027_MB3518</i>	2.55	<i>rsha</i>	-2.72
<i>folE</i>	2.55	<i>BQ2027_MB0032</i>	-2.61
<i>prfA</i>	2.52	<i>BQ2027_MB0368</i>	-2.61
<i>mmpL6</i>	2.52	<i>pks6b</i>	-2.61
<i>mmaA2</i>	2.40	<i>rplV</i>	-2.61
<i>BQ2027_MB2440C</i>	2.40	<i>BQ2027_MB1054</i>	-2.61
<i>mtr</i>	2.40	<i>BQ2027_MB1660</i>	-2.61
<i>BQ2027_MB2999C</i>	2.40	<i>BQ2027_MB1758C</i>	-2.61
<i>cyp138</i>	2.34	<i>baca</i>	-2.61
<i>mymT</i>	2.34	<i>BQ2027_MB1942C</i>	-2.61
<i>moeA2</i>	2.34	<i>BQ2027_MB1945C</i>	-2.61
<i>BQ2027_MB0505</i>	2.34	<i>BQ2027_MB2712C</i>	-2.61
<i>BQ2027_MB0983</i>	2.34	<i>fprA</i>	-2.61
<i>BQ2027_MB1114C</i>	2.34	<i>BQ2027_MB3460C</i>	-2.61
<i>BQ2027_MB1168</i>	2.34	<i>acs</i>	-2.61
<i>BQ2027_MB1223</i>	2.34	<i>kstd</i>	-2.59
<i>BQ2027_MB1386</i>	2.34	<i>BQ2027_MB2849C</i>	-2.53
<i>cmk</i>	2.34	<i>lrpa</i>	-2.53
<i>BQ2027_MB1768C</i>	2.34	<i>sigL</i>	-2.43
<i>fadD15</i>	2.34	<i>BQ2027_MB0232C</i>	-2.42
<i>glnQ</i>	2.34	<i>pyrE</i>	-2.42
<i>BQ2027_MB2804C</i>	2.34	<i>PPE9</i>	-2.42
<i>cobB</i>	2.34	<i>xseA</i>	-2.42

<i>BQ2027_MB2951C</i>	2.34	<i>mmpL13</i>	-2.42
<i>BQ2027_MB3218</i>	2.34	<i>rocA</i>	-2.42
<i>BQ2027_MB3368C</i>	2.34	<i>rrf</i>	-2.42
<i>eccc4</i>	2.34	<i>cut1</i>	-2.42

Table 5.18: Top 50 up and down-regulated BCG *lux* genes at 0 vs 24 hrs according to log₂ fold change.

Up-regulated		Down-regulated	
Gene ID	Log ₂ fold change	Gene ID	Log ₂ fold change
<i>hsp</i>	6.00	<i>leuC</i>	-4.72
<i>lppM</i>	3.96	<i>vapc10</i>	-3.90
<i>pks2</i>	3.94	<i>rplY</i>	-3.90
<i>mbtB</i>	3.88	<i>leuD</i>	-3.84
<i>sodA</i>	3.67	<i>bfrB</i>	-3.81
<i>papA1</i>	3.63	<i>parB</i>	-3.78
<i>aprA</i>	3.42	<i>BQ2027_MB2976</i>	-3.71
<i>lat</i>	3.41	<i>PPE51</i>	-3.71
<i>usfY</i>	3.20	<i>fadD30</i>	-3.54
<i>dnaE2</i>	3.19	<i>rrf</i>	-3.50
<i>BQ2027_MB3532C</i>	3.19	<i>BQ2027_MB0368</i>	-3.42
<i>BQ2027_MB0833C</i>	3.12	<i>xseA</i>	-3.42
<i>BQ2027_MB1988A</i>	3.05	<i>BQ2027_MB1277C</i>	-3.42
<i>BQ2027_MB3317C</i>	3.05	<i>BQ2027_MB3370</i>	-3.42
<i>BQ2027_MB3617</i>	3.05	<i>glpQ1</i>	-3.36
<i>BQ2027_MB3650C</i>	3.05	<i>BQ2027_MB1803</i>	-3.21
<i>BQ2027_MB3260C</i>	2.98	<i>ltp3</i>	-3.12
<i>BQ2027_MB0112C</i>	2.90	<i>rplF</i>	-3.07
<i>mbtD</i>	2.90	<i>lpqX</i>	-3.05
<i>mbtJ</i>	2.90	<i>sigC</i>	-3.05
<i>mbtI</i>	2.90	<i>BQ2027_MB3469C</i>	-3.05
<i>BQ2027_MB2643C</i>	2.90	<i>dpre1</i>	-3.05
<i>fadE27</i>	2.90	<i>BQ2027_MB0048C</i>	-3.04
<i>moaB2</i>	2.84	<i>BQ2027_MB0475</i>	-3.04
<i>BQ2027_MB3759</i>	2.74	<i>BQ2027_MB0043C</i>	-2.98
<i>BQ2027_MB2748C</i>	2.74	<i>mgta</i>	-2.96
<i>BQ2027_MB0278C</i>	2.74	<i>BQ2027_MB0498</i>	-2.94
<i>eccd3</i>	2.74	<i>BQ2027_MB0900C</i>	-2.94
<i>guaB1</i>	2.74	<i>BQ2027_MB1335</i>	-2.94
<i>BQ2027_MB2110</i>	2.74	<i>pks9</i>	-2.94
<i>35kd_ag</i>	2.68	<i>glnA4</i>	-2.94
<i>mbtE</i>	2.63	<i>ilvN</i>	-2.94
<i>lipF</i>	2.58	<i>BQ2027_MB3030C</i>	-2.94
<i>accA1</i>	2.57	<i>vapc44</i>	-2.94
<i>BQ2027_MB0069</i>	2.55	<i>BQ2027_MB3460C</i>	-2.94
<i>BQ2027_MB0109</i>	2.55	<i>rplR</i>	-2.92
<i>adhE1</i>	2.55	<i>echA21</i>	-2.92
<i>BQ2027_MB0256C</i>	2.55	<i>accD4</i>	-2.92
<i>esxG</i>	2.55	<i>BQ2027_MB0251</i>	-2.82
<i>BQ2027_MB0995</i>	2.55	<i>proC</i>	-2.82
<i>BQ2027_MB1017C</i>	2.55	<i>fum</i>	-2.82
<i>vapB16</i>	2.55	<i>lpqY</i>	-2.82
<i>BQ2027_MB2689C</i>	2.55	<i>cysM</i>	-2.82
<i>nrdr</i>	2.55	<i>erg3</i>	-2.82
<i>BQ2027_MB3518</i>	2.55	<i>BQ2027_MB3121C</i>	-2.82
<i>mce4C</i>	2.55	<i>BQ2027_MB0222</i>	-2.79

<i>BQ2027_MB3891</i>	2.55	<i>BQ2027_MB0441</i>	-2.73
<i>gnd2</i>	2.50	<i>ctpC</i>	-2.73
<i>glpKb</i>	2.50	<i>BQ2027_MB3122</i>	-2.72
<i>whiB6</i>	2.47	<i>pks6b</i>	-2.69

5.2.7 Main dual RNA-seq experiment

Due to the COVID-19 pandemic, only the initial steps, i.e. the WBA and flow cytometry, of the main dual RNA-seq experiment were able to be carried out. Samples from 6 donors were analysed for flow cytometry and corresponding samples were frozen for RNA extraction.

Inoculation measurements of BCG lux for the dual RNA-seq experiment

Unlike the previous WBA experiments, the bacterial inoculation for the main dual RNA-seq experiment was not within the expected range of 6 – 10 RLU:CFU ratio (**Table 5.18**). Here, an RLU:CFU ratio of 20 – 24 was measured which corresponds to 10 times fewer bacteria present in the blood samples than what was planned. It is unclear why the bacteria grew differently here than in previous experiments. Unfortunately, this MOI is too low for dual RNA-seq and the experiment would need to be repeated to ensure enough bacterial RNA, compared to host RNA, would be available.

Table 5.19: RLU:CFU ratio calculation to determine BCG lux MOI used for inoculation of the WBA.

The 1 ml blood samples were used for flow cytometry analysis and the 2 ml samples were used for RNA extraction, which was later not performed.

Blood sample volume	RLU	CFU	RLU/CFU ratio
1 ml	6.5E+06	3.3E+05	20:1
2 ml	1.5E+07	6.3E+05	24:1

Flow cytometry analysis of dual RNA-seq samples

The cell proportions of the main cell types involved in TB disease were measured in the dual RNA-seq samples. Flow cytometry analysis of the dual RNA-seq samples was performed immediately after sample storage for RNA extraction, and therefore before the inoculation measurements were taken. This was to identify the changes in cell proportions across time in infected samples, to account for different cell proportions in samples when analysing the transcriptomics data, and to confirm cell death was not significant during the time course of the experiment. The cell types analysed in this experiment included two neutrophil subsets (CD14+CD11b- and CD14+CD11b-CD15+ neutrophils), NK cells (CD56+CD19-), two monocyte subsets (CD14+CD11b+), T cells (CD3+CD19-), and B cells (CD3-CD19+). For simplicity, cell subsets will be referred to by their cell names and not fluorophores used.

Within the innate immune cells, both the percentage of total and live CD11b+ monocytes remained low during 0 and 6 hrs post-infection, although still within the expected range of 2 – 4% of the total cell count (**Figure 5.21A**). At 24 hrs post-infection, however, these cells expanded in infected samples compared to uninfected samples. A similar trend was observed in the NK cell population, with a stable number at 0 and 6 hrs post-infection (**Figure 5.21B**). At 24 hrs these cells also expanded, but in uninfected samples compared to infected samples. CD11b- neutrophils took up a large proportion of the total cell population, as expected (**Figure 5.21C**). A slow decrease in CD11b- neutrophils was observed over the time-course, likely reflecting their short life cycle, however no significant difference was observed between infected and uninfected samples. The same trend was observed in CD15+ neutrophils, with a decrease in cell population over time (**Figure 5.21D**). However, over time the percentage of live CD15+ neutrophils varied significantly, with very few being present or detected at 0 hrs, to a significant increase by 6 hrs, followed by a decrease to extremely few/no live cells by 24 hrs.

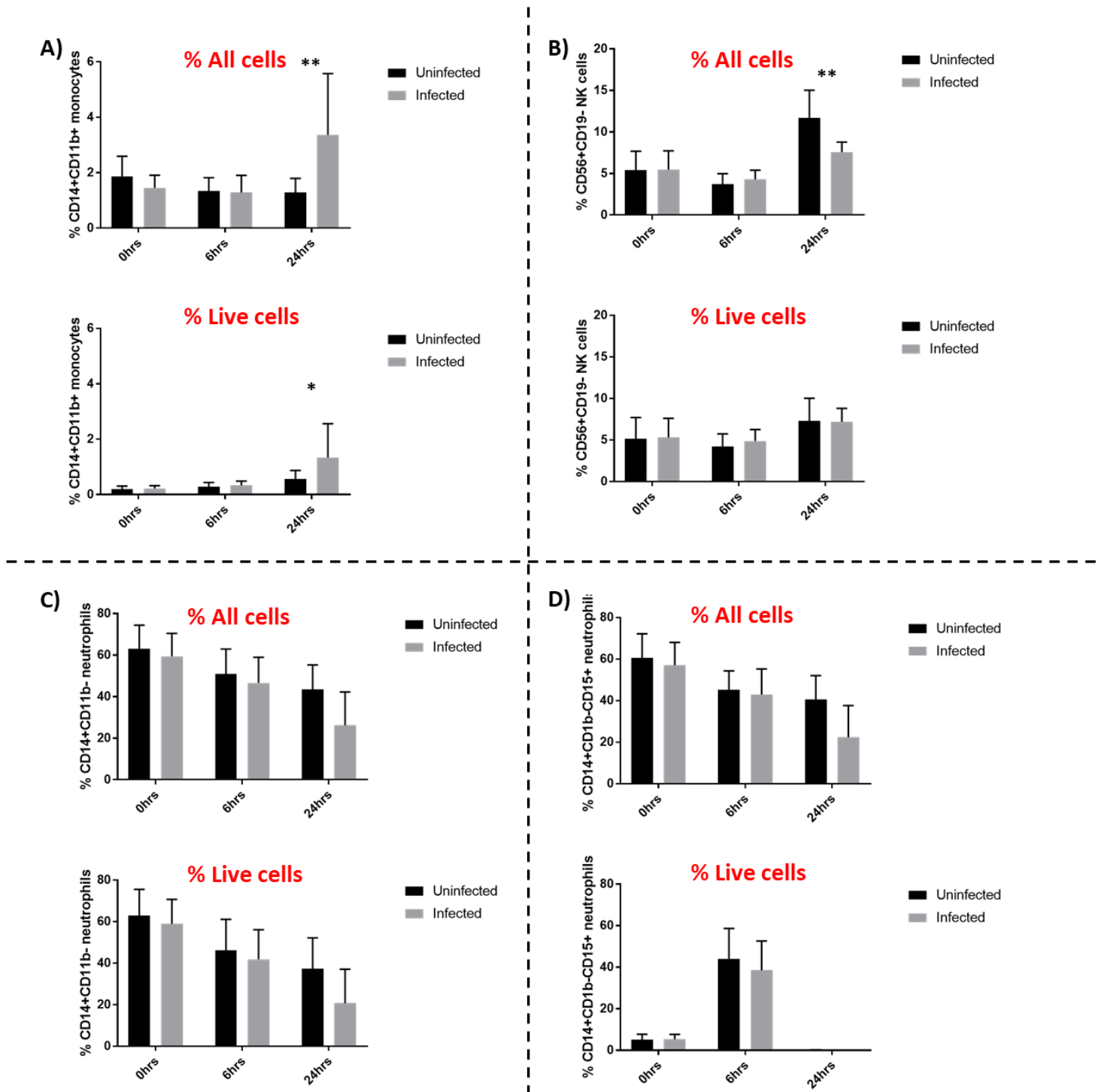


Figure 5.21: Innate immune cell population composition upon whole blood infection with BCG lux for the main dual RNA-seq experiment. A) Percentage of monocytes as a percentage of total cells and percentage of live monocytes from total monocytes in uninfected and infected samples at 0-, 6- and 24-hours post-infection. **B)** Percentage of NK cells as a percentage of total cells and percentage of live NK cells from total NK cells in uninfected and infected samples at 0-, 6- and 24-hours post-infection. **C)** Percentage of CD11b- neutrophils as a percentage of total cells and percentage of live CD11b- neutrophils from total CD11b- neutrophils in uninfected and infected samples at 0-, 6- and 24-hours post-infection. **D)** Percentage of CD15+ neutrophils as a percentage of total cells and percentage of live CD15+ neutrophils from total CD15+ neutrophils in uninfected and infected samples at 0-, 6- and 24-hours post-infection. N = 6. * = p-value < 0.05; ** = p-value < 0.01.

Within the adaptive immune cell populations, the percentage of all T cells and live T cells increased over time, but no difference was observed between uninfected and infected samples (**Figure 5.22A**). T cells consisted of approximately 30% of the total cell population which expanded to approximately 50% of the total cell population over time. The percentage of B cells, however, transiently increased at 6 hrs from approximately 2% to 7.5% before decreasing again closer to their initial percentage at 24 hrs (**Figure 5.22B**). The percentage of live B cells however remained relatively stable at approximately 2.5% over the time-course. B cells consisted of a small proportion of the total cell population except at 6 hrs where they rose to approximately 7.5%. No difference was observed in the B cell population between uninfected and infected samples.

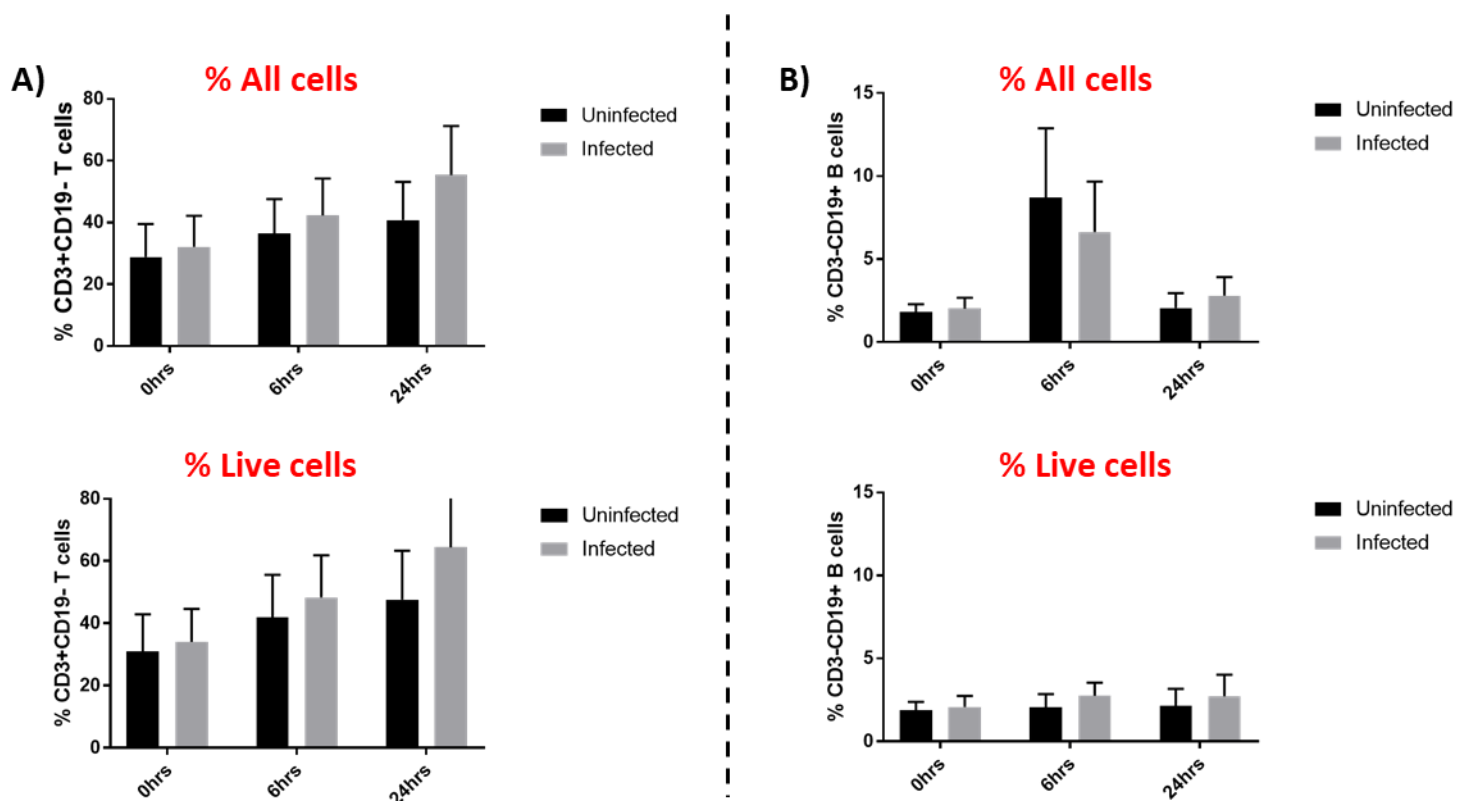


Figure 5.22: Adaptive immune cell population composition upon whole blood infection with BCG lux for the main dual RNA-seq experiment. A) Percentage of T cells as a percentage of total cells and percentage of live T cells from total T cells in uninfected and infected samples at 0-, 6- and 24-hours post-infection. **B)** Percentage of B cells as a percentage of total cells and percentage of live B cells from total B cells in uninfected and infected samples at 0-, 6- and 24-hours post-infection. *N* = 6.

5.3 Discussion

WBAs have shown great potential for the study of complex infections, such as TB. Here, an *in vitro* WBA was further characterised and optimised for its application using dual RNA-seq. This technique is a relatively new approach which was used here to analyse host-pathogen interactions between BCG *lux* and an amalgamation of host immune cells. To perform this dual RNA-seq experiment, several optimisation steps were required to ensure that the bacterial RNA was not over-diluted by host RNA and that the host immune cells survived across the time-course.

Through these optimisation steps, I identified an optimal protocol for preparing and extracting both host and pathogen RNA from BCG *lux* infected whole blood. Although increasing the MOI from 1:1 to 10:1 increased the proportion of bacterial RNA to host RNA, the flow cytometry analysis showed that this significantly impacted the number of monocytes present in the assay. While there was little difference in live monocyte populations, the higher infection rate presumably infected a greater number of monocytes resulting in greater death which lowered the number of overall monocytes at the 24 and 48-hour time-points. This large decrease in monocyte cell population would prevent any transcriptional changes from being observed as a result of infection of these cells or would likely only produce transcriptional signals reflecting cell death pathways, limiting our biological interpretation of the RNA-seq results. Furthermore, the number of granulocytes also decreased at the higher MOI compared to the lower MOI at the later time-points, although not significant. While the concentration of bacterial RNA appeared significantly lower, the following analyses were continued with a MOI of 1:1 as the level of cell death using a MOI of 10:1 would likely skew the transcriptomic results towards cell death signals not reflective of true infection.

In addition to the MOI, the flow cytometry results also indicated a decrease in cell survival throughout the assay time-course. Both the monocyte and lymphocyte populations expanded over time, however the granulocytes decreased significantly between 0 and 24 hours, likely reflecting the short lifespan of these cells. However, to prevent the identification of cell death signalling as the predominant transcriptomic response, earlier time-points up to 24 hours were chosen for future experiments.

Finally, the extraction procedure was also optimised to increase the proportion of bacterial RNA present in samples. The final extraction protocol included bead-beating all samples and applying DNase treatment. Despite low Nanodrop ratios and RNA concentrations, sufficient host RNA sequencing reads can be obtained using this protocol. The quantity of bacterial reads obtained however was still lower than estimated for dual RNA-seq (Westermann et al., 2012). However, this

was partly due to the large proportion of globin RNA sequenced which prevented further bacterial RNA from being sequenced. Furthermore, a higher read depth may have resulted in sequencing of sufficient reads, as lower quantities have previously proved sufficient (Westermann et al., 2016). Importantly, the inclusion of the bead-beating step did not degrade host or bacterial RNA and was in fact essential to increase the depth of bacterial RNA reads obtained, whilst also improving the depth of host RNA reads obtained.

Using these optimised approaches, a pilot dual RNA-seq experiment was performed which identified several biological pathways which changed over time in both the host and the bacterium. For host gene expression analysis, both bead-beat and no-bead-beat samples were used as biological replicates due to low sample numbers, and so it should be noted that additional genes captured by the bead-beat procedure will not be reflected in this analysis. In the host, there was large number of significantly up-regulated genes at the earlier time-points, which is expected after initial infection. These changes reduced significantly when comparing 11 hrs and 24 hrs, suggesting a lack of immune cell response during these time-points. It is likely that most of the innate immune response is measured in the samples prior to 11 hours and that further adaptive immune responses would require more than 24 hours to develop. Previous WBA studies have also observed this trend, with higher numbers of significantly differentially expressed genes at early (< 48 hours) and late (> 48 hours) time-points (von Both et al., 2018). However, it is likely that a proportion of these large gene expression changes at later time-points is a result of cell death due to *in vitro* incubation of whole blood, instead of representing changes in response to infection.

Of the host gene expression changes, many genes involved in the immune response and inflammation were significantly up-regulated at the earlier time-points, as observed previously (von Both et al., 2018). Significantly down-regulated genes were also involved in the immune response including apoptosis and pre-B cells. Some of the top differentially expressed genes were also down-regulated in previous studies, however usually at later time-points of around 30 hours. At the later time-point there was an overall up-regulation of immune related genes such as *IFNG*, which encodes IFN γ and is one of the most important pathways for TB disease, but also genes not directly involved in the immune response. Interestingly, specific genes such as *IL1A* were confirmed as significantly up-regulated at early time-points in a previous microarray WBA Mtb infection model, but other genes such as *CXCL10* were not identified in the entire dataset (von Both et al., 2018). This discrepancy could be due to the difference in infecting with BCG *lux* compared with Mtb, despite this previous publication indicating an overall correlation in expression changes between the two bacteria, or it could also reflect the

differences in microarray and RNA-seq technology and the advantage of using RNA-seq to measure all transcripts present in a sample. This highlights the importance of using RNA-seq, where the most significantly up-regulated gene, *CXCL10*, was potentially missed using microarray.

In contrast to the host response, the bacterial response could not be measured using differential gene expression analysis using DESeq2. The non-bead-beat samples produced very few reads compared to the bead-beat samples, and so could not be used as biological replicates. As a result, no significance analysis was performed for bacterial gene expression changes and instead log₂ fold change values were calculated between the same time-point comparisons. Unfortunately, this resulted in difficulties estimating if the extraction procedure produced sufficient read depth for differential gene expression analysis, if biological replicates were present. Overall, the genes which were differentially expressed were from a combination of pathways including an up-regulation of genes involved in the response to infection and down-regulation of genes involved in cell metabolism at early time-points, and an up-regulation of iron transport, growth, and cell envelope biosynthesis and down-regulation of immune defence related genes at later time-points. While these trends were identified, many of the top differentially expressed transcripts were uncharacterised, resulting in limited interpretation.

While this experiment provided interesting insights into the changes in gene expression which occur in the host and bacteria over time, the obvious limitation is the lack of biological replicates. This not only prevented a more in-depth differential gene expression analysis from being performed, but also prevented the deconvolution of transcriptional changes due to differences in cell populations between samples as additional replicates are required for this. The findings warrant validation in further analyses. A more extensive dual RNA-seq experiment involving 6 donors for both infected and uninfected samples was planned to take place during this PhD however the COVID-19 pandemic prevented this from proceeding. Nevertheless, the optimisation steps of the WBA in this study provide the basis of a workflow to use the WBA for future dual RNA-seq experiments.

In addition, while this study primarily uses BCG *lux* as a surrogate strain for Mtb infection, there may be some differences in infection between these bacteria. While BCG differs from Mtb in several ways, one of the most prominent differences is the presence of the TbD1 region in BCG compared with 'modern' *Mycobacterium tuberculosis* strains, as well as almost all BCG strains having a SNP in the TbD1 region at codon position 551 of the *mmpL6* gene which differentiates these strains from *M. canettii*, *M. africanum*, and ancient *Mycobacterium tuberculosis* strains (Brosch et al., 2002). BCG also contains RD7, RD8, RD9 and RD10 deletions, distinguishing BCG strains from the previously mentioned

MTBC strains. In addition to these differences, BCG also lacks the ESX-1 secretion system, a key virulence factor in *Mtb* strains which is used to break down the phagosome membrane when mycobacteria are phagocytosed by macrophages (Nadolinska et al., 2020). Therefore, there are certain key virulence genes which are absent in BCG. Previous studies have also shown that there are some differences in the transcriptome of THP-1-infected macrophages when infected with H37Rv compared with BCG (Pu et al., 2021). In contrast, Newton, et. al., showed that the majority of host genes were concordantly expressed in BCG-infected whole blood compared with whole blood infected with H37Rv using the WBA (von Both et al., 2018). Overall, the majority of TB disease pathways can be measured using BCG in this system, but certain pathways are not represented. As the WBA has been previously validated for infection with *Mtb*, and the optimisation of the assay for dual RNA-seq has been performed here, an additional experiment using *Mtb* would be beneficial to understand what findings described here translate to TB disease.

Another possible limitation, although not tested, was the depth of bacterial RNA reads obtained. Previous studies suggested a minimum of $2 - 5 \times 10^6$ bacterial RNA reads after rRNA depletion for dual RNA-seq analysis (Westermann et al., 2012). Here, only 3×10^4 bacterial reads were obtained from the pilot dual RNA-seq experiment. This suggests that there were too few reads for differential gene expression analysis. Here we found that many transcripts were identified with high $\log_2$ fold changes (absolute 2) and therefore it is likely that some of these findings would be significant when using biological replicates. It should also be noted that the significant sequencing of the globin beta subunit gene, *HBB*, will have affected the number of bacterial reads sequenced. Therefore, appropriate removal of globin RNA would have alleviated this problem to a degree.

Overall, this chapter provides the basis for future work investigating host-pathogen interactions using the WBA and a dual RNA-seq approach. This is the first time whole blood has been used to model host-pathogen interactions using a dual RNA-seq approach, allowing for complex interactions between different immune cells to be measured over time. Further work should increase the number of replicates and donors used, however should limit the time-course to early time-points and therefore likely focus on more innate immune responses to infection. Furthermore, now that these approaches have been established in a category level 2 laboratory, a *lux* strain of *Mtb* should be used instead of BCG *lux* in a BSL 3 laboratory and will produce data which are more relevant for TB disease.

5.4 Appendix 1: FASTQC report

FastQC Report

Wed 25 Sep 2019
WTCHG_723588_73695153_1.fastq.gz

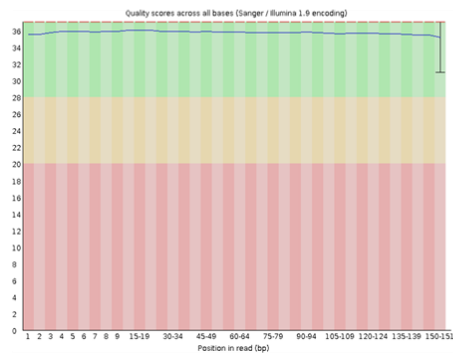
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

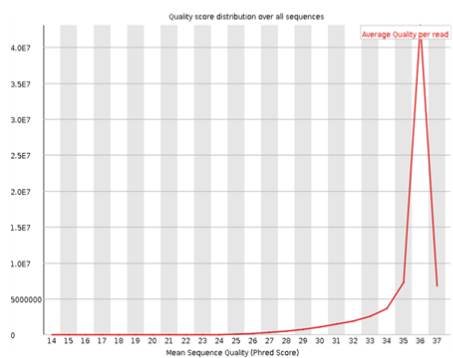
Basic Statistics

Measure	Value
Filename	WTCHG_723588_73695153_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	70089377
Sequences flagged as poor quality	0
Sequence length	151
%GC	56

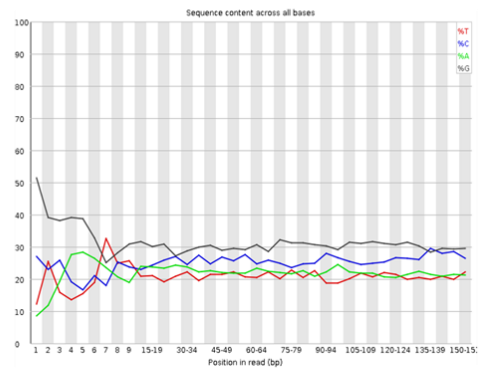
Per base sequence quality



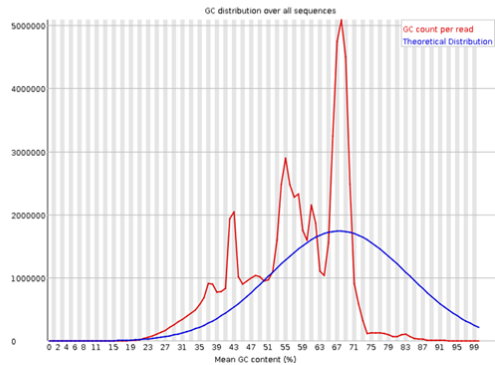
Per sequence quality scores



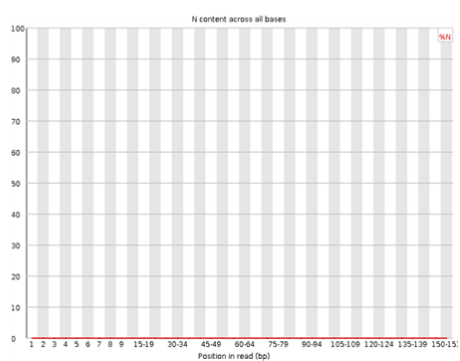
Per base sequence content



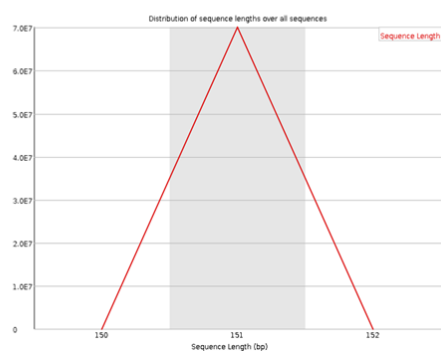
Per sequence GC content



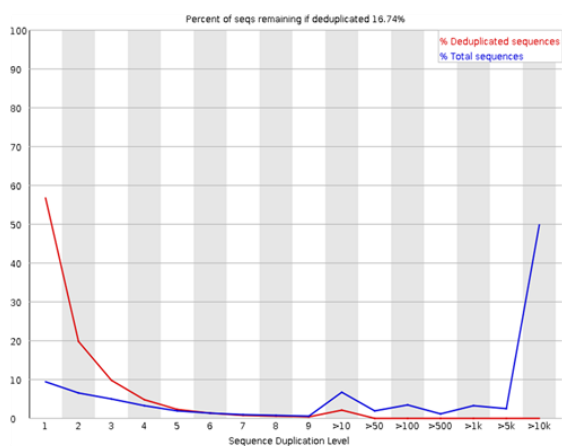
Per base N content



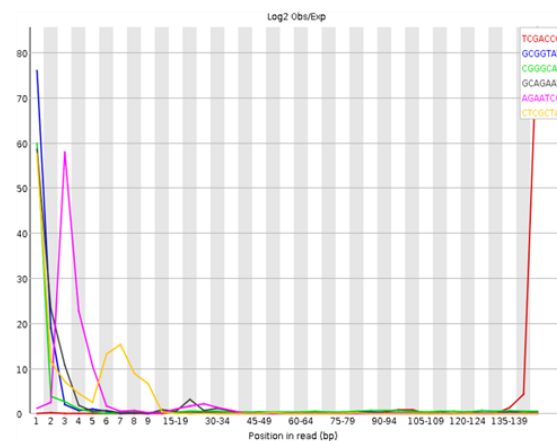
Sequence Length Distribution



Sequence Duplication Levels



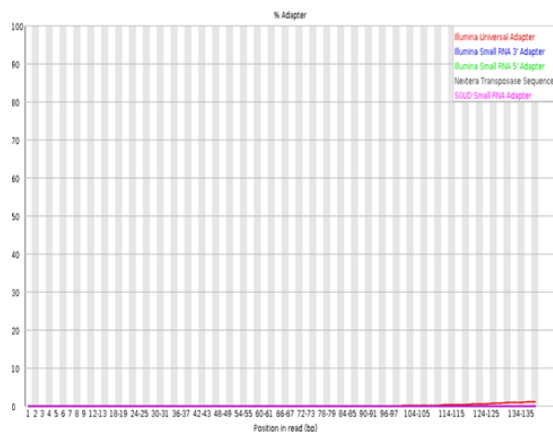
Kmer Content



Overrepresented sequences

Sequence

Sequence	Count	Percentage	Possible Source
GCAGAAATCCAGATGCTCAAGGCCCTTCATAATATCCCGAGTTAGTAGT	1294802	1.8473584092493789	No Hit
GTCAGCACGGTGCTACAGAAGCCAGGAATTTGTCAGGAGGCGTGAC	859659	1.2265182496913904	No Hit
GTGAATCTTTGCCAAAGTGATGGGCCAGCACACAGCCAGCACGTTGCC	657260	0.9377455302534649	No Hit
GTGAATCGGCGGGAGGTGGCGGCCAGGGTCACAGAGGAGGAGTGCT	636069	0.9075112766375423	No Hit
GTGGCTTAGGAGCTTGAAGTTGACCGGTCACCGAAGCTTGTGCGCGT	622978	0.8888336958680628	No Hit
GGCAGAAATCCAGATGCTCAAGGCCCTTCATAATATCCCGAGTTAGTAG	511443	0.7297011642720123	No Hit
GCGAGCTTAGTGATACTTGTGGGCCAGGGCATTAGCCACACAGCCACCA	478670	0.6829422952354106	No Hit
GGCATGTGTCACAGTGCGCCAGCGGTTGGTCAGCGGTCGCGCACCTT	469079	0.6692583385353817	No Hit
GGAGGTACGACGGTGCTCAGAGAAGCCAGGAATTTGTCAGGAGGCGT	455740	0.6502269238318383	No Hit
GGGAGGTGGCGGCCAGGGTCACAGCAGGAGGAGTGGCTTAGGAGCTTGA	439391	0.6269009924285673	No Hit
GCTTGAAGTTGACCGGGTCCACCGAAGCTTGTGCGGTCAGGTCGCTC	434789	0.6203350901521067	No Hit



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TCGACCG	24435	0.0	85.51644	145
GCGGTAT	8860	0.0	76.120255	1
CGGGCAT	28280	0.0	59.95372	1
GCAGAAAT	336805	0.0	58.621414	1
AGAATCC	341940	0.0	57.95368	3
CTCGCTA	23150	0.0	57.60786	1
GAATCCA	349695	0.0	57.0931	4
CAGAAATC	347035	0.0	56.828175	2
AATCCAG	357250	0.0	55.85122	5
ATCCAGA	358660	0.0	55.688248	6
TCGCTAT	24130	0.0	55.357975	2
TCCAGAT	366290	0.0	54.42512	7
CGCTATG	25090	0.0	53.37546	3
CCAGATG	376055	0.0	53.060062	8
CAGATGC	381345	0.0	52.335682	9
GCGCGGT	11205	0.0	51.646637	1

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GACGGTA	5705	0.0	43.981644	1
TGCGAGC	20060	0.0	43.01967	1
CCGGTAT	5440	0.0	40.258636	1
CGGAGGT	33900	0.0	40.19556	1

Figure S5.1: Example FASTQC report from the pilot dual RNA-seq experiment.

5.4 Appendix 2: Comparing host-pathogen interactions in an *in vitro* whole blood infection assay with *in vivo* Tuberculosis infection

5.4.1 Background

As the COVID-19 pandemic began, my dual RNA-seq analyses were halted. With several unknowns surrounding the completion of my laboratory work, I instead designed a novel approach to understanding host-pathogen interactions of the WBA. While the WBA has been used to analyse several aspects of TB disease (Martineau et al., 2012; Newton et al., 2011; von Both et al., 2018), there have not been any validations of how representative the assay is with true *in vivo* infection. Therefore, while the WBA is likely to accurately reflect certain aspects of *in vivo* infection, full characterisation should be performed and findings in previous studies should be carefully considered. Furthermore, another confounding factor to the interpretation of analysis of the WBA is the composition of each sample, particularly over time. Previous publications using the WBA show little or no estimation of the composition of immune cells, despite predicting their function in bulk. It is possible that certain immune cells are driving particular host responses and being able to differentiate between cells and responses would benefit the understanding of the analysis.

In contrast to the lack of immune cell composition analysis, WBA studies have frequently used the assay as a time-course to understand changing host responses over time (Newton et al., 2011; Roy et al., 2019). This is particularly important for TB as there are several different presentations of the disease, including latent and active disease, which can develop over time. However, the frequently used time-points of the WBA have not been fully characterised or directly compared to different stages of *in vivo* infection. In addition, the WBA represents a unique model where different aspects of TB disease can be studied using different donors (Bok et al., 2021). The most frequent use of the assay is with healthy adult donors which are then infected with Mtb *in vitro*. However, this experimental design may not reflect chronic *in vivo* infection, and instead reflect a new acute infection. This aspect of the assay has again not been characterised.

Therefore, the aims of this chapter were to identify the key differences in gene expression changes in the *in vitro* WBA compared to *in vivo* TB infection. Furthermore, the biological pathways which are

accurately represented, and those which are not, in the WBA with TB infection are identified. This will allow further analyses to focus on aspects of the host response to infection which are known to be reliable compared to true infection. This analysis addressed an important gap in the research landscape while building on the interpretation of dual RNA-seq results. This project was conceived by myself and Dr Myrsini Kaforou, and analyses were performed by Petra Bachanová, a BSc student.

5.4.2 Results – Comparison of cohort datasets

Several publicly available datasets were used in this analysis. An *in vitro* WBA dataset using whole blood from adult donors which was infected with Mtb was compared with two *in vivo* datasets from adult and paediatric patient cohorts (**Figure S5.4.2**). The *in vitro* dataset included bulk gene expression data for 6-, 24-, 48-, 72-, and 96-hour time-points for infected samples and bulk gene expression data for 0-, 6-, 48-, 72-, and 96-hour time-points for uninfected samples. The *in vivo* dataset included bulk gene expression data from adults and children diagnosed with active TB or latent infection from South Africa, Kenya, and Malawi. In these datasets, HIV uninfected patients were used and those infected with HIV were excluded.

Before comparing the *in vitro* and *in vivo* datasets, sample pre-processing, differential gene expression analysis, and pathway analysis were performed on the datasets individually. While this was performed on these datasets in their original publications, the bioinformatics workflow will be tailored to the individual datasets and may not be appropriate for comparisons across datasets. Therefore, a standardised bioinformatic workflow was used here.

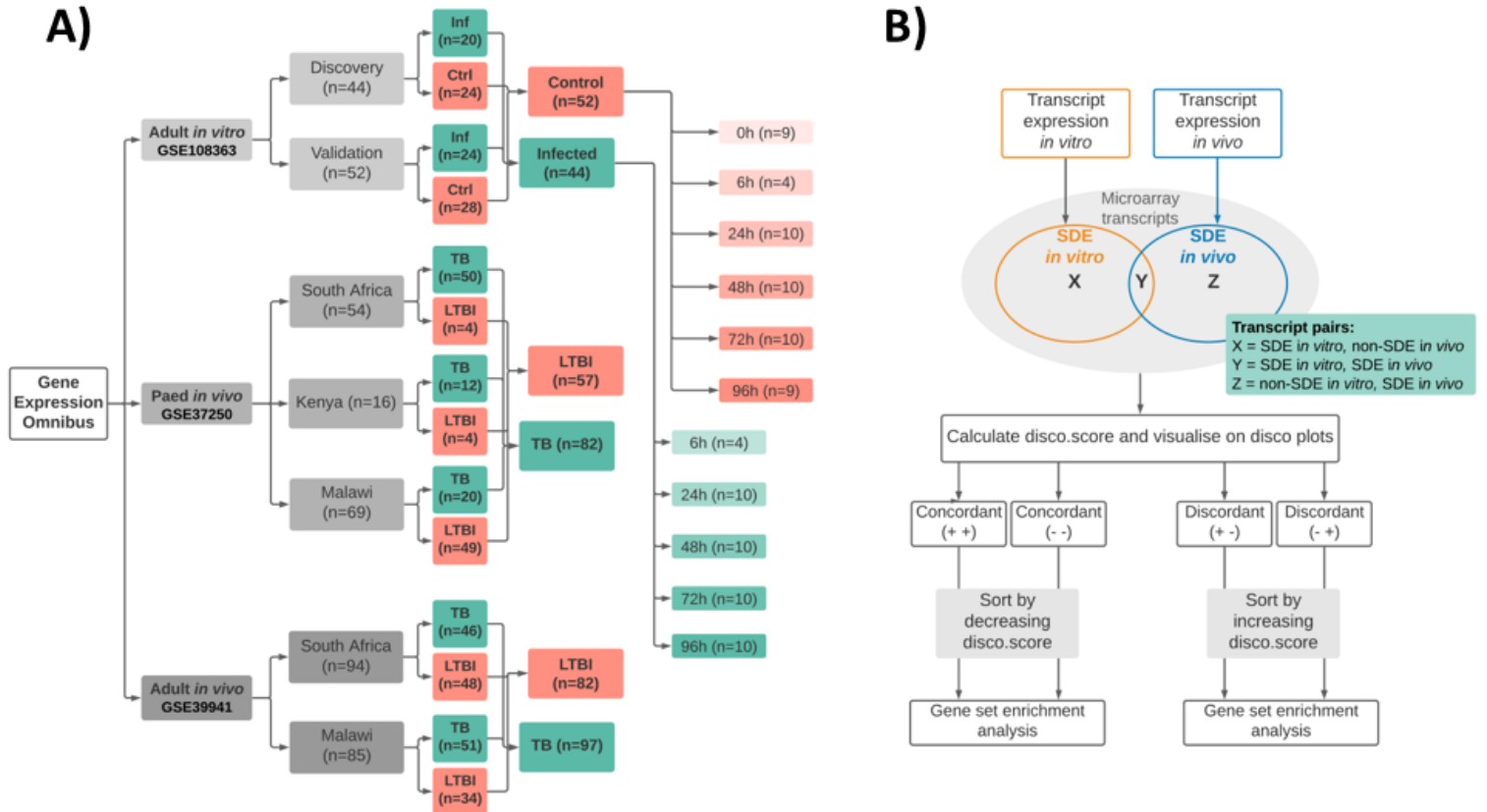


Figure S5.4.2: Concordance and discordance analysis study design. A) Publicly available datasets from GEO which were downloaded and used in this analysis. Samples extracted from these datasets, including time-points used, are described. Inf = *Mtb* infected samples; Ctrl = uninfected samples; TB = active TB patients; LTBI = latent TB patients. **B)** Concordance and discordance bioinformatic analysis workflow. Differential gene expression analysis was performed for each dataset individually before identifying transcript ‘unions’. Transcript unions comprise of genes identified as differentially expressed in one dataset (Y), being selected across all datasets even if not significantly differentially expressed (X and Z) and used for concordance and discordance analysis. Concordance and discordance analysis was performed using the disco.score and visualised using disco plots. Concordant genes were those with the same $\log_2FC$ expression direction (e.g. positive) in both datasets which can be further grouped into concordantly upregulated genes (where $\log_2FC$ expression is positive in both datasets) and concordantly downregulated genes (where $\log_2FC$ expression is negative in both datasets). In contrast, discordant genes were those with opposite $\log_2FC$ expression direction across datasets (e.g. positive $\log_2FC$ expression in one dataset but negative $\log_2FC$ expression in another dataset). Gene Set Enrichment Analysis was then performed from the discordant and concordant genes. SDE = significantly differentially expressed genes.

5.4.3 Results – Pre-processing of *in vitro* and *in vivo* datasets

Both datasets were pre-processed as described in **Chapter 2.4.9**, to ensure that datasets were treated with the same steps and scaled similarly. To identify sample outliers, and to ensure pre-processing had been performed correctly, the samples were visualised using PCA. The *in vitro* dataset formed 2 main clusters which grouped Mtb-infected samples and uninfected samples separately (**Figure S5.4.3A**). This corresponded to PC2, meaning this clustering was only the second greatest source of variation in the dataset. The greatest source of variation resulted from the different time-points of the assay (**Figure S5.4.3B**). Time-points clustered sequentially, with the greatest distance being observed between 6 and 24 hours, and 24 and 48 hours, suggesting the greatest transcriptional changes occurred at these intervals. The paediatric *in vivo* dataset showed 2 main clusters corresponding to active TB and latent TB patients (**Figure S5.4.3C**). While these clusters demonstrated slight overlap, they were mainly distinct from each other. The adult *in vivo* dataset showed a similar pattern, with 2 main clusters, although there was minimal overlap (**Figure S5.4.3D**). In this dataset, however, 3 LTBI samples appeared to be misclassified as they clustered within the active TB group.

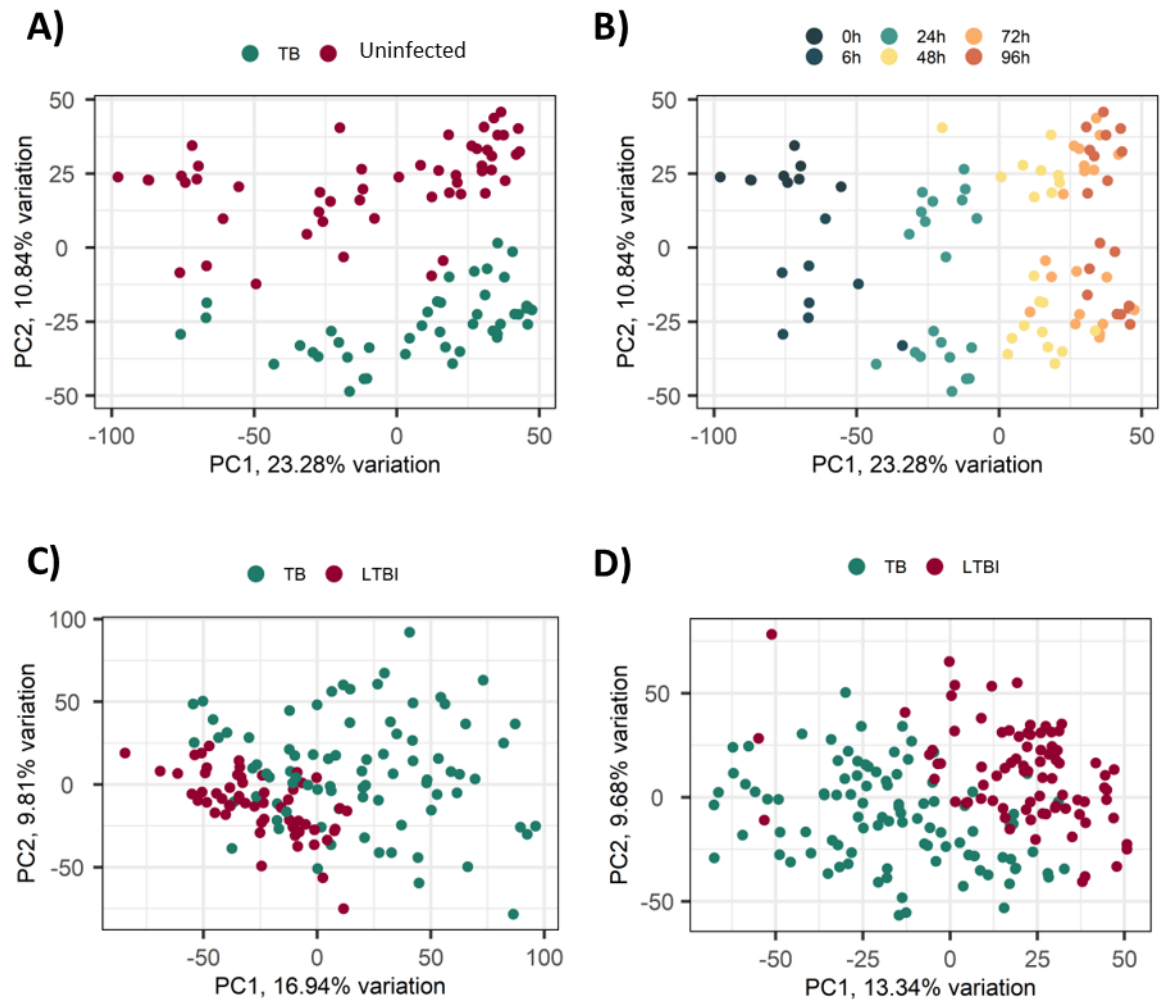


Figure S5.4.3: PCA of in vitro and in vivo datasets after pre-processing. **A)** PCA plot of in vitro dataset coloured by infection status. TB = *Mtb* infected samples. **B)** PCA plot of in vitro dataset coloured by time-point. **C)** PCA plot of paediatric in vivo dataset coloured by infection status. TB = active TB patients; LTBI = latent TB patients. **D)** PCA plot of adult in vivo dataset coloured by infection status. TB = active TB patients; LTBI = latent TB patients.

5.4.4 Results – Differential gene expression analysis of the individual datasets

Differential gene expression analysis of the *in vitro* dataset was performed on the 6-, 24-, 48-, 72-, and 96-hour time-points of Mtb-infected whole blood samples and was compared to uninfected controls at the same time-points. The number of significantly differentially expressed genes identified for each time-point can be seen in **Table S5.4.4**. In this dataset, the majority of genes were downregulated at 24 hours post-infection (65.2% of genes), 48 hours post-infection (63.6%), and at 72 hours post-infection (69.2%). The expression profile of each time-point can be seen in **Figure S5.4.4**.

Table S5.4.4: Significantly differentially expressed genes in the *in vitro* and *in vivo* datasets. Adj *p*-val = adjusted *p*-value.

Dataset	Significantly expressed transcripts (adj <i>p</i> -val < 0.05)
<i>In vitro</i> 6 h vs 0h	45
<i>In vitro</i> 24 h vs 0h	2,152
<i>In vitro</i> 48 h vs 0h	1,767
<i>In vitro</i> 72 h vs 0h	1,476
<i>In vitro</i> 96 h vs 0h	2,093
Paediatric <i>in vivo</i> TB vs LTBI	13,051
Adult <i>in vivo</i> TB vs LTBI	10,007

Differential gene expression analysis of the *in vivo* datasets was performed to compare active TB patients with latent TB patients, for adult and paediatric cohorts individually. Previous gene expression analyses comparing active TB, LTBI, and healthy controls has demonstrated that the expression profile of latent TB patients is similar to healthy controls (Broderick et al., 2021). Latent TB patients were therefore used as a proxy for healthy or uninfected controls in this analysis. The number of significantly differentially expressed genes in these datasets can be seen in **Table S5.4.4**. In contrast to the *in vitro* dataset, the majority of genes in the *in vivo* dataset analyses were found to be upregulated in active TB patients (66.0% in the paediatric cohort and 62.6% in the adult cohort). The expression profiles of the comparisons can be seen in **Figure S5.4.4**.

Pathway analysis was performed on the genes which were significantly up or downregulated in each dataset using Gene Ontology (GO) analysis. For the *in vitro* dataset, immune activation, antigen processing, and defence response to bacteria were enriched in downregulated genes in the 24 – 96-hour time-points (Appendix 3, **Tables S5.4.11 and S5.4.12**). In contrast, lymphocyte chemotaxis, neutrophil chemotaxis, and cellular response to IL-1 were pathways enriched in upregulated genes for the same time-points. For the *in vivo* dataset, regulation of cytokine production, cytoskeleton organisation, and Fc receptor signalling were pathways enriched in downregulated genes while defence response to bacteria, innate immune response in mucosa, and positive regulation of IL-1B were pathways which were enriched with upregulated genes (Appendix 3, **Tables S5.4.13 and S5.4.14**).

5.4.5 Results – Comparison of gene expression signatures between *in vitro* and *in vivo* datasets

To compare the datasets with each other, concordance and discordance analysis was performed. To do this, the genes which were significantly differentially expressed in at least one dataset were selected from all datasets and described as gene ‘unions’. These unions were used for disco.score analysis (**Figure S5.4.5**). The discordance-concordance plots were separated into quadrants for each dataset comparison which corresponded to concordant upregulation (quadrant I), concordant downregulation (quadrant III), discordant as genes were upregulated *in vivo* but downregulated *in vitro* (quadrant II), and discordant as genes were downregulated *in vivo* but upregulated *in vitro* (quadrant IV). When comparing the *in vitro* dataset with the adult *in vivo* dataset, the top concordant genes included *CARD17*, *SOD2*, and *CASP5*. Concordant downregulated genes included *CCR3* and *DCANP1*. Discordant genes included *VPREB3*, *SLAMF1*, and *CCR7*.

Pathway analysis was performed for genes in each of the described quadrants for each *in vitro-in vivo* comparison using Gene Set Enrichment Analysis. Gene Set Enrichment Analysis identified gene modules, which are counts of the identified gene sets, which can be seen in **Figure S5.4.6**. Earlier time-points (6 – 48 hours) presented with a greater amount of concordance, while later time-points (48 – 96 hours) presented with a greater amount of discordance. The comparison of the *in vitro* dataset with the paediatric *in vivo* dataset presented with a greater number of concordantly downregulated modules (**Figure S5.4.6C**) than the comparison of the *in vitro* dataset with the adult *in vivo* dataset, which presented with a greater number of concordantly upregulated modules (**Figure S5.4.6A**). The majority of modules were found to be discordantly upregulated in the *in vitro* dataset and downregulated in the *in vivo* dataset.

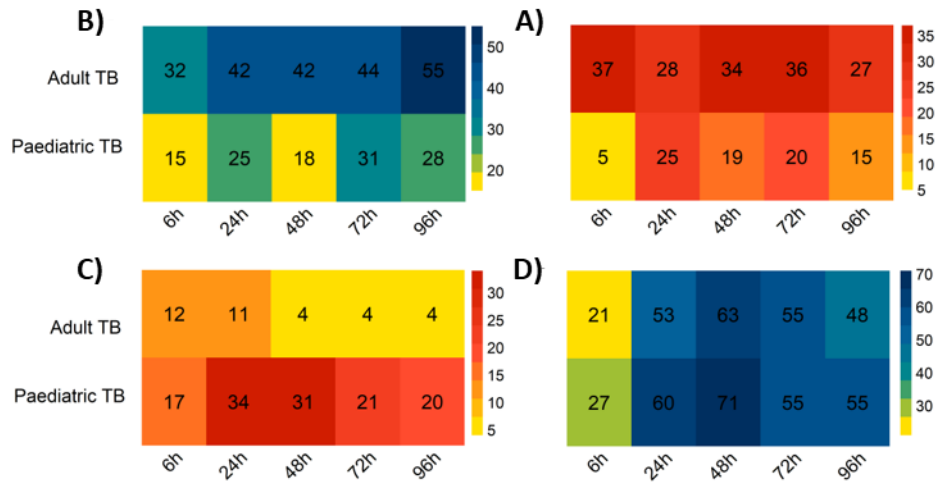


Figure S5.4.6: Concordant and discordant gene module counts. Plots are presented in the same order as the disco plots, i.e. with QII, QI, QIV, then QIII. **A)** Counts of concordantly upregulated gene modules (QI). **B)** Counts of discordantly regulated gene modules, with modules being upregulated in the in vivo datasets and downregulated in the in vitro dataset (QII). **C)** Counts of concordantly downregulated gene modules (QIII). **D)** Counts of discordantly regulated gene modules, with modules being downregulated in the in vivo datasets and upregulated in the in vitro dataset (QIV). Concordant modules show increased concordance with increased intensity of red, whereas discordant modules show increased discordance with increased intensity of blue.

5.4.6 Results – Concordance and discordance analysis of the *in vitro* and adult *in vivo* datasets

To explore the concordant and discordant gene modules further, gene modules were summarised by similar biological pathways and visualised through dot plots (**Figures S5.4.7 and S5.4.8**). Concordant pathways identified in the comparison of the *in vitro* and adult *in vivo* datasets were predominantly found in upregulated modules (**Figure S5.4.7B**, corresponding to quadrant I), confirming our previous analysis. The most upregulated gene module was inflammation, with an average effect size of 12.8 (average p -value of 4.36×10^{-22}) across all time-points. Other highly upregulated concordant gene modules included monocyte enrichment (average effect size of 9.34 and average p -value of 4.57×10^{-11}), chemokines (average effect size of 6.67 and average p -value of 4.7×10^{-7}), and cytokines, specifically the type I IFN response (average effect size of 2.98 and average p -value of 4.6×10^{-4}).

In contrast, few concordantly downregulated gene modules were found in this comparison (**Figure S5.4.7C**, corresponding to quadrant III). Of the concordantly downregulated gene modules, T and B cell modules were the strongest in comparisons with 6 – 24-hour time-points (an average effect size of 14.5 and an average p -value of 6.1×10^{-13} for the 6-hour time-point and an average effect size of 17.95 and an average p -value of 5.09×10^{-18} for the 24-hour time-point). Other pathways which were identified as concordantly downregulated included monocyte-related pathways although this was significantly lower in other modules compared to the T cell modules described.

Discordant summarised gene modules which are upregulated in the *in vitro* dataset but downregulated in the adult *in vivo* dataset can be seen in **Figure S5.4.7D** (corresponding to quadrant IV). Discordance was found to be the highest in T and B cell modules and increased over time (T cells effect size of 17.54 and p -value of 8.27×10^{-25} at 6 hours compared with an effect size of 46.87 and a p -value of 1.67×10^{-94} at 24 hours; B cells effect size of 2.13 and a p -value of 0.025 at 6 hours compared with an effect size of 23.5 and a p -value of 8.56×10^{-39} at 96 hours). The NK cell gene modules also showed an increase in discordance between the 2 datasets, with the greatest effect size of 17 and a p -value of 2.3×10^{-27} . Other discordantly regulated modules included those involved in lymphocyte activation and function.

Discordant summarised gene modules which are upregulated in the adult *in vivo* dataset but downregulated in the *in vitro* dataset can be seen in **Figure S5.4.7A** (corresponding to quadrant II). These included monocyte modules (an average effect size of 9.09 and an average p -value of 2.87×10^{-11}).

⁴⁴⁾ and inflammation modules (an average effect size of 11.08 and an average p-value of 6.38×10^{-26}) which presented with the greatest effect sizes, as well as dendritic cells and antigen presentation modules which presented with lower effect sizes.

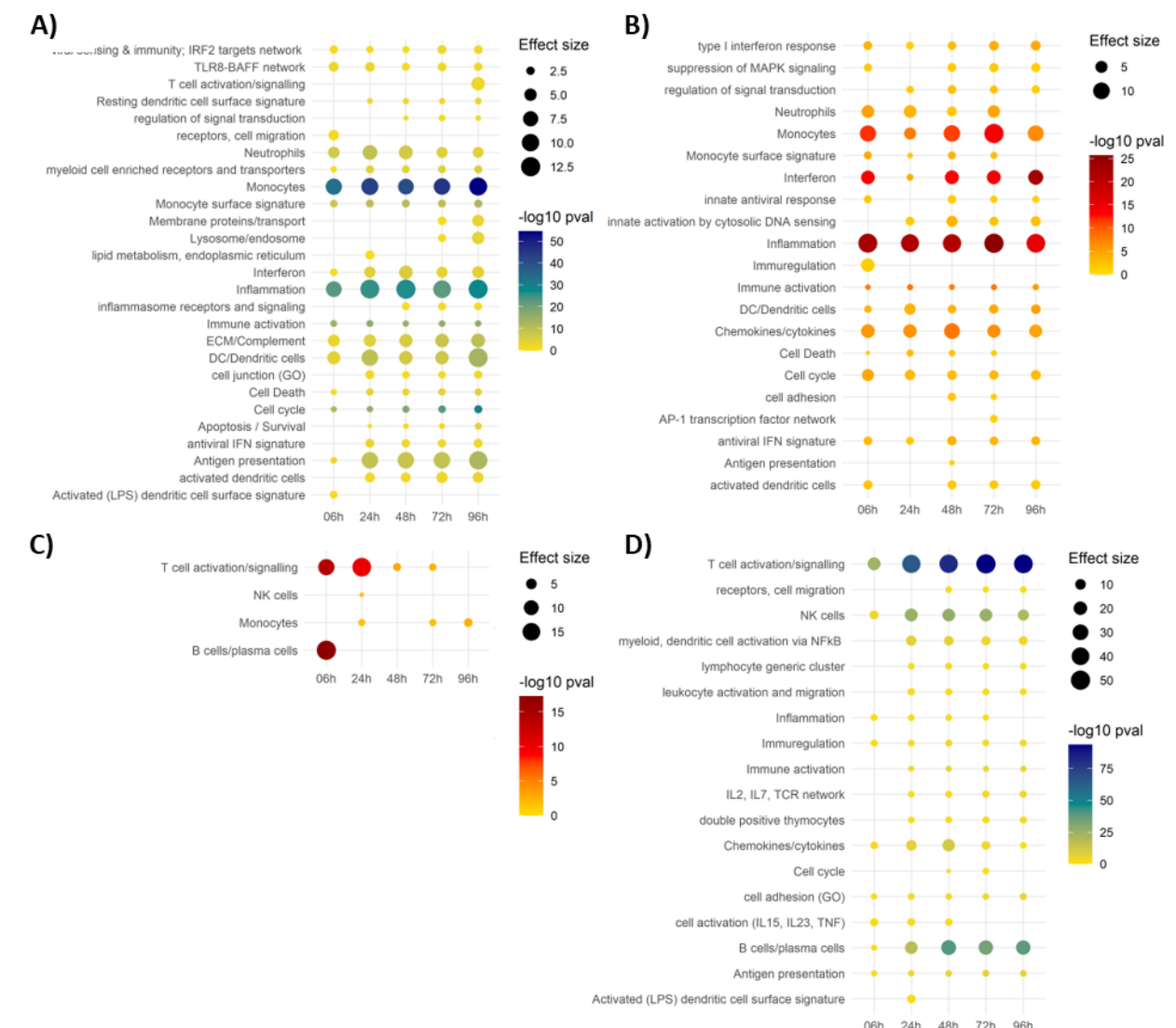


Figure S5.4.7: Dot plots of summarised gene modules from discordance and concordance analysis comparing *in vitro* with the adult *in vivo* dataset. A) Discordant gene modules which are upregulated in the *in vivo* dataset and downregulated in the *in vitro* dataset. B) Concordantly upregulated gene modules. C) Concordantly downregulated gene modules. D) Discordant gene modules which are downregulated in the *in vivo* dataset and upregulated in the *in vitro* dataset. Pval = p-value.

5.4.7 Results – Concordance and discordance analysis of the *in vitro* and paediatric *in vivo* datasets

Similarly, to the previous section, pathway analysis was performed using quadrants from the disco plots for the comparison of the *in vitro* and paediatric *in vivo* datasets and gene modules summarised for ease of interpretation. Concordantly upregulated modules can be seen in **Figure S5.4.8B** (corresponding to quadrant I). The gene modules for chemokines and cytokines showed the greatest concordant upregulation with an average effect size of 11.2 and an average p-value of 7.43×10^{-8} , followed by the inflammation module (with an average effect size of 5.09 and an average p-value of 3.01×10^{-8}) then the monocyte module (with an average effect size of 4.07 and an average p-value of 1.28×10^{-3}).

There were significantly more concordantly downregulated modules in this comparison (**Figure S5.4.8C**, corresponding to quadrant III) than in the comparison of the *in vitro* dataset with the adult *in vivo* dataset. Concordantly downregulated gene modules included T cell activation which peaked in effect size at 6 hours (with an effect size of 25.4 and a p-value of 3.1×10^{-16}). Other gene modules which were concordantly downregulated included those involved in monocytes and antigen presentation.

Discordant modules which were upregulated in the *in vitro* dataset and downregulated in the paediatric *in vivo* dataset can be seen in **Figure S5.4.8D** (corresponding to quadrant IV). These modules were mainly involved in T cell signalling which was also observed in the *in vitro* comparison with the adult *in vivo* dataset. In contrast, the B cell modules observed in the previous comparison was less pronounced in this comparison with the paediatric dataset with an average effect size of 5.24 and an average p-value of 0.003. Further discordant modules included NK cells and chemokine and cytokine regulation modules.

Discordant modules which were upregulated in the paediatric *in vivo* dataset and downregulated in the *in vitro* dataset can be seen in **Figure S5.4.8A** (corresponding to quadrant II). These modules included monocytes, which was also observed in the *in vitro* comparison with the adult *in vivo* dataset, and inflammation and dendritic cell modules. There was an overall increase in effect size of these modules at later time-points.

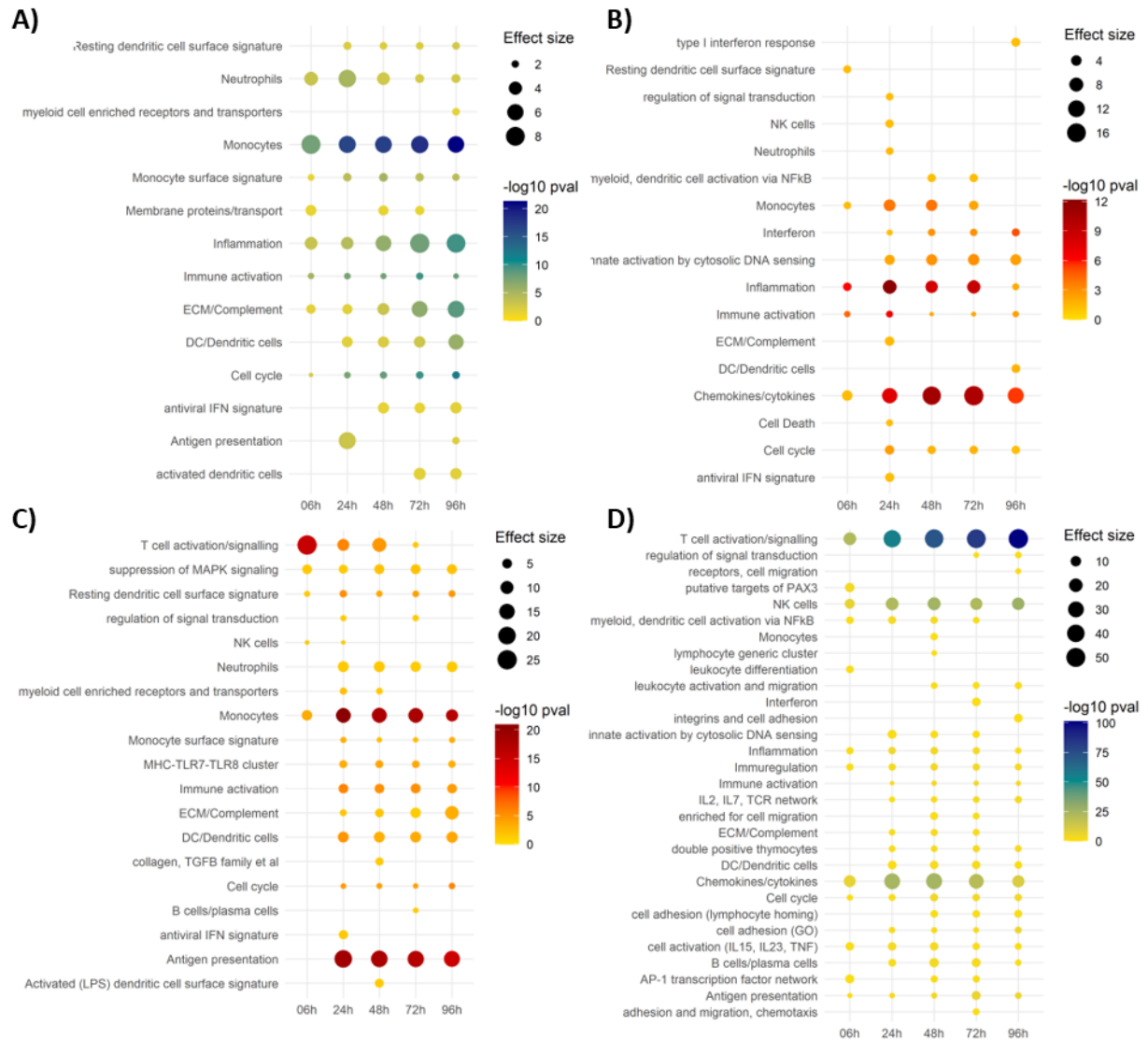


Figure S5.4.8: Dot plots of summarised gene modules from discordance and concordance analysis comparing *in vitro* with the paediatric *in vivo* dataset. A) Discordant gene modules which are upregulated in the *in vivo* dataset and downregulated in the *in vitro* dataset. B) Concordantly upregulated gene modules. C) Concordantly downregulated gene modules. D) Discordant gene modules which are downregulated in the *in vivo* dataset and upregulated in the *in vitro* dataset. Pval = p-value.

5.4.8 Results – Concordance and discordance analysis of *in silico* cell populations from *in vitro* and *in vivo* datasets

In addition to disco.score analysis, *in silico* flow cytometry analysis was also performed using the gene expression data from each dataset to identify similarities and differences in cell populations. *In silico* flow cytometry analysis was performed using CibersortX (Newman et al., 2019) and the results were summarised into innate and adaptive immune cell types.

Within the adult and paediatric *in vivo* datasets, the proportion of eosinophils, M1 and M2 macrophages, mast cells, and activated NK cells were low and showed no differences between active TB and latent TB patients (**Figures S5.4.9 and S5.4.10**). In both datasets, M0 macrophages, monocytes, and neutrophils were found to be higher in active TB patients compared to latent TB patients. In contrast, resting NK cells were lower in active TB patients compared to latent TB patients in both datasets. In addition, in the paediatric *in vivo* dataset, activated dendritic cells and resting mast cells were increased in active TB patients compared to latent TB patients.

When investigating innate immune cells in the *in vitro* dataset, activated dendritic cells were increased in infected samples compared with uninfected samples from 24 hours onwards (**Figure S5.4.9**), similar to the paediatric *in vivo* dataset. These cells also saw a peak at 24 hours before decreasing over time. M0 macrophage cells started lower in infected samples but increased over time in both infected and uninfected samples. M2 macrophages were increased in uninfected samples after 24 hours compared with infected samples, which was not seen in either *in vivo* datasets. Similarly, monocytes were increased in uninfected samples compared with infected samples across almost all time-points which was in contrast to what was seen in the *in vivo* datasets. Activated mast cells were increased in infected samples compared to control, but in both decreased over time, which again was not observed in the *in vivo* datasets.

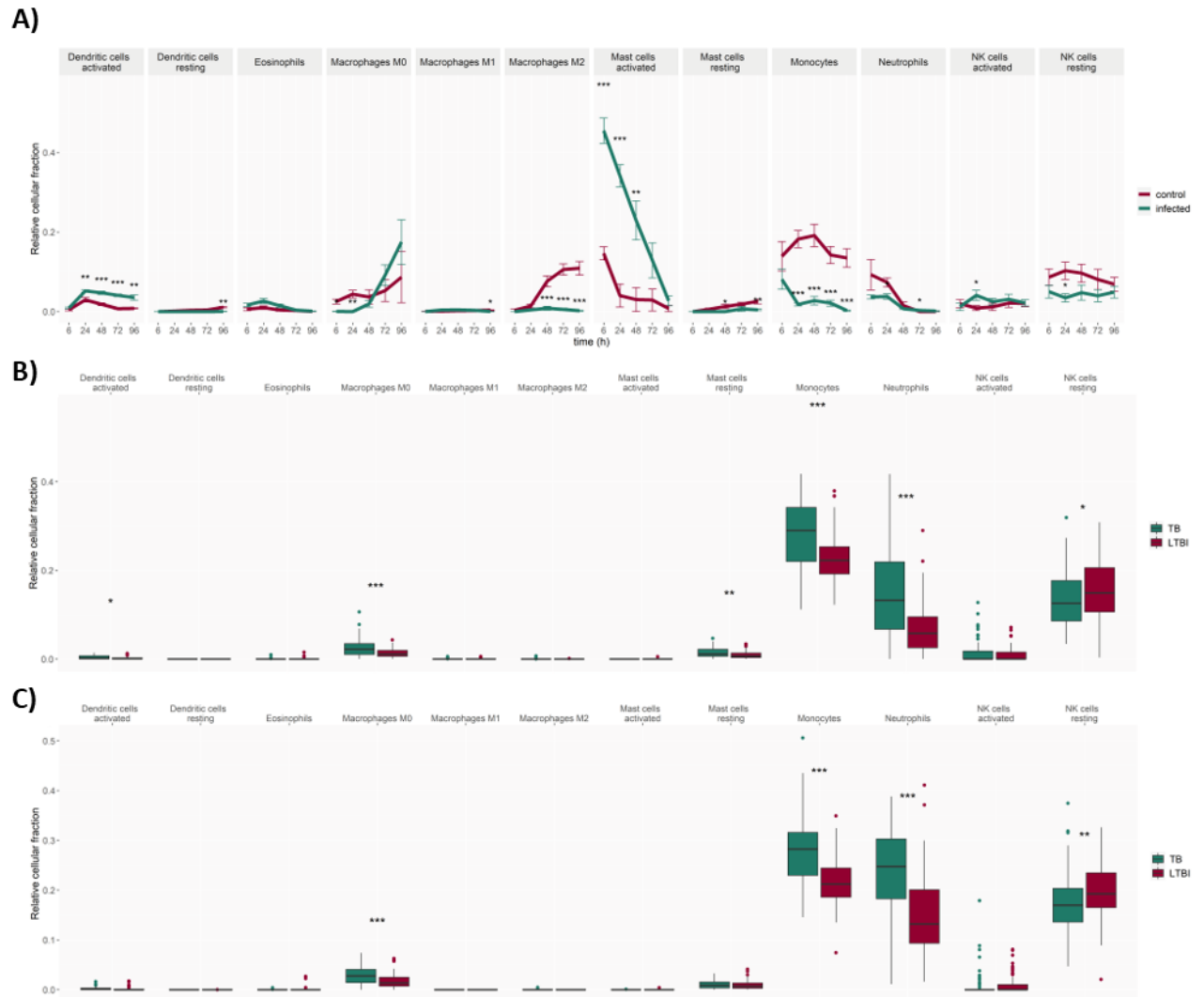


Figure S5.4.9: In silico flow cytometry analysis of innate immune cell populations in in vitro and in vivo datasets. A) Changes in cell populations in *Mtb*-infected and uninfected whole blood from the in vitro dataset over time. **B)** Changes in cell populations in active TB and latent TB patients in the paediatric in vivo dataset. **C)** Changes in cell populations in active TB and latent TB patients in the adult in vivo dataset. Error bars indicate the standard deviation from the mean. * $p < 0.05$; ** $p < 0.005$; *** $p < 0.001$.

The adaptive cells in both *in vivo* datasets were in less agreement than what was seen in the innate cells. In both *in vivo* datasets, memory B cells, CD8 T cells, and naïve CD4 T cells were increased in latent TB patients than active TB patients (**Figures S5.4.10B and S5.4.10C**). In the paediatric *in vivo* dataset, the remaining adaptive immune cells saw no, or close to no, difference between patient groups. In the adult *in vivo* dataset, naïve B cells, resting CD4 memory T cells, and Treg cells were also increased in latent TB patients compared to active TB patients. In contrast, activated CD4 memory T cells were increased in active TB patients compared to latent TB patients in this dataset.

There were few significant differences observed in the *in vitro* dataset when investigating differences in adaptive immune cells, likely due to large variances across samples (**Figure S5.4.10**). However, activated CD4 memory T cells were found to be increased in infected samples compared to uninfected samples at 48-, 72-, and 96-hours post-infection which is similar to what was observed in the adult *in vivo* dataset.

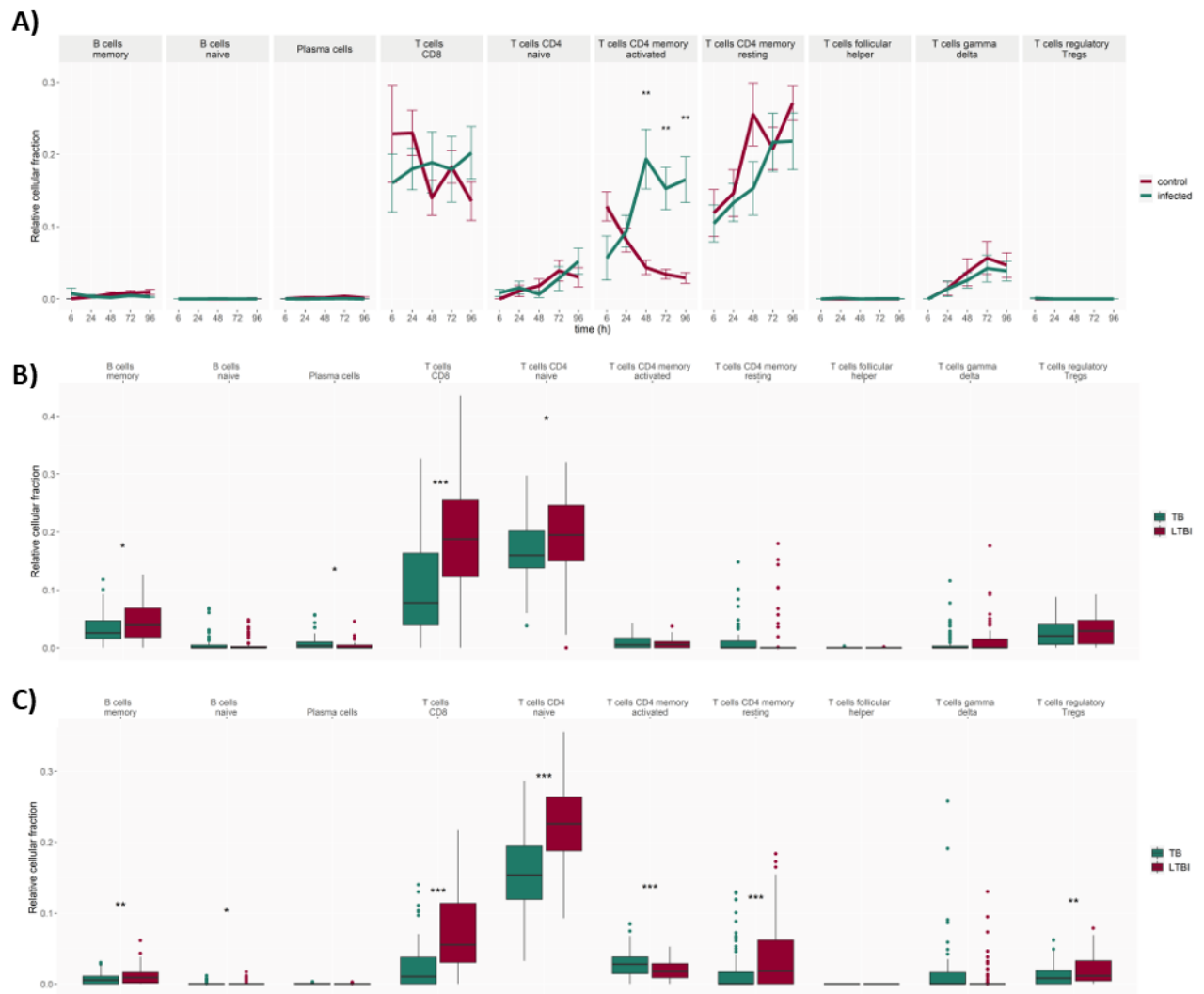


Figure S5.4.10: In silico flow cytometry analysis of adaptive immune cell populations in in vitro and in vivo datasets. A) Changes in cell populations in *Mtb*-infected and uninfected whole blood from the in vitro dataset over time. **B)** Changes in cell populations in active TB and latent TB patients in the paediatric in vivo dataset. **C)** Changes in cell populations in active TB and latent TB patients in the adult in vivo dataset. Error bars indicate the standard deviation from the mean. * $p < 0.05$; ** $p < 0.005$; *** $p < 0.001$.

5.4 Discussion

Despite the COVID-19 pandemic preventing laboratory work, using publicly available datasets allowed the bioinformatic comparison of the *in vitro* WBA with *in vivo* Mtb infection. This analysis is the first time the WBA has been directly compared to true infection and, therefore, provides important information regarding which aspects of the WBA model *in vivo* infection well and which aspects would require further validation. While the WBA dataset used here recruited adult donors, infection of whole blood with Mtb using this model may reflect short-term acute infection more closely than long-term adult infection. Therefore, both adult (i.e. possibly longer term and less naïve in immune response) and paediatric (acute, naïve immune response) cohorts were compared with the WBA dataset.

Up-regulated pathways were generally more concordant in the comparison with the paediatric *in vivo* dataset, while down-regulated pathways were more concordant in the comparison with the adult *in vivo* dataset. This suggests that pathways known to be up-regulated in adult infection using the WBA, or down-regulated in paediatric infection using the WBA, should not be modelled using this system.

Specifically, for the comparison of the *in vitro* assay with the adult *in vivo* dataset, chemokines, cytokines, the type I IFN response, B cells, and T cells were all concordant and therefore relevant to true infection. In contrast, innate immune pathways including antigen presentation and dendritic cells were discordant between the datasets, alongside B cell and T cell pathways which were not modelled well at later time-points. This suggests that modelling of B and T cell subsets should be performed at earlier time-points, although it is unlikely that they will be very involved in the initial immune response. Monocytes and inflammation were both concordant and discordant, suggesting that more detailed analysis into specific monocyte and inflammation pathways should be performed to understand which aspects of these cells in the WBA are relevant for adult *in vivo* infection.

When comparing the *in vitro* assay with the paediatric *in vivo* dataset antigen presentation and T cell activation pathways were all modelled well with high concordance, while pathways involving B cells, NK cells, chemokines, and cytokines, were not modelled well with high discordance. Monocytes, however, showed high concordance and discordance, similarly with the adult *in vivo* dataset comparison.

While providing substantial information regarding cell populations in these datasets, there are some concerns with using bioinformatic approaches to predict cell populations. Recently, CibersortX has been shown to be accurate at estimating several cell populations including most innate immune cells,

B and T cells, but less accurate at estimating NK cells and Tregs (Miao et al., 2020). In addition, CIBERSORTX does not provide estimated for subset cell populations such as T helper cells. To overcome this limitation, flow cytometry analysis should be performed alongside RNA-seq analysis where possible.

Overall, this analysis has shown that aspects of both the adult and paediatric *in vivo* response to infection can be modelled using the WBA. Many pathways were modelled with higher concordance at earlier time-points, particularly for adaptive immune cells. While there are some inconclusive results regarding key pathways, such as monocytes and inflammation, concordance was also shown. This suggests that certain aspects of these responses are relevant for *in vivo* infection, but that more in-depth pathway analysis is needed for these pathways in particular. However, this is the first time the *in vitro* WBA has been compared to *in vivo* infection, and further laboratory and bioinformatic studies should be performed to validate these findings in more detail.

5.4 Appendix 3: GO term analysis

Table S5.4.11: GO term analysis for upregulated terms in the *in vitro* WBA dataset

go	cluster	score	size	term	Time-point
GO:0009987	1	12.01323	3	cellular process	24
GO:0030593	2	4.376751	87	neutrophil chemotaxis	24
GO:0065007	3	3.853872	0	biological regulation	24
GO:0050789	3	3.522879	0	regulation of biological process	24
GO:0045765	5	3.508638	34	regulation of angiogenesis	24
GO:0043903	4	3.508638	0	regulation of symbiotic process	24
GO:0071347	2	3.161151	86	cellular response to interleukin-1 negative regulation of cell population	24
GO:0008285	6	3.040959	460	proliferation	24
GO:0007267	7	2.818156	231	cell-cell signaling	24
GO:0070098	2	2.739929	78	chemokine-mediated signaling pathway	24
GO:0050727	8	2.728158	88	regulation of inflammatory response	24
GO:0071222	4	2.66354	181	cellular response to lipopolysaccharide	24
GO:0002548	2	2.614394	52	monocyte chemotaxis	24
GO:0045766	5	2.61261	175	positive regulation of angiogenesis positive regulation of lymphocyte	24
GO:0050671	6	2.55752	6	proliferation	24
GO:1903706	5	2.337242	4	regulation of hemopoiesis	24
GO:0006954	8	2.326058	404	inflammatory response positive regulation of cell population	24
GO:0008284	6	2.294136	603	proliferation	24
GO:0035195	9	2.291579	896	gene silencing by miRNA	24
GO:0009987	1	12.01323	3	cellular process	48
GO:0030593	2	4.376751	87	neutrophil chemotaxis	48
GO:0065007	3	3.853872	0	biological regulation	48
GO:0050789	3	3.522879	0	regulation of biological process	48
GO:0045765	5	3.508638	34	regulation of angiogenesis	48
GO:0043903	4	3.508638	0	regulation of symbiotic process	48
GO:0071347	2	3.161151	86	cellular response to interleukin-1 negative regulation of cell population	48
GO:0008285	6	3.040959	460	proliferation	48
GO:0007267	7	2.818156	231	cell-cell signaling	48
GO:0070098	2	2.739929	78	chemokine-mediated signaling pathway	48
GO:0050727	8	2.728158	88	regulation of inflammatory response	48
GO:0071222	4	2.66354	181	cellular response to lipopolysaccharide	48
GO:0002548	2	2.614394	52	monocyte chemotaxis	48
GO:0045766	5	2.61261	175	positive regulation of angiogenesis positive regulation of lymphocyte	48
GO:0050671	6	2.55752	6	proliferation	48
GO:1903706	5	2.337242	4	regulation of hemopoiesis	48
GO:0006954	8	2.326058	404	inflammatory response	48

GO:0008284	6	2.294136	603	positive regulation of cell population proliferation	48
GO:0035195	9	2.291579	896	gene silencing by miRNA	48
GO:0009987	1	6.376751	3	cellular process	72
GO:0045765	2	3.823909	34	regulation of angiogenesis	72
GO:0032675	2	3.638272	11	regulation of interleukin-6 production	72
GO:0006954	3	3.173925	404	inflammatory response	72
GO:0009615	4	3.173925	109	response to virus	72
GO:0007166	5	2.835647	284	cell surface receptor signaling pathway	72
GO:0050789	6	2.761954	0	regulation of biological process	72
GO:0030593	4	2.50724	87	neutrophil chemotaxis	72
GO:0002250	7	2.49485	357	adaptive immune response	72
GO:0006959	7	2.371611	60	humoral immune response	72
GO:0043410	8	2.339135	126	positive regulation of MAPK cascade	72
GO:0050671	9	2.326058	6	positive regulation of lymphocyte proliferation	72
GO:0072594	10	2.320572	10	establishment of protein localization to organelle	72
GO:0071222	4	2.294136	181	cellular response to lipopolysaccharide	72
GO:0042129	9	2.188425	13	regulation of T cell proliferation	72
GO:0031399	8	2.166853	0	regulation of protein modification process	72
GO:0008284	9	2.125518	603	positive regulation of cell population proliferation	72
GO:1903708	2	2.105684	0	positive regulation of hemopoiesis	72
GO:0043547	11	2.047208	333	positive regulation of GTPase activity	72
GO:0009987	1	6.376751	3	cellular process	96
GO:0045765	2	3.823909	34	regulation of angiogenesis	96
GO:0032675	2	3.638272	11	regulation of interleukin-6 production	96
GO:0006954	3	3.173925	404	inflammatory response	96
GO:0009615	4	3.173925	109	response to virus	96
GO:0007166	5	2.835647	284	cell surface receptor signaling pathway	96
GO:0050789	6	2.761954	0	regulation of biological process	96
GO:0030593	4	2.50724	87	neutrophil chemotaxis	96
GO:0002250	7	2.49485	357	adaptive immune response	96
GO:0006959	7	2.371611	60	humoral immune response	96
GO:0043410	8	2.339135	126	positive regulation of MAPK cascade	96
GO:0050671	9	2.326058	6	positive regulation of lymphocyte proliferation	96
GO:0072594	10	2.320572	10	establishment of protein localization to organelle	96
GO:0071222	4	2.294136	181	cellular response to lipopolysaccharide	96
GO:0042129	9	2.188425	13	regulation of T cell proliferation	96
GO:0031399	8	2.166853	0	regulation of protein modification process	96
GO:0008284	9	2.125518	603	positive regulation of cell population proliferation	96
GO:1903708	2	2.105684	0	positive regulation of hemopoiesis	96
GO:0043547	11	2.047208	333	positive regulation of GTPase activity	96

Table S5.4.12: GO term analysis for downregulated terms in the *in vitro* WBA dataset

go	cluster	score	size	term	Time-point
GO:0042593	1	2.920819	118	glucose homeostasis	24
GO:0002253	2	2.552842	2	activation of immune response	24
GO:0006952	3	2.327902	75	defense response	24
GO:0031349	3	2.017729	0	positive regulation of defense response	24
GO:0071674	2	1.935542	3	mononuclear cell migration	24
GO:0043604	4	1.899629	2	amide biosynthetic process	24
GO:0045089	5	1.899629	24	positive regulation of innate immune response	24
GO:0009605	6	1.876148	0	response to external stimulus	24
GO:0009607	5	1.801343	0	response to biotic stimulus	24
GO:0034622	7	1.785156	22	cellular protein-containing complex assembly	24
GO:0022613	7	1.777284	1	ribonucleoprotein complex biogenesis	24
GO:0006820	8	1.744727	18	anion transport	24
GO:0071222	6	1.744727	181	cellular response to lipopolysaccharide	24
GO:0055082	1	1.735182	1	cellular chemical homeostasis	24
GO:0043207	5	1.694649	2	response to external biotic stimulus	24
GO:0051707	5	1.694649	3	response to other organism	24
GO:0048771	9	1.686133	8	tissue remodeling	24
GO:0002833	5	1.66354	0	positive regulation of response to biotic stimulus	24
GO:0043603	4	1.661544	1	cellular amide metabolic process	24
GO:0050896	1	5.638272	69	response to stimulus	48
GO:0050829	2	4.60206	90	defense response to Gram-negative bacterium	48
GO:0055088	3	3.769551	47	lipid homeostasis	48
GO:0051607	2	3.69897	236	defense response to virus	48
GO:0009755	4	3.508638	52	hormone-mediated signaling pathway	48
GO:0045087	2	3.366532	599	innate immune response	48
GO:0019318	6	3.366532	1	hexose metabolic process	48
GO:0007265	5	3.366532	84	Ras protein signal transduction	48
GO:0030595	4	3.346787	24	leukocyte chemotaxis	48
GO:0046434	7	3.318759	1	organophosphate catabolic process	48
GO:0051701	2	3.29243	0	interaction with host	48
GO:0042593	3	3.267606	118	glucose homeostasis	48
GO:0030522	5	3.21467	26	intracellular receptor signaling pathway	48
GO:0050830	2	3.19382	120	defense response to Gram-positive bacterium	48
GO:0070374	7	3.19382	253	positive regulation of ERK1 and ERK2 cascade	48
GO:0071222	2	3.161151	181	cellular response to lipopolysaccharide	48
GO:0050878	3	2.571865	4	regulation of body fluid levels	48
GO:0045936	7	2.522879	0	negative regulation of phosphate metabolic process	48
GO:0009987	1	21.03621	3	cellular process	72
GO:0006469	2	3.49485	87	negative regulation of protein kinase activity	72
GO:0032870	4	3.431798	29	cellular response to hormone stimulus	72
GO:0051607	3	3.431798	236	defense response to virus	72
GO:0050896	4	3.036212	69	response to stimulus	72
GO:0050789	5	2.735182	0	regulation of biological process	72

GO:0048839	6	2.712198	47	inner ear development	72
GO:0007610	7	2.661544	8	behavior	72
GO:0071222	8	2.659556	181	cellular response to lipopolysaccharide	72
GO:0051701	8	2.465974	0	interaction with host	72
GO:0008217	9	2.411168	69	regulation of blood pressure	72
GO:0042742	8	2.337242	213	defense response to bacterium	72
GO:0006952	3	2.318759	75	defense response	72
GO:0065007	5	2.275724	0	biological regulation	72
GO:0006897	10	2.273273	197	endocytosis	72
GO:0070374	2	2.266001	253	positive regulation of ERK1 and ERK2 cascade	72
GO:0071407	4	2.237321	45	cellular response to organic cyclic compound	72
GO:0051239	11	2.197911	0	regulation of multicellular organismal process	72
GO:0015748	10	2.19382	0	organophosphate ester transport	72
GO:0071704	1	3.173925	0	organic substance metabolic process	96
GO:0065007	2	2.879426	0	biological regulation	96
GO:0006575	3	2.798603	1	cellular modified amino acid metabolic process	96
GO:0002699	4	2.600326	0	positive regulation of immune effector process	96
GO:0043030	4	2.590067	10	regulation of macrophage activation	96
GO:0044255	5	2.575118	16	cellular lipid metabolic process	96
GO:0003018	6	2.527244	0	vascular process in circulatory system	96
GO:1902652	7	2.457175	0	secondary alcohol metabolic process	96
GO:0070374	8	2.443697	253	positive regulation of ERK1 and ERK2 cascade	96
GO:1901575	1	2.420216	0	organic substance catabolic process	96
GO:0150076	9	2.404504	4	neuroinflammatory response	96
GO:0044237	1	2.396856	1	cellular metabolic process	96
GO:0001649	10	2.388277	110	osteoblast differentiation	96
GO:0008610	5	2.365523	15	lipid biosynthetic process	96
GO:0050778	4	2.333482	13	positive regulation of immune response	96
GO:0050789	2	2.294136	0	regulation of biological process	96
GO:0038061	8	2.286509	72	NIK/NF-kappaB signaling	96
GO:0008015	6	2.269218	34	blood circulation	96

Table S5.4.13: GO term analysis for paediatric *in vivo* dataset

go	cluster	score	size	term	Direction
GO:0044237	1	6.958607	1	cellular metabolic process	Up
GO:0050832	2	6.568636	32	defense response to fungus	Up
GO:0071704	1	6.39794	0	organic substance metabolic process	Up
GO:0006807	1	5.522879	7	nitrogen compound metabolic process	Up
GO:0031640	2	5.346787	41	killing of cells of other organism	Up
GO:0050794	3	4.619789	2	regulation of cellular process	Up
GO:0002227	2	4.443697	32	innate immune response in mucosa	Up
GO:0065007	3	4.443697	0	biological regulation	Up
GO:0071383	4	4.244125	13	cellular response to steroid hormone stimulus	Up
GO:0035821	2	4.004365	2	modulation of process of other organism	Up
GO:0009755	4	3.69897	52	hormone-mediated signaling pathway	Up
GO:0061844	2	3.69897	90	antimicrobial humoral immune response mediated by antimicrobial peptide	Up
GO:0050829	2	3.69897	90	defense response to Gram-negative bacterium	Up
GO:0050830	2	3.69897	120	defense response to Gram-positive bacterium	Up
GO:0022618	5	3.657577	2	ribonucleoprotein complex assembly	Up
GO:0071222	4	3.568636	181	cellular response to lipopolysaccharide	Up
GO:0019731	2	3.552842	57	antibacterial humoral response	Up
GO:0051702	2	3.167491	5	interaction with symbiont	Up
GO:0031329	1	2.568636	0	regulation of cellular catabolic process	Down
GO:0045786	2	2.537602	45	negative regulation of cell cycle	Down
GO:0006996	3	2.481486	11	organelle organization	Down
GO:0030168	5	2.39794	108	platelet activation	Down
GO:0090501	4	2.39794	20	RNA phosphodiester bond hydrolysis	Down
GO:0044237	6	2.366532	1	cellular metabolic process	Down
GO:0065007	7	2.356547	0	biological regulation	Down
GO:0007411	3	2.318759	252	axon guidance	Down
GO:0010605	2	2.259637	0	negative regulation of macromolecule metabolic process	Down
GO:0006364	8	2.251812	172	rRNA processing	Down
GO:0009314	9	2.161151	24	response to radiation	Down
GO:0016043	3	2.148742	0	cellular component organization	Down
GO:0031324	2	2.09691	1	negative regulation of cellular metabolic process	Down
GO:0016236	1	2.031517	115	macroautophagy	Down
GO:0042273	8	2.004365	34	ribosomal large subunit biogenesis	Down
GO:0051172	2	1.995679	0	negative regulation of nitrogen compound metabolic process	Down
GO:0030308	2	1.978811	132	negative regulation of cell growth	Down
GO:0010506	1	1.946922	75	regulation of autophagy	Down

Table S5.4.14: GO term analysis for adult *in vivo* dataset

go	cluster	score	size	term	Direction
GO:0050794	1	7.481486	2	regulation of cellular process	Up
GO:0065007	1	5.638272	0	biological regulation	Up
GO:0007165	2	4.065502	1056	signal transduction	Up
				cellular response to mechanical	
GO:0071260	3	4.031517	80	stimulus	Up
				positive regulation of interleukin-1	
GO:0032731	4	4.008774	60	beta production	Up
				positive regulation of peptidyl-	
GO:0050731	5	3.769551	92	tyrosine phosphorylation	Up
GO:0006950	6	3.744727	0	response to stress	Up
GO:0071704	7	3.69897	0	organic substance metabolic process	Up
GO:0019216	5	3.657577	95	regulation of lipid metabolic process	Up
				receptor signaling pathway via JAK-	
GO:0007259	2	3.568636	34	STAT	Up
GO:0051716	2	3.309804	0	cellular response to stimulus	Up
GO:0043408	5	3.119186	23	regulation of MAPK cascade	Up
				positive regulation of inflammatory	
GO:0050729	6	3.119186	94	response	Up
				positive regulation of biological	
GO:0048518	4	3.031517	0	process	Up
GO:0046631	8	2.966576	1	alpha-beta T cell activation	Up
GO:0051179	9	2.935542	0	localization	Up
				negative regulation of secretion by	
GO:1903531	9	2.931814	2	cell	Up
GO:0006955	8	2.866461	472	immune response	Up
GO:0009987	1	6.522879	3	cellular process	Down
GO:0052548	2	3.481486	8	regulation of endopeptidase activity	Down
GO:0032535	3	2.501689	1	regulation of cellular component size	Down
				carbohydrate derivative metabolic	
GO:1901135	4	2.477556	1	process	Down
				cellular protein-containing complex	
GO:0034622	3	2.412289	22	assembly	Down
				regulation of actin filament	
GO:0110053	3	2.312471	2	organization	Down
				negative regulation of protein	
GO:0031400	2	2.298432	0	modification process	Down
GO:0038093	5	2.191789	4	Fc receptor signaling pathway	Down
GO:0019752	6	2.14752	5	carboxylic acid metabolic process	Down
				negative regulation of cellular	
GO:0051129	3	2.022734	0	component organization	Down
				immune response-activating cell	
GO:0002429	5	1.982549	4	surface receptor signaling pathway	Down
				immune response-activating signal	
GO:0002757	5	1.982549	1	transduction	Down
GO:0022604	7	1.902049	24	regulation of cell morphogenesis	Down
GO:0060341	8	1.898253	6	regulation of cellular localization	Down
GO:0051276	3	1.816161	16	chromosome organization	Down
GO:0002253	5	1.746662	2	activation of immune response	Down
GO:0005975	9	1.706859	126	carbohydrate metabolic process	Down
GO:0016043	3	1.659159	0	cellular component organization	Down
				positive regulation of cellular	
GO:0051130	3	1.606425	0	component organization	Down

Chapter 6: General Discussion

With the onset of the COVID-19 pandemic, the research of many other high priority infectious diseases has been delayed or neglected. Prior to SARS-CoV-2, TB was the highest cause of death from an infectious disease globally (WHO, 2020), and many countries were still not on track to reach the End TB goals by 2035, set out by the WHO (WHO, 2014). Despite these goals and progress by many towards strategies to tackle this disease, many aspects of TB disease were still understudied, including Mtb lineages and paediatric TB. This thesis takes a unique approach, through investigating different disciplines of TB research, to form a more comprehensive understanding of this disease.

6.1 TB research requires an interdisciplinary outlook

Transcriptomic TB biomarker signatures have shown great potential to be developed into diagnostic tests, with the discovery of several signatures in recent years. However, despite the signatures published, one of the main findings from this thesis is the difficulty for the paediatric biomarker signatures to meet the WHO TPP guidelines. The reasons for this have been discussed, including using limiting training cohorts, lack of validation, and small sample sizes. However, one aspect of these studies which is often overlooked is the lack of biological understanding. While it may not be surprising that some of the top genes in many of the published signatures are involved in IFN γ signalling, the authors often fail to discuss other pathways which may be involved, particularly those of the bacterium. It could be argued that the primary aim of these diagnostic biomarker studies is clinical application, and therefore not understanding the underlying biology, and some experts in the field believe there is no biological significance of the genes selected in these signatures.

When comparing the top differentially expressed genes in the paediatric South African dataset in Chapter 3 with findings of previous studies there was several similarities in the top differentially expressed genes in both the TB vs OD and TB vs LTBI comparisons. In contrast, comparing the top differentially expressed genes from the paediatric South African dataset with the host pilot WBA dataset in Chapter 5 found none of the top genes were common across the datasets. There may be several reasons for this (described below). However, as latent TB has shown to have a similar transcriptomic profile to healthy controls (Broderick et al., 2021), similarities between the WBA

dataset and the TB vs LTBI comparison of the paediatric South African dataset could be expected. Looking higher than at gene level may identify similarities. Pathway analysis was not performed for the pilot WBA dataset but extrapolation to pathways can still be estimated. Between both datasets common pathways which change included the up-regulation of immune and inflammatory genes (such as *IL1A* in the WBA dataset between 0 and 11 hrs and *GBP* genes in the South African dataset in the TB vs OD comparison). This is expected as the characteristic TB response is of up-regulation of immune responses and inflammation, particularly the IFN γ pathway (Sia & Rengarajan, 2019).

Another similar pathway identified was the up-regulation of non-coding genes in both the WBA dataset at later time-points (24 and 48 hrs) and the South African dataset in the TB vs LTBI comparison. These included genes such as *LNIC01705* in the WBA dataset and many *SNOR* genes in the South African dataset. The functions of many non-coding RNAs are still unknown, so their downstream effects in these datasets should be investigated. Other long non-coding RNA functions include reduced T cell and helper T cell activation, involvement in cAMP and calcium signalling, and NK cell function (Almatroudi, 2022). Similarly, little is known of the function of the *SNOR* genes in TB. Coincidentally, one study investigating transcriptomic differences in severe COVID-19 patients with healthy controls identified both *SNORD17* and *SNORA53*, two of the top *SNOR* RNAs in the South African dataset, as being differentially down-regulated in severe COVID-19 patients, suggesting a potential role in lung pathology (Budhraja et al., 2022).

However, there could be several explanations for the lack of concordant or discordant genes between these two datasets. One possibility is that the South African dataset is from a paediatric cohort, while our WBA dataset is an adult cohort of donor, and adult and paediatric TB responses are known vary considerably. Despite this, the two models may reflect similar stages of TB disease. Paediatric TB disease is usually acute and develops into active disease rapidly. Similarly, as the infection of whole blood in the WBA is relatively short-term compared to adult TB disease which may last decades, it can also be considered acute. Nevertheless, the lack of biological replicates in the pilot WBA dataset means that any findings must be validated in a larger experiment before any biological interpretation.

Instead of comparing the genes identified in the paediatric South African cohort with the genes in the WBA, a broader investigation looking at adult *in vivo* studies may find similarities with the WBA. Across most biomarker studies there are consistently chosen genes, including *BATF2*, *GBP5*, *GBP6*, *C1QB*, and

FCGR1A. Some of these genes were also identified in the top differentially expressed genes in the South African dataset in Chapter 3 (*GBP5*, *BATF2*, and *FCGR1A* were up-regulated in the TB vs OD comparison). However, again none of these genes were identified in the WBA dataset presented in Chapter 5.

Despite certain genes being consistently identified as the top biomarker genes, all the currently published biomarker signatures fail to meet the WHO guidelines. It is likely that signatures underperform in specific countries due to differences in circulating Mtb lineages in those regions, and thereby the differences in host response to those lineages. Specifically, signatures underperform in countries such as The Gambia, where the dominant Mtb lineages are lineages 5 and 6 (Gagneux, 2018; Singer et al., 2022; Suliman et al., 2018). These lineages are unique to Western Africa, whose countries are frequently not used as training cohorts for biomarker studies. In the few studies where cohorts are sourced from these countries, the different host responses in these cohorts compared to cohorts from other countries such as South Africa result in skewed biomarker signatures which are accurate in one population and not the other (Suliman et al., 2018).

To identify if differing circulating Mtb lineages in cohorts in these studies are the cause of the lack of accuracy in biomarkers in different cohorts, Mtb lineage typing should be performed during patient recruitment for future biomarker studies. Furthermore, studies should focus on cohorts recruited from regions of the world with different Mtb lineages, such as West Africa, Ethiopia, Europe, and Asia. This, alongside further studies including paediatric cohorts, which are often overlooked, will advance the potential of biomarker signatures to being developed into diagnostic tests.

Alongside lack of microbial phenotyping in biomarker studies, there is also a lack of functional annotations of the Mtb and BCG genome, transcriptome, and lack of available resources to study these mycobacteria compared to those available for host research. This has restricted the analysis in both Chapters 4 and 5. In Chapter 4, there are few pathway analysis tools which can functionally annotate Mtb genes, resulting in either broad analyses of less than 10 categories, which is not sufficient to identify differences between lineages, or many categories extracted from an online database instead of being available as an independent analysis tool with statistical significant testing features. Furthermore, many Mtb genes and metabolites themselves are uncharacterised, limiting the interpretation of any significant findings.

Despite the limitations, the work in Chapter 4 is only the second study integrating Mtb genomic and metabolomic data. This analysis provides a more easily accessible and straightforward approach to integrate these datatypes and can be used on a dataset with a small sample size. An important finding was the correlation between the rs6112 nsSNP block and ergothioneine. None of the genes within the nsSNP block were directly involved in the ergothioneine biosynthesis pathway, however there were several indirect pathways which should be investigated. Future studies can use the approaches described here to interrogate this pathway. For example, the findings from Chapter 4 highlight several Mtb genes (*rpsJ*, *sucD*, *metE*, *pknH*, *vapC10*, *ppsE*, and *desA3*) which may be involved in the difference in ergothioneine production in Mtb strains in lineage 1 compared to lineages 2 and 4. To investigate these genes further, knock-out lineage 1 Mtb strains could be generated for those genes which are not essential. Currently, *pknH*, *vapC10*, and *ppsE* are not known to be essential (R. Zhang et al., 2004), and deletion mutants of *pknH* and *ppsE* have previously been studied (although the *ppsE* deletion was one of several deletions in this Mtb mutant) (Papavinasasundaram et al., 2005; Passemar et al., 2014). The *pknH* mutant showed decreased mycobacterial growth in THP-1 cells, although no transcriptomics was performed to measure changes in other pathways. These strains could then be investigated in culture to identify any differences in ergothioneine production or related pathways. To supplement this, they can then be used to infect whole blood from healthy donors through the WBA to measure the bacterial and host response in a more complex immune environment. Dual RNA-seq transcriptomics of this experiment could show if ergothioneine synthesis pathways are altered in the bacteria. As ergothioneine has been shown to protect against oxidative stress, bacterial oxidative responses can also be measured using this approach (Cumming et al., 2018). Host pathways involved in establishing an oxidative stress response toward Mtb may also change upon infection with knock-out strains compared with control strains. Findings from these experiments could then be compared with *in vivo* infection of TB patients infected with Mtb strains from lineage 1 and lineages 2 and 4 to confirm aberrations in ergothioneine regulation and other pathways identified from the previous experiments.

The small sample size in this Chapter prevented identification of further lineage-specific SNP drivers of metabolic differences, as a large number of SNPs correlated with multiple metabolites. This work should be expanded to include strains from each Mtb lineage, and a greater sample size, to identify further correlations.

6.2 The shortcomings and future possibilities of the WBA

The WBA offers a unique method of understanding host-pathogen interactions, through including several immune cells which can interact over time. Furthermore, the assay is affordable compared to *in vivo* models as it does not require animal husbandry or housing. Despite the benefits of using the WBA over cell models, the complexity of whole blood presents challenges in experimental design and analysis of results. One of the challenges encountered in this thesis was accounting for different cell compositions across samples in a single experiment. This was accounted for with detailed flow cytometry analysis, however this resulted in doubling of sample numbers which led to logistical difficulties. As a result, it was possible to analyse 6- and 24-hour time-points, but further time-points would be impossible. A more detailed analysis considering the composition of immune cells could be the use of the WBA with single cell RNA-seq to help understand the transcriptional changes caused by different cell populations. This would allow for the analysis of transcriptional signatures which are specific to immune cells of interest. Furthermore, immune cells and resulting transcripts which are modelled poorly by the WBA, as described in Chapter 6, can be excluded from the analysis using this approach, increasing the relevance of the WBA to true infection.

A larger limitation of the WBA and dual RNA-seq approach is the lack of biological characterisation of the WBA and the novelty, and therefore also lack of characterisation, of dual RNA-seq. This thesis has attempted to characterise the WBA further through optimising efficient RNA extraction protocols, understanding the survival of immune cell subsets over time, and comparing the WBA to *in vivo* infection datasets. However, more should be done to characterise the assay such as understanding why some cell populations expand during the assay, the proportion of cells which become infected and which cells, and more detailed comparisons with *in vivo* infection should be carried out.

While dual RNA-seq has been performed on several different models of TB, it has not previously been carried out using whole blood. This allowed for a unique analysis of host-pathogen interactions over time in a complex system. In this thesis, it was shown that this experimental approach is possible, and, despite low bacterial read counts, differential gene expression analysis could be achieved here and statistical significance is likely with an experiment with biological replicates. Furthermore, this approach did not compromise the host RNA quality and, in fact, improved the number of host reads

obtained. Another highlight of this approach is the identification of novel transcripts which are up or down-regulated in the host response to infection due to the use of RNA-seq. This builds on previous publications using a similar approach but with microarray analysis. This approach could easily be tailored towards other RNA species which are less accessible using microarrays, such as small non-coding RNAs or other RNA species.

To this end, a larger dual RNA-seq approach could be used to investigate the non-coding RNAs identified in the analyses in Chapter 3, and therefore help identify further functions of these RNAs. This was one of the first uses for dual RNA-seq, in the early study by Westermann, *et. al.*, who identified the small regulatory RNA, *PinT*, as important for *Salmonella* virulence (Westermann et al., 2012).

The dual-RNA seq WBA approach could also be used to investigate differences in host-pathogen interactions as a result of infection with different Mtb lineages, validating and supplementing the results in Chapter 4. This could be performed using whole blood from patients from countries with different Mtb lineages, as described above, or additionally through using whole blood from healthy donors to then be infected with Mtb strains from different lineages. Multiple types of data could then be collected, genomic, transcriptomic, or metabolomic. This would enable the validation of the results in Chapter 4 in an *in vitro* system more closely reflecting *in vivo* infection, as the current genomic and metabolomic analysis was performed using Mtb cultures. This approach would likely also produce new metabolomic insights, as a whole blood model will allow Mtb to metabolise different carbon sources compared with in liquid culture. Mtb will up-regulate and down-regulate different metabolic pathways depending on the carbon source available, and particularly as cholesterol in host immune cells are used as an energy source by the bacteria (Aguilar-Ayala et al., 2017).

This thesis provides the starting steps to establish further experiments using the WBA and dual-RNA-seq. Future work applying the WBA dual RNA-seq approach to blood infected with Mtb would be more relevant for the study of TB disease. Furthermore, this can be expanded through using whole blood from TB patients to understand infection in those with chronic disease. This could include paediatric patients as well as adult patients, as previous studies have shown the WBA can be used with smaller blood volumes from paediatric cohorts (R. B. Roy et al., 2019).

6.3 Conclusions

The results of this thesis have forwarded several fields of TB research. Current TB biomarker signatures have been evaluated on an independent dataset, and the lack of consideration of paediatric cohorts addressed. Furthermore, the main limitations in the field have been identified and expanded upon. This is the second time genomic and metabolomic Mtb data have been integrated, through an approach which can be easily replicated by others. This identified a novel interaction between Mtb lineage 1 strains, and resulting nsSNPs, with ergothioneine which should be further validated. Finally, a workflow for performing dual RNA-seq on the WBA was optimised and developed. This sets the stage for further dual RNA-seq analyses to be performed using the WBA, and indeed other intracellular pathogens can also be studied using this approach. Finally, the WBA was also further characterised by comparing the *in vitro* host response to *in vivo* infection in both adult and paediatric datasets. This is the first time comparison between the WBA and true infection has been performed and this identified which immune pathways should, and should not be, studies using this assay. Overall, this work has built on several fields of, often neglected, TB research and should encourage further study of these topics.

Chapter 7: Bibliography

- Abramovitch, R. B., Rohde, K. H., Hsu, F.-F., & Russell, D. G. (2011). *aprABC: A Mycobacterium tuberculosis complex-specific locus that modulates pH-driven adaptation to the macrophage phagosome*. <https://doi.org/10.1111/j.1365-2958.2011.07601.x>
- Adrian, A., & Jorg, R. (2021). *topGO: Enrichment Analysis for Gene Ontology*.
- Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., Bourexis, D., Brister, J. R., Bryant, S. H., Canese, K., Charowhas, C., Clark, K., Dicuccio, M., Dondoshansky, I., Federhen, S., Feolo, M., Funk, K., Geer, L. Y., Gorelenkov, V., ... Zbicz, K. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44(D1), D7–D19. <https://doi.org/10.1093/NAR/GKV1290>
- Aguilar-Ayala, D. A., Tilleman, L., van Nieuwerburgh, F., Deforce, D., Palomino, J. C., Vandamme, P., Gonzalez-Y-Merchand, J. A., & Martin, A. (2017). The transcriptome of Mycobacterium tuberculosis in a lipid-rich dormancy model through RNAseq analysis. *Scientific Reports* 2017 7:1, 7(1), 1–13. <https://doi.org/10.1038/s41598-017-17751-x>
- Ahmed, N., & Hasnain, S. E. (2011). Molecular epidemiology of tuberculosis in India: Moving forward with a systems biology approach. *Tuberculosis*, 91(5), 407–413. <https://doi.org/10.1016/J.TUBE.2011.03.006>
- Akgul, C., Moulding, D. A., & Edwards, S. W. (2001). Molecular control of neutrophil apoptosis. *FEBS Letters*, 487(3), 318–322. [https://doi.org/10.1016/S0014-5793\(00\)02324-3](https://doi.org/10.1016/S0014-5793(00)02324-3)
- Alessandra, M., Chiara, P., & Manuela, S. (2018). Genetic susceptibility to infectious disease: Current status and future perspectives from genome-wide approaches. *Infection, Genetics and Evolution*, 66, 286–307. <https://reader.elsevier.com/reader/sd/pii/S1567134817303349?token=F35B1F34C86069031E6FC8363E4DFB98A94EA3DF83FE1DED14ADB818992B12C8FDC3FC009F83229D31DD3FAAB67B36B1&originRegion=eu-west-1&originCreation=20210809103539>
- Almatroudi, A. (2022). Non-Coding RNAs in Tuberculosis Epidemiology: Platforms and Approaches for Investigating the Genome's Dark Matter. *International Journal of Molecular Sciences*, 23(8). <https://doi.org/10.3390/IJMS23084430>
- Amorim Franco, T. M., & Blanchard, J. S. (2017). Bacterial Branched-Chain Amino Acid Biosynthesis: Structures, Mechanisms, and Drugability. *Biochemistry*, 56(44), 5849. <https://doi.org/10.1021/ACS.BIOCHEM.7B00849>
- Anders, S., & Huber, W. (2010). *Differential expression analysis for sequence count data*. <https://doi.org/10.1038/npre.2010.4282.1>
- Anderson, S. T., Kaforou, M., Brent, A. J., Wright, V. J., Banwell, C. M., Chagaluka, G., Crampin, A. C., Dockrell, H. M., French, N., Hamilton, M. S., Hibberd, M. L., Kern, F., Langford, P. R., Ling, L., Mlotha, R., Ottenhoff, T. H. M., Pienaar, S., Pillay, V., Scott, J. A. G., ... Eley, B. (2014). Diagnosis of Childhood Tuberculosis and Host RNA Expression in Africa. *New England Journal of Medicine*, 370(18), 1712–1723. <https://doi.org/10.1056/nejmoa1303657>

- Andrews, S. (2010). v0.11.8, Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Asai, M., Li, Y., Khara, J. S., Robertson, B. D., Langford, P. R., & Newton, S. M. (2019). *Galleria mellonella*: An Infection Model for Screening Compounds Against the Mycobacterium tuberculosis Complex. *Frontiers in Microbiology*, 10, 2630. <https://doi.org/10.3389/FMICB.2019.02630/BIBTEX>
- Autissier, P., Soulas, C., Burdo, T. H., & Williams, K. C. (2010). Evaluation of a 12-color flow cytometry panel to study lymphocyte, monocyte, and dendritic cell subsets in humans. *Cytometry Part A*, 77A(5), 410–419. <https://doi.org/10.1002/CYTO.A.20859>
- Awuh, J. A., Trude, •, & Flo, H. (2017). Molecular basis of mycobacterial survival in macrophages. *Cellular and Molecular Life Sciences*, 74. <https://doi.org/10.1007/s00018-016-2422-8>
- Baguma, R., Penn-Nicholson, A., Smit, E., Erasmus, M., Day, J., Makhethhe, L., de Kock, M., Jane Hughes, E., van Rooyen, M., Pienaar, B., Stone, L., Hanekom, W., Brennan, M. J., Wallis, R. S., Hatherill, M., & Scriba, T. J. (2017). *Application of a whole blood mycobacterial growth inhibition assay to study immunity against Mycobacterium tuberculosis in a high tuberculosis burden population*. <https://doi.org/10.1371/journal.pone.0184563>
- Baguma, R., Penn-Nicholson, A., Smit, E., Erasmus, M., Day, J., Makhethhe, L., Kock, M. de, Hughes, E. J., Rooyen, M. van, Pienaar, B., Stone, L., Hanekom, W., Brennan, M. J., Wallis, R. S., Hatherill, M., & Scriba, T. J. (2017). Application of a whole blood mycobacterial growth inhibition assay to study immunity against Mycobacterium tuberculosis in a high tuberculosis burden population. *PLOS ONE*, 12(9), e0184563. <https://doi.org/10.1371/JOURNAL.PONE.0184563>
- Baliu-Piqué, M., Verheij, M. W., Drylewicz, J., Ravesloot, L., de Boer, R. J., Koets, A., Tesselaar, K., & Borghans, J. A. M. (2018). Short lifespans of memory T-cells in bone marrow, blood, and lymph nodes suggest that T-cell memory is maintained by continuous self-renewal of recirculating cells. *Frontiers in Immunology*, 9(SEP), 2054. <https://doi.org/10.3389/FIMMU.2018.02054/BIBTEX>
- Barh, D., Tiwari, S., Andrade, B. S., Weener, M. E., Góes-Neto, A., Azevedo, V., Ghosh, P., Blum, K., & Ganguly, N. K. (2021). A novel multi-omics-based highly accurate prediction of symptoms, comorbid conditions, and possible long-term complications of COVID-19. *Molecular Omics*, 17(2), 317–337. <https://doi.org/10.1039/D0MO00189A>
- Barh, D., Tiwari, S., Weener, M. E., Azevedo, V., Góes-Neto, A., Gromiha, M. M., & Ghosh, P. (2020). Multi-omics-based identification of SARS-CoV-2 infection biology and candidate drugs against COVID-19. *Computers in Biology and Medicine*, 126, 104051. <https://doi.org/10.1016/J.COMPBIOMED.2020.104051>
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1), D991–D995. <https://doi.org/10.1093/NAR/GKS1193>

- Basu Roy, R., Whittaker, E., & Kampmann, B. (2012). Current understanding of the immune response to tuberculosis in children. *Current Opinion in Infectious Diseases*, 25(3), 250–257. <https://doi.org/10.1097/QCO.0B013E3283529AF9>
- Basu Roy, R., Whittaker, E., Seddon, J. A., & Kampmann, B. (2019). Children and Mycobacterium tuberculosis: a review of susceptibility and protection. *The Lancet. Infectious Diseases*, 19(3), e96. [https://doi.org/10.1016/S1473-3099\(18\)30157-9](https://doi.org/10.1016/S1473-3099(18)30157-9)
- Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Silva, A. da, Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Castro, L. G., ... Zhang, J. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), D480–D489. <https://doi.org/10.1093/NAR/GKAA1100>
- Becton, Dickinson, & Company. (2021). *FlowJo™ Software for Windows [software application] Version 10.8.1*. Ashland, OR: Becton, Dickinson and Company; 2021.
- Benjak, A., Sala, C., & Hartkoorn, R. C. (2015). Whole-Transcriptome Sequencing for High-Resolution Transcriptomic Analysis in *Mycobacterium tuberculosis*. *Methods in Molecular Biology*, 1285, 17–30. https://doi.org/10.1007/978-1-4939-2450-9_2
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242. <https://doi.org/10.1093/NAR/28.1.235>
- Berry, M. P. R., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A. A., Oni, T., Wilkinson, K. A., Banchereau, R., Skinner, J., Wilkinson, R. J., Quinn, C., Blankenship, D., Dhawan, R., Cush, J. J., Mejias, A., Ramilo, O., Kon, O. M., Pascual, V., Banchereau, J., ... O'Garra, A. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature*, 466(7309), 973–977. <https://doi.org/10.1038/nature09247>
- Bespyatykh, J., Shitikov, E., Bespyatykh, D., Guliaev, A., Klimina, K., Veselovsky, V., Arapidi, G., Dogonadze, M., Zhuravlev, V., Ilina, E., & Govorun, V. (2020). Metabolic Changes of Mycobacterium tuberculosis during the Anti-Tuberculosis Therapy. *Pathogens* 2020, Vol. 9, Page 131, 9(2), 131. <https://doi.org/10.3390/PATHOGENS9020131>
- Bespyatykh, J., Shitikov, E., Guliaev, A., Smolyakov, A., Klimina, K., Veselovsky, V., Malakhova, M., Arapidi, G., Dogonadze, M., Manicheva, O., Bespyatykh, D., Mokrousov, I., Zhuravlev, V., Ilina, E., & Govorun, V. (2019). System OMICs analysis of Mycobacterium tuberculosis Beijing B0/W148 cluster. *Scientific Reports* 2019 9:1, 9(1), 1–11. <https://doi.org/10.1038/s41598-019-55896-z>
- Beste, D. J. V., Peters, J., Hooper, T., Avignone-Rossa, C., Bushell, M. E., & McFadden, J. (2005). Compiling a Molecular Inventory for Mycobacterium bovis BCG at Two Growth Rates: Evidence for Growth Rate-Mediated Regulation of Ribosome Biosynthesis and Lipid Metabolism. *Journal of Bacteriology*, 187(5), 1677. <https://doi.org/10.1128/JB.187.5.1677-1684.2005>
- Birhanu, A. G., Yimer, S. A., Kalayou, S., Riaz, T., Zegeye, E. D., Holm-Hansen, C., Norheim, G., Aseffa, A., Abebe, M., & Tønnum, T. (2019). Ample glycosylation in membrane and cell envelope proteins may explain the phenotypic diversity and virulence in the Mycobacterium tuberculosis complex. *Scientific Reports* 2019 9:1, 9(1), 1–15. <https://doi.org/10.1038/s41598-019-39654-9>

- Bizzell, E., Sia, J. K., Quezada, M., Enriquez, A., Georgieva, M., & Rengarajan, J. (2018). Deletion of BCG Hip1 protease enhances dendritic cell and CD4 T cell responses. *Journal of Leukocyte Biology*, 103(4), 739–748. <https://doi.org/10.1002/JLB.4A0917-363RR>
- Bohlin, J., & H-O Pettersson, J. (2019). Evolution of Genomic Base Composition: From Single Cell Microbes to Multicellular Animals. *Computational and Structural Biotechnology Journal*, 17, 362–370. <https://doi.org/10.1016/j.csbj.2019.03.001>
- Bok, J., Hofland, R. W., & Evans, C. A. (2021). Whole Blood Mycobacterial Growth Assays for Assessing Human Tuberculosis Susceptibility: A Systematic Review and Meta-Analysis. *Frontiers in Immunology*, 12, 904. <https://doi.org/10.3389/FIMMU.2021.641082/BIBTEX>
- Boro, M., & Balaji, K. N. (2017). CXCL1 and CXCL2 Regulate NLRP3 Inflammasome Activation via G-Protein–Coupled Receptor CXCR2. *The Journal of Immunology*, 199(5), 1660–1671. <https://doi.org/10.4049/JIMMUNOL.1700129>
- Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S. A., Bryant, J. M., Harris, S. R., Schuenemann, V. J., Campbell, T. J., Majander, K., Wilbur, A. K., Guichon, R. A., Steadman, D. L. W., Cook, D. C., Niemann, S., Behr, M. A., Zumarraga, M., ... Krause, J. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* 2014 514:7523, 514(7523), 494–497. <https://doi.org/10.1038/nature13591>
- Brace, P. T., Tezera, L. B., Bielecka, M. K., Mellows, T., Garay, D., Tian, S., Rand, L., Green, J., Jogai, S., Steele, A. J., Millar, T. M., Sanchez-Elsner, T., Friedland, J. S., Proud, C. G., & Elkington, P. T. (2017). Mycobacterium tuberculosis subverts negative regulatory pathways in human macrophages to drive immunopathology. *PLOS Pathogens*, 13(6), e1006367. <https://doi.org/10.1371/JOURNAL.PPAT.1006367>
- Brites, D., & Gagneux, S. (2015). Co-evolution of Mycobacterium tuberculosis and Homo sapiens. *Immunological Reviews*, 264(1), 6–24. <https://doi.org/10.1111/IMR.12264>
- Broderick, C., Cliff, J. M., Lee, J. S., Kaforou, M., & Moore, D. A. (2021). Host transcriptional response to TB preventive therapy differentiates two sub-groups of IGRA-positive individuals. *Tuberculosis*, 127, 102033. <https://doi.org/10.1016/J.TUBE.2020.102033>
- Brosch, R., Gordon, S. v., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L. M., Pym, A. S., Samper, S., Soolingen, D. van, & Cole, S. T. (2002). A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proceedings of the National Academy of Sciences*, 99(6), 3684–3689. <https://doi.org/10.1073/PNAS.052548299>
- Brudey, K., Driscoll, J. R., Rigouts, L., Prodinger, W. M., Gori, A., Al-Hajoj, S. A., Allix, C., Aristimuño, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J. T., Fauville-Dufaux, M., Ferdinand, S., ... Sola, C. (2006). Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiology*, 6. <https://doi.org/10.1186/1471-2180-6-23>
- Bucsan, A. N., Mehra, S., Khader, S. A., & Kaushal, D. (2019). The current state of animal models and genomic approaches towards identifying and validating molecular determinants of

- Mycobacterium tuberculosis infection and tuberculosis disease. *Pathogens and Disease*, 77(4). <https://doi.org/10.1093/FEMSPD/FTZ037>
- Budhraj, A., Basu, A., Gheware, A., Abhilash, D., Rajagopala, S., Pakala, S., Sumit, M., Ray, A., Subramaniam, A., Mathur, P., Nambirajan, A., Kumar, S., Gupta, R., Wig, N., Trikha, A., Guleria, R., Sarkar, C., Gupta, I., & Jain, D. (2022). Molecular signature of postmortem lung tissue from COVID-19 patients suggests distinct trajectories driving mortality. *DMM Disease Models and Mechanisms*, 15(5). <https://doi.org/10.1242/DMM.049572/275032/AM/MOLECULAR-SIGNATURE-OF-POSTMORTEM-LUNG-TISSUE-FROM>
- Bulut, Y., Michelsen, K. S., Hayrapetian, L., Naiki, Y., Spallek, R., Singh, M., & Ardit, M. (2005). Mycobacterium tuberculosis heat shock proteins use diverse Toll-like receptor pathways to activate pro-inflammatory signals. *The Journal of Biological Chemistry*, 280(22), 20961–20967. <https://doi.org/10.1074/JBC.M411379200>
- Calmette, A., & Guérin, C. (1920). *Ann. Inst. Pasteur*, 34, 553–561.
- Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., & Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications* 2021 12:1, 12(1), 1–12. <https://doi.org/10.1038/s41467-020-20430-7>
- Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S. R., Ignatyeva, O., Kontsevaya, I., Corander, J., Bryant, J., Parkhill, J., Nejentsev, S., Horstmann, R. D., Brown, T., & Drobniowski, F. (2014). Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature Genetics* 2014 46:3, 46(3), 279–286. <https://doi.org/10.1038/ng.2878>
- Chackerian, A. A., & Behar, S. M. (2003). Susceptibility to Mycobacterium tuberculosis: lessons from inbred strains of mice. *Tuberculosis*, 83(5), 279–285. [https://doi.org/10.1016/S1472-9792\(03\)00017-9](https://doi.org/10.1016/S1472-9792(03)00017-9)
- Chai, Q., Wang, L., Liu, C. H., & Ge, B. (2020). New insights into the evasion of host innate immunity by Mycobacterium tuberculosis. *Cellular & Molecular Immunology* 2020 17:9, 17(9), 901–913. <https://doi.org/10.1038/s41423-020-0502-z>
- Chan, J., Mehta, S., Bharrhan, S., Chen, Y., Achkar, J. M., Casadevall, A., & Flynn, J. A. (2014). The role of B cells and humoral immunity in Mycobacterium tuberculosis infection. *Seminars in Immunology*, 26(6), 588–600. <https://doi.org/10.1016/J.SMIM.2014.10.005>
- Chauvel, C., Novoloaca, A., Veyre, P., Reynier, F., & Becker, J. (2020). Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings in Bioinformatics*, 21(2), 541–552. <https://doi.org/10.1093/BIB/BBZ015>
- Chen, Y.-Y., Chang, J.-R., Huang, W.-F., Hsu, S.-C., Kuo, S.-C., Sun, J.-R., & Dou, H.-Y. (2014). The Pattern of Cytokine Production In Vitro Induced by Ancient and Modern Beijing Mycobacterium tuberculosis Strains. *PLOS ONE*, 9(4), e94296. <https://doi.org/10.1371/JOURNAL.PONE.0094296>
- Choi, Y., & Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, 31(16), 2745–2747. <https://doi.org/10.1093/BIOINFORMATICS/BTV195>
- Cilloni, L., Fu, H., Vesga, J. F., Dowdy, D., Pretorius, C., Ahmedov, S., Nair, S. A., Mosneaga, A., Masini, E., Sahu, S., & Arinaminpathy, N. (2020). The potential impact of the COVID-19 pandemic on

- the tuberculosis epidemic a modelling analysis. *EClinicalMedicine*, 28, 100603.
<https://doi.org/10.1016/J.ECLINM.2020.100603>
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 6(2), 80–92.
https://doi.org/10.4161/FLY.19695/SUPPL_FILE/KFLY_A_10919695_SM0001.ZIP
- Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. v., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., ... Barrell, B. G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 1998 396:6707, 396(6707), 190–190. <https://doi.org/10.1038/24206>
- Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., Kato-Maeda, M., Galagan, J., Niemann, S., & Gagneux, S. (2011). Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics* 2011 44:1, 44(1), 106–110. <https://doi.org/10.1038/ng.1038>
- Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G., Yeboah-Manu, D., Bothamley, G., Mei, J., Wei, L., Bentley, S., Harris, S. R., Niemann, S., Diel, R., Aseffa, A., ... Gagneux, S. (2013). Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature Genetics*, 45(10), 1176–1182. <https://doi.org/10.1038/ng.2744>
- Correa-Macedo, W., Cambri, G., & Schurr, E. (2019). The Interplay of Human and *Mycobacterium Tuberculosis* Genomic Variability. *Frontiers in Genetics*, 0, 865.
<https://doi.org/10.3389/FGENE.2019.00865>
- Coscolla, M., Gagneux, S., Menardo, F., Loiseau, C., Ruiz-Rodriguez, P., Borrell, S., Darko Otchere, I., Asante-Poku, A., Asare, P., Sánchez-Busó, L., Gehre, F., Sanoussi, D., Antonio, M., Affolabi, D., Fyfe, J., Beckert, P., Niemann, S., Alabi, A. S., Grobusch, M. P., ... Brites, D. (2021). Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. *Microbial Genomics*, 7, 477. <https://doi.org/10.1099/mgen.0.000477>
- Couvin, D., David, A., Zozio, T., & Rastogi, N. (2019). Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the *Mycobacterium tuberculosis* genotyping database. *Infection, Genetics and Evolution*, 72, 31–43.
<https://doi.org/10.1016/J.MEEGID.2018.12.030>
- Cristina, M. G., Brisse, S., Brosch, R., Fabre, M., Omaïs, B., Marmiesse, M., Supply, P., & Vincent, V. (2005). Ancient Origin and Gene Mosaicism of the Progenitor of *Mycobacterium tuberculosis*. *PLOS Pathogens*, 1(1), e5. <https://doi.org/10.1371/JOURNAL.PPAT.0010005>
- Cumming, B. M., Addicott, K. W., Adamson, J. H., & Steyn, A. J. C. (2018). *Mycobacterium tuberculosis* induces decelerated bioenergetic metabolism in human macrophages. *ELife*, 7.
<https://doi.org/10.7554/ELIFE.39169>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and

- VCFTools. *Bioinformatics*, 27(15), 2156–2158.
<https://doi.org/10.1093/BIOINFORMATICS/BTR330>
- de Araujo, L. S., Vaas, L. A. I., Ribeiro-Alves, M., Geffers, R., Mello, F. C. Q., de Almeida, A. S., Moreira, A. da S. R., Kritski, A. L., Lapa e Silva, J. R., Moraes, M. O., Pessler, F., & Saad, M. H. F. (2016). Transcriptomic Biomarkers for Tuberculosis: Evaluation of DOCK9, EPHA4, and NPC2 mRNA Expression in Peripheral Blood. *Frontiers in Microbiology*, 7(OCT), 1586.
<https://doi.org/10.3389/fmicb.2016.01586>
- de Carvalho, L. P. S., Fischer, S. M., Marrero, J., Nathan, C., Ehrt, S., & Rhee, K. Y. (2010). Metabolomics of Mycobacterium tuberculosis Reveals Compartmentalized Co-Catabolism of Carbon Substrates. *Chemistry & Biology*, 17(10), 1122–1131.
<https://doi.org/10.1016/J.CHEMBIOL.2010.08.009>
- de Martino, M., Lodi, L., Galli, L., & Chiappini, E. (2019). Immune Response to Mycobacterium tuberculosis: A Narrative Review. *Frontiers in Pediatrics*, 0, 350.
<https://doi.org/10.3389/fped.2019.00350>
- de Voss, J. J., Rutter, K., Schroeder, B. G., Su, H., Zhu, Y., & Barry, C. E. (2000). The salicylate-derived mycobactin siderophores of Mycobacterium tuberculosis are essential for growth in macrophages. *Proceedings of the National Academy of Sciences of the United States of America*, 97(3), 1252–1257. <https://doi.org/10.1073/PNAS.97.3.1252/ASSET/2267DDB2-9FB8-4866-B15A-AE885A7D27C3/ASSETS/GRAPHIC/PQ0201955005.JPEG>
- di Pietrantonio, T., Correa, J. A., Orlova, M., Behr, M. A., & Schurr, E. (2011). Joint effects of host genetic background and mycobacterial pathogen on susceptibility to infection. *Infection and Immunity*, 79(6), 2372–2378. <https://doi.org/10.1128/IAI.00985-10>
- DiNardo, A. R., Gandhi, T., Heyckendorf, J., Grimm, S. L., Rajapakshe, K., Nishiguchi, T., Reimann, M., Kahari, J., Dlamini, Q., Lange, C., Marwitz, S., cohort study group, D., Cirillo, J. D., Netea, M. G., van Crevel, R., Mandalakas, A. M., Coarfa, C., Avsar, K., DiNardo, A., ... van, F. (2021). Gene expression signatures identify biologically and clinically distinct tuberculosis endotypes. *MedRxiv*, 19, 2020.05.13.20100776. <https://doi.org/10.1101/2020.05.13.20100776>
- Dinarello, C. A. (2018). Overview of the IL-1 family in innate inflammation and acquired immunity. *Immunological Reviews*, 281(1), 8–27. <https://doi.org/10.1111/IMR.12621>
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21.
<https://doi.org/10.1093/bioinformatics/bts635>
- Domaszewska, T., Scheuermann, L., Hahnke, K., Mollenkopf, H., Dorhoi, A., Kaufmann, S. H. E., & Weiner, J. (2017). Concordant and discordant gene expression patterns in mouse strains identify best-fit animal model for human tuberculosis OPEN. <https://doi.org/10.1038/s41598-017-11812-x>
- Dorman, S. E., Schumacher, S. G., Alland, D., Nabeta, P., Armstrong, D. T., King, B., Hall, S. L., Chakravorty, S., Cirillo, D. M., Tukvadze, N., Bablishvili, N., Stevens, W., Scott, L., Rodrigues, C., Kazi, M. I., Joloba, M., Nakiyingi, L., Nicol, M. P., Ghebrekristos, Y., ... Xie, Y. (2018). Xpert MTB/RIF Ultra for detection of Mycobacterium tuberculosis and rifampicin resistance: a prospective multicentre diagnostic accuracy study. *The Lancet Infectious Diseases*, 18(1), 76–84. [https://doi.org/10.1016/S1473-3099\(17\)30691-6](https://doi.org/10.1016/S1473-3099(17)30691-6)

- Eoh, H., & Rhee, K. Y. (2014). Methylcitrate cycle defines the bactericidal essentiality of isocitrate lyase for survival of *Mycobacterium tuberculosis* on fatty acids. *Proceedings of the National Academy of Sciences*, 111(13), 4976–4981. <https://doi.org/10.1073/PNAS.1400390111>
- Eum, S. Y., Kong, J. H., Hong, M. S., Lee, Y. J., Kim, J. H., Hwang, S. H., Cho, S. N., Via, L. E., & Barry, C. E. (2010). Neutrophils Are the Predominant Infected Phagocytic Cells in the Airways of Patients With Active Pulmonary TB. *Chest*, 137(1), 122–128. <https://doi.org/10.1378/CHEST.09-0903>
- Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
- Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., Warren, R. M., Streicher, E. M., Calver, A., Sloutsky, A., Kaur, D., Posey, J. E., Plikaytis, B., Oggioni, M. R., Gardy, J. L., Johnston, J. C., Rodrigues, M., Tang, P. K. C., Kato-Maeda, M., ... Murray, M. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nature Genetics* 2013 45:10, 45(10), 1183–1189. <https://doi.org/10.1038/ng.2747>
- Feng, J., Cao, Z., Wang, L., Wan, Y., Peng, N., Wang, Q., Chen, X., Zhou, Y., & Zhu, Y. (2017). Inducible GBP5 Mediates the Antiviral Response via Interferon-Related Pathways during Influenza A Virus Infection. *Journal of Innate Immunity*, 9(4), 419–435. <https://doi.org/10.1159/000460294>
- Fletcher, H. A., Tanner, R., Wallis, R. S., Meyer, J., Manjaly, Z. R., Harris, S., Satti, I., Richard F, S., Hoft, D., Kampmann, B., Walker, K. B., Dockrell, H. M., Fruth, U., Barker, L., Brennan, M. J., & McShane, H. (2013). Inhibition of mycobacterial growth in vitro following primary but not secondary vaccination with *Mycobacterium bovis* BCG. *Clinical and Vaccine Immunology*, 20(11), 1683–1689. <https://doi.org/10.1128/CVI.00427-13>
- Flynn, J. A. L. (2006). Lessons from experimental *Mycobacterium tuberculosis* infections. *Microbes and Infection*, 8(4), 1179–1188. <https://doi.org/10.1016/J.MICINF.2005.10.033>
- Fonseca, K. L., Rodrigues, P. N. S., Olsson, I. A. S., & Saraiva, M. (2017). Experimental study of tuberculosis: From animal models to complex cell systems and organoids. *PLOS Pathogens*, 13(8), e1006421. <https://doi.org/10.1371/JOURNAL.PPAT.1006421>
- Fontán, P., Aris, V., Ghanny, S., Soteropoulos, P., & Smith, I. (2008). Global transcriptional profile of *Mycobacterium tuberculosis* during THP-1 human macrophage infection. *Infection and Immunity*, 76(2), 717–725. <https://doi.org/10.1128/IAI.00974-07>
- Gagneux, S. (2018). Ecology and evolution of *Mycobacterium tuberculosis*. In *Nature Reviews Microbiology* (Vol. 16, Issue 4, pp. 202–213). Nature Publishing Group. <https://doi.org/10.1038/nrmicro.2018.8>
- Geginat, J., Paroni, M., Maglie, S., Alfen, J. S., Kastirr, I., Gruarin, P., de Simone, M., Pagani, M., & Abrignani, S. (2014). Plasticity of human CD4 T cell subsets. *Frontiers in Immunology*, 5(DEC), 630. <https://doi.org/10.3389/FIMMU.2014.00630/BIBTEX>
- Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., Deng, C., Varusai, T., Ragueneau, E., Haider, Y., May, B., Shamovsky, V., Weiser, J., Brunson, T., Sanati, N., ... D'Eustachio, P. (2022). The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1), D687–D692. <https://doi.org/10.1093/NAR/GKAB1028>

- GjØen, J. E., Jenum, S., Sivakumaran, D., Mukherjee, A., MacAden, R., Kabra, S. K., Lodha, R., Ottenhoff, T. H. M., Haks, M. C., Doherty, T. M., Ritz, C., & Grewal, H. M. S. (2017). Novel transcriptional signatures for sputum-independent diagnostics of tuberculosis in children. *Scientific Reports*, 7(1), 1–9. <https://doi.org/10.1038/s41598-017-05057-x>
- Gliddon, H. D., Herberg, J. A., Levin, M., & Kaforou, M. (2018). Genome-wide host RNA signatures of infectious diseases: discovery and clinical translation. *Immunology*, 153(2), 171–178. <https://doi.org/10.1111/IMM.12841>
- Graham, S. M., Ahmed, T., Amanullah, F., Browning, R., Cardenas, V., Casenghi, M., Cuevas, L. E., Gale, M., Gie, R. P., Grzemska, M., Handelsman, E., Hatherill, M., Hesselning, A. C., Jean-Philippe, P., Kampmann, B., Kabra, S. K., Lienhardt, C., Lighter-Fisher, J., Madhi, S., ... Wingfield, C. (2012). Evaluation of Tuberculosis Diagnostics in Children: 1. Proposed Clinical Case Definitions for Classification of Intrathoracic Tuberculosis Disease. Consensus From an Expert Panel. *The Journal of Infectious Diseases*, 205(suppl_2), S199–S208. <https://doi.org/10.1093/INFDIS/JIS008>
- Graustein, A. D., Horne, D. J., Arentz, M., Bang, N. D., Chau, T. T. H., Thwaites, G. E., Caws, M., Thuong, N. T. T., Dunstan, S. J., & Hawn, T. R. (2015). TLR9 gene region polymorphisms and susceptibility to tuberculosis in Vietnam. *Tuberculosis*, 95(2), 190–196. <https://doi.org/10.1016/J.TUBE.2014.12.009>
- Gupta, R. K., Turner, C. T., Venturini, C., Esmail, H., Rangaka, M. X., Copas, A., Lipman, M., Abubakar, I., & Noursadeghi, M. (2020). Concise whole blood transcriptional signatures for incipient tuberculosis: a systematic review and patient-level pooled meta-analysis. *The Lancet Respiratory Medicine*, 8(4), 395–406. [https://doi.org/10.1016/S2213-2600\(19\)30282-6](https://doi.org/10.1016/S2213-2600(19)30282-6)
- Haas, M. K., & Belknap, R. W. (2019). Diagnostic Tests for Latent Tuberculosis Infection. *Clinics in Chest Medicine*, 40(4), 829–837. <https://doi.org/10.1016/J.CCM.2019.07.007>
- Hamilton, J. A. (2019). GM-CSF-Dependent Inflammatory Pathways. *Frontiers in Immunology*, 10, 2055. <https://doi.org/10.3389/FIMMU.2019.02055/BIBTEX>
- Han, Y., Gao, S., Muegge, K., Zhang, W., & Zhou, B. (2015). Advanced Applications of RNA Sequencing and Challenges: <https://doi.org/10.4137/BBI.S28991>, 9, 29–46. <https://doi.org/10.4137/BBI.S28991>
- Harrell Jr, F. E. (2021). *Frank E Harrell Jr, 2021. Hmisc: Harrell Miscellaneous. R package version 4.6-0.* <https://CRAN.R-project.org/package=Hmisc>.
- Herberg, J. A., Kaforou, M., Wright, V. J., Shales, H., Eleftherohorinou, H., Hoggart, C. J., Cebey-López, M., Carter, M. J., Janes, V. A., Gormley, S., Shimizu, C., Tremoulet, A. H., Barendregt, A. M., Salas, A., Kanegaye, J., Pollard, A. J., Faust, S. N., Patel, S., Kuijpers, T., ... Consortium, for the I. (2016). Diagnostic Test Accuracy of a 2-Transcript Host RNA Signature for Discriminating Bacterial vs Viral Infection in Febrile Children. *JAMA*, 316(8), 835–845. <https://doi.org/10.1001/JAMA.2016.11236>
- Hmama, Z., Peña-Díaz, S., Joseph, S., & Av-Gay, Y. (2015). Immuno-evasion and immunosuppression of the macrophage by Mycobacterium tuberculosis. *Immunological Reviews*, 264(1), 220–232. <https://doi.org/10.1111/IMR.12268>
- Hołowka, J., Trojanowski, D., Gind, K., Wojtaś, B., Gielniewski, B., Jakimowicz, D., & Zakrzewska-Czerwińska, J. (2017). HupB is a bacterial nucleoid-associated protein with an indispensable

- eukaryotic-like tail. *MBio*, 8(6). https://doi.org/10.1128/MBIO.01272-17/SUPPL_FILE/MBO006173577ST1.DOCX
- Hong, E. P., & Park, J. W. (2012). Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics & Informatics*, 10(2), 117–122. <https://doi.org/10.5808/GI.2012.10.2.117>
- Huang, L., Nazarova, E. v, Tan, S., Liu, Y., & Russell, D. G. (2018). Growth of Mycobacterium tuberculosis in vivo segregates with host macrophage metabolism and ontogeny. *J. Exp. Med*, 215, 1135–1152. <https://doi.org/10.1084/jem.20172020>
- Irene, C.-C. M., Germán, R.-C. J., Gamaliel, L.-L., Adriana, M.-E. D., Isabel, B. E., Nohemí, M. B., Jorge, B. P., Lucía, Z.-B. Z., Myriam, B. del V., Fernanda, C.-G., Adrian, O.-L., Isabel, M. M., & Rogelio, H.-P. (2021). Profiling the immune response to Mycobacterium tuberculosis Beijing family infection: a perspective from the transcriptome. <https://doi.org/10.1080/21505594.2021.1936432>, 12(1), 1689–1704.
- Ittisoponpisan, S., Islam, S. A., Khanna, T., Alhuzimi, E., David, A., & Sternberg, M. J. E. (2019). Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *Journal of Molecular Biology*, 431(11), 2197–2212. <https://doi.org/10.1016/J.JMB.2019.04.009>
- Jagielski, T., van Ingen, J., Rastogi, N., Dziadek, J., Mazur, P. K., & Bielecki, J. (2014). *Current Methods in the Molecular Typing of Mycobacterium tuberculosis and Other Mycobacteria*. <https://doi.org/10.1155/2014/645802>
- January Weiner. (2020). *January Weiner (2020). tmod: Feature Set Enrichment Analysis for Metabolomics and Transcriptomics. R package version 0.46.2.* <https://CRAN.R-project.org/package=tmod>.
- Jenum, S., Dhanasekaran, S., Lodha, R., Mukherjee, A., Kumar Saini, D., Singh, S., Singh, V., Medigeshi, G., Haks, M. C., Ottenhoff, T. H. M., Doherty, T. M., Kabra, S. K., Ritz, C., & Grewal, H. M. S. (2016). Approaching a diagnostic point-of-care test for pediatric tuberculosis through evaluation of immune biomarkers across the clinical disease spectrum. *Scientific Reports*, 6(1), 1–11. <https://doi.org/10.1038/srep18520>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 2021 596:7873, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kanabalan, R. D., Lee, L. J., Lee, T. Y., Chong, P. P., Hassan, L., Ismail, R., & Chin, V. K. (2021). Human tuberculosis and Mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery. *Microbiological Research*, 246, 126674. <https://doi.org/10.1016/J.MICRES.2020.126674>
- Kapopoulou, A., Lew, J. M., & Cole, S. T. (2011). The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis*, 91(1), 8–13. <https://doi.org/10.1016/J.TUBE.2010.09.006>
- Karatzoglou, A., Hornik, K., Smola, A., & Zeileis, A. (2004). kernlab - An S4 package for kernel methods in R. *Journal of Statistical Software*, 11(1), 1–20. <https://doi.org/10.18637/jss.v011.i09>

- Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., & Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, 20(4), 1085–1093. <https://doi.org/10.1093/BIB/BBX085>
- Keiser, T. L., & Purdy, G. E. (2017). Killing Mycobacterium tuberculosis In Vitro: What Model Systems Can Teach Us. *Microbiology Spectrum*, 5(3). <https://doi.org/10.1128/MICROBIOLSPEC.TBTB2-0028-2016>
- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* 2015 10:6, 10(6), 845–858. <https://doi.org/10.1038/nprot.2015.053>
- Khan, S. F., Damerell, V., Omar, R., du Toit, M., Khan, M., Maranyane, H. M., Mlaza, M., Bleloch, J., Bellis, C., Sahm, B. D. B., Peres, J., ArulJothi, K. N., & Prince, S. (2020). The roles and regulation of TBX3 in development and disease. *Gene*, 726, 144223. <https://doi.org/10.1016/J.GENE.2019.144223>
- Ki, H. P., & Shingadia, D. (2017). Tuberculosis in children. *Paediatrics and Child Health*, 27(3), 109–115. <https://doi.org/10.1016/J.PAED.2016.12.004>
- Kim, B.-H., Chee, J. D., Bradfield, C. J., Park, E.-S., Kumar, P., & MacMicking, J. D. (2016). Interferon-induced guanylate-binding proteins in inflammasome activation and host defense. *Nature Immunology* 2016 17:5, 17(5), 481–489. <https://doi.org/10.1038/ni.3440>
- Knolle, J., Pierau, M., Hebel, K., Lampe, K., Jorch, G., Kropf, S., Arens, C., & Brunner-Weinzierl, M. C. (2020). Children From the Age of Three Show a Developmental Switch in T-Cell Differentiation. *Frontiers in Immunology*, 11, 1640. <https://doi.org/10.3389/FIMMU.2020.01640/BIBTEX>
- Krämer, A., Green, J., Pollard, J., & Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, 30(4), 523–530. <https://doi.org/10.1093/BIOINFORMATICS/BTT703>
- Krapp, C., Hotter, D., Gawanbacht, A., McLaren, P. J., Kluge, S. F., Stürzel, C. M., Mack, K., Reith, E., Engelhart, S., Ciuffi, A., Hornung, V., Sauter, D., Telenti, A., & Kirchhoff, F. (2016). Guanylate Binding Protein (GBP) 5 Is an Interferon-Inducible Inhibitor of HIV-1 Infectivity. *Cell Host & Microbe*, 19(4), 504–514. <https://doi.org/10.1016/J.CHOM.2016.02.019>
- Krishnan, N., Malaga, W., Constant, P., Caws, M., Thi, T., Chau, H., Salmons, J., Thi, N., Lan, N., Duc Bang, N., Daffé, M., Young, D. B., Robertson, B. D., Guilhot, C., & Thwaites, G. E. (2011). *Mycobacterium tuberculosis* Lineage Influences Innate Immune Response and Virulence and Is Associated with Distinct Cell Envelope Lipid Profiles. <https://doi.org/10.1371/journal.pone.0023870>
- Krishnan, S., Queiroz, A. T. L., Gupta, A., Gupte, N., Bisson, G. P., Kumwenda, J., Naidoo, K., Mohapi, L., Mave, V., Mngqibisa, R., Lama, J. R., Hosseinipour, M. C., Andrade, B. B., & Karakousis, P. C. (2021). Integrative Multi-Omics Reveals Serum Markers of Tuberculosis in Advanced HIV. *Frontiers in Immunology*, 12, 2060. <https://doi.org/10.3389/FIMMU.2021.676980/BIBTEX>
- Kurthkoti, K., Tare, P., Paitchowdhury, R., Gowthami, V. N., Garcia, M. J., Colangeli, R., Chatterji, D., Nagaraja, V., & Rodriguez, G. M. (2015). The mycobacterial iron-dependent regulator IdeR induces ferritin (bfrB) by alleviating Lsr2 repression. *Molecular Microbiology*, 98(5), 864–877. <https://doi.org/10.1111/MMI.13166/SUPPINFO>

- Lai, R. P. J., Cortes, T., Marais, S., Rockwood, N., Burke, M. L., Garza-Garcia, A., Horswell, S., Sesay, A. K., O'garra, A., Young, D. B., & Wilkinson, R. J. (2021). *Transcriptomic Characterization of Tuberculous Sputum Reveals a Host Warburg Effect and Microbial Cholesterol Catabolism*.
- Lee, J., Boyce, S., Powers, J., Baer, C., Sasseti, C. M., & Behar, S. M., Jr. (2020). CD11cHi monocyte-derived macrophages are a major cellular compartment infected by Mycobacterium tuberculosis. *PLOS Pathogens*, 16(6), e1008621. <https://doi.org/10.1371/JOURNAL.PPAT.1008621>
- Lee, S., Kang, J., & Kim, J. (2019). Structural and biochemical characterization of Rv0187, an O-methyltransferase from Mycobacterium tuberculosis. *Scientific Reports* 2019 9:1, 9(1), 1–12. <https://doi.org/10.1038/s41598-019-44592-7>
- Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6), 882. <https://doi.org/10.1093/BIOINFORMATICS/BTS034>
- Leong, S., Zhao, Y., Ribeiro-Rodrigues, R., Jones-López, E. C., Acuña-Villaorduña, C., Rodrigues, P. M., Palaci, M., Alland, D., Dietze, R., Ellner, J. J., Johnson, W. E., & Salgame, P. (2020). Cross-validation of existing signatures and derivation of a novel 29-gene transcriptomic signature predictive of progression to TB in a Brazilian cohort of household contacts of pulmonary TB. *Tuberculosis*, 120, 101898. <https://doi.org/10.1016/j.tube.2020.101898>
- Levy, S. E., & Myers, R. M. (2016). Advancements in Next-Generation Sequencing. <http://Dx.Doi.Org/10.1146/Annurev-Genom-083115-022413>, 17, 95–115. <https://doi.org/10.1146/ANNUREV-GENOM-083115-022413>
- Ley, S. D., de Vos, M., van Rie, A., & Warren, R. M. (2019). Deciphering Within-Host Microevolution of Mycobacterium tuberculosis through Whole-Genome Sequencing: the Phenotypic Impact and Way Forward. *Microbiology and Molecular Biology Reviews*, 83(2). https://doi.org/10.1128/MMBR.00062-18/SUPPL_FILE/MMBR.00062-18-SD002.XLSX
- Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. <http://github.com/lh3/bwa>.
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5), 589–595. <https://doi.org/10.1093/BIOINFORMATICS/BTP698>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, Q., Yu, L., Xin, T., Du, S., Yi, X., Zhang, B., Wu, B., & Sun, H. (2015). Increased IL-9 mRNA expression as a biomarker to diagnose childhood tuberculosis in a high burden settings. In *Journal of Infection* (Vol. 71, Issue 2, pp. 273–276). W.B. Saunders Ltd. <https://doi.org/10.1016/j.jinf.2015.04.002>
- Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7), 923–930. <https://doi.org/10.1093/bioinformatics/btt656>
- Liaw, A., & Wiener, M. (2002). *Classification and Regression by randomForest*. 2(3). <http://www.stat.berkeley.edu/>

- Lin Aung, H., Berney, M., & Cook, G. M. (2014). *Hypoxia-Activated Cytochrome bd Expression in Mycobacterium smegmatis Is Cyclic AMP Receptor Protein Dependent*. <https://doi.org/10.1128/JB.01771-14>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Maecker, H. T., Frey, T., Nomura, L. E., & Trotter, J. (2004). Selecting fluorochrome conjugates for maximum sensitivity. *Cytometry Part A*, 62A(2), 169–173. <https://doi.org/10.1002/CYTO.A.20092>
- Mahairas, G. G., Sabo, P. J., Hickey, M. J., Singh, D. C., & Stover, C. K. (1996). Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. *Journal of Bacteriology*, 178(5), 1274–1282. <https://doi.org/10.1128/JB.178.5.1274-1282.1996>
- Mandal, N., Anand, P. K., Gautam, S., Das, S., & Hussain, T. (2017). Diagnosis and treatment of paediatric tuberculosis: An insight review. [Http://Dx.Doi.Org/10.1080/1040841X.2016.1262813](http://Dx.Doi.Org/10.1080/1040841X.2016.1262813), 43(4), 466–480. <https://doi.org/10.1080/1040841X.2016.1262813>
- Marrakchi, H., Lan  elle, G., & Qu  emard, A. (2000). InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II. *Microbiology*, 146(2), 289–296. <https://doi.org/10.1099/00221287-146-2-289/CITE/REFWORKS>
- Martineau, A. R., Newton, S. M., Wilkinson, K. A., Kampmann, B., Hall, B. M., Nawroly, N., Packe, G. E., Davidson, R. N., Griffiths, C. J., & Wilkinson, R. J. (2007). Neutrophil-mediated innate immune resistance to mycobacteria. *Journal of Clinical Investigation*, 117(7), 1988–1994. <https://doi.org/10.1172/JCI31097>
- Martineau, A. R., Wilkinson, R. J., Wilkinson, K. A., Newton, S. M., Kampmann, B., Hall, B. M., Packe, G. E., Davidson, R. N., Eldridge, S. M., Maunsell, Z. J., Rainbow, S. J., Berry, J. L., & Griffiths, C. J. (2012). A Single Dose of Vitamin D Enhances Immunity to Mycobacteria. <https://doi.org/10.1164/Rccm.200701-007OC>, 176(2), 208–213. <https://doi.org/10.1164/RCCM.200701-007OC>
- McNab, F. W., Berry, M. P. R., Graham, C. M., Bloch, S. A. A., Oni, T., Wilkinson, K. A., Wilkinson, R. J., Kon, O. M., Banchereau, J., Chaussabel, D., & O’Garra, A. (2011). Programmed death ligand 1 is over-expressed by neutrophils in the blood of patients with active tuberculosis. *European Journal of Immunology*, 41(7), 1941–1947. <https://doi.org/10.1002/EJI.201141421>
- McNab, F. W., Ewbank, J., Howes, A., Moreira-Teixeira, L., Martirosyan, A., Ghilardi, N., Saraiva, M., & O’Garra, A. (2014). Type I IFN Induces IL-10 Production in an IL-27–Independent Manner and Blocks Responsiveness to IFN-  for Production of IL-12 and Bacterial Killing in *Mycobacterium tuberculosis*–Infected Macrophages. *The Journal of Immunology*, 193(7), 3600–3612. <https://doi.org/10.4049/JIMMUNOL.1401088/-/DCSUPPLEMENTAL>
- McNerney, R., Maeurer, M., Abubakar, I., Marais, B., McHugh, T. D., Ford, N., Weyer, K., Lawn, S., Grobusch, M. P., Memish, Z., Squire, S. B., Pantaleo, G., Chakaya, J., Casenghi, M., Migliori, G. B., Mwaba, P., Zijenah, L., Hoelscher, M., Cox, H., ... Zumla, A. (2012). Tuberculosis Diagnostics and Biomarkers: Needs, Challenges, Recent Advances, and Opportunities. *The Journal of Infectious Diseases*, 205(suppl_2), S147–S158. <https://doi.org/10.1093/INFDIS/JIR860>

- Meeting report of the WHO expert consultation on the definition of extensively drug-resistant tuberculosis. (2020). *Meeting report of the WHO expert consultation on the definition of extensively drug-resistant tuberculosis*. <https://www.who.int/publications/i/item/meeting-report-of-the-who-expert-consultation-on-the-definition-of-extensively-drug-resistant-tuberculosis>
- Meier, N. R., Volken, T., Geiger, M., Heininger, U., Tebruegge, M., & Ritz, N. (2019). Risk factors for indeterminate interferon-gamma release assay for the diagnosis of tuberculosis in children - A systematic review and meta-analysis. *Frontiers in Pediatrics*, 7, 208. <https://doi.org/10.3389/FPED.2019.00208/BIBTEX>
- Meissner-Roloff, R. J., Koekemoer, G., Warren, R. M., & Loots, D. T. (2012). A metabolomics investigation of a hyper- and hypo-virulent phenotype of Beijing lineage M. tuberculosis. *Metabolomics* 2012 8:6, 8(6), 1194–1203. <https://doi.org/10.1007/S11306-012-0424-6>
- Menardo, F., Duchêne, S., Brites, D., & Gagneux, S. (2019). The molecular clock of Mycobacterium tuberculosis. *PLOS Pathogens*, 15(9), e1008067. <https://doi.org/10.1371/JOURNAL.PPAT.1008067>
- Mendoza-Coronel, E., & Castanon-Arreola, M. (2016). Comparative evaluation of in vitro human macrophage models for mycobacterial infection study. *Pathogens and Disease*, 74(6), 52. <https://doi.org/10.1093/FEMSPD/FTW052>
- Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., B Blum, M. G., Rüscher-Gerdes, S., Mokrousov, I., Aleksic, E., Allix-Béguec, C., Antierens, A., Augustynowicz-Kopeć, E., Ballif, M., Barletta, F., Peter Beck, H., Barry, C. E., Bonnet, M., Borroni, E., ... Wirth, T. (2015). Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nature Publishing Group*. <https://doi.org/10.1038/ng.3195>
- Miao, Y.-R., Zhang, Q., Lei, Q., Luo, M., Xie, G.-Y., Wang, H., Guo, A.-Y., Miao, Y., Zhang, Q., Lei, Q., Luo, M., Xie, G., Guo, A., & Wang, H. (2020). 902880 (1 of 10) *ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy*. <https://doi.org/10.1002/adv.201902880>
- Minassian, A. M., Satti, I., Poulton, I. D., Meyer, J., Hill, A. V. S., & Mcshane, H. (2011). A Human Challenge Model for Mycobacterium tuberculosis Using Mycobacterium bovis Bacille Calmette-Guérin. <https://doi.org/10.1093/infdis/jis012>
- Montaldo, C., Messina, F., Abbate, I., Antonioli, M., Bordoni, V., Aiello, A., Ciccocanti, F., Colavita, F., Farroni, C., Najafi Fard, S., Giombini, E., Goletti, D., Matusali, G., Rozera, G., Rueca, M., Sacchi, A., Piacentini, M., Agrati, C., Fimia, G. M., ... Ippolito, G. (2021). Multi-omics approach to COVID-19: a domain-based literature review. *Journal of Translational Medicine* 2021 19:1, 19(1), 1–18. <https://doi.org/10.1186/S12967-021-03168-8>
- Montoya, D. J., Andrade, P., Silva, B. J. A., Teles, R. M. B., Ma, F., Bryson, B., Sadanand, S., Noel, T., Lu, J., Sarno, E., Arnvig, K. B., Young, D., Lahiri, R., Williams, D. L., Fortune, S., Bloom, B. R., Pellegrini, M., & Modlin, R. L. (2019). Dual RNA-Seq of Human Leprosy Lesions Identifies Bacterial Determinants Linked to Host Immune Response. *Cell Reports*, 26(13), 3574–3585.e3. <https://doi.org/10.1016/j.celrep.2019.02.109>
- Mourik, B. C., de Steenwinkel, J. E. M., de Knecht, G. J., Huizinga, R., Verbon, A., Ottenhoff, T. H. M., van Soolingen, D., & Leenen, P. J. M. (2019). Mycobacterium tuberculosis clinical isolates of the

- Beijing and East-African Indian lineage induce fundamentally different host responses in mice compared to H37Rv. *Scientific Reports* 2019 9:1, 9(1), 1–11. <https://doi.org/10.1038/s41598-019-56300-6>
- Muñoz-Elías, E. J., & McKinney, J. D. (2005). Mycobacterium tuberculosis isocitrate lyases 1 and 2 are jointly required for in vivo growth and virulence. *Nature Medicine* 2005 11:6, 11(6), 638–644. <https://doi.org/10.1038/nm1252>
- Muñoz-Elías, E. J., Upton, A. M., Cherian, J., & McKinney, J. D. (2006). Role of the methylcitrate cycle in Mycobacterium tuberculosis metabolism, intracellular growth, and virulence. *Molecular Microbiology*, 60(5), 1109–1122. <https://doi.org/10.1111/J.1365-2958.2006.05155.X>
- Muttucumaru, D. G. N., Roberts, G., Hinds, J., Stabler, R. A., & Parish, T. (2004). Gene expression profile of Mycobacterium tuberculosis in a non-replicating state. *Tuberculosis*, 84(3–4), 239–246. <https://doi.org/10.1016/J.TUBE.2003.12.006>
- Nadolinskaia, N. I., Karpov, D. S., & Goncharenko, A. v. (2020). Vaccines Against Tuberculosis: Problems and Prospects (Review). *Applied Biochemistry and Microbiology*, 56(5), 497–504. <https://doi.org/10.1134/S0003683820050129/FIGURES/2>
- Napier, G., Campino, S., Merid, Y., Abebe, M., Woldeamanuel, Y., Aseffa, A., Hibberd, M. L., Phelan, J., & Clark, T. G. (2020). Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies. *Genome Medicine* 2020 12:1, 12(1), 1–10. <https://doi.org/10.1186/S13073-020-00817-3>
- Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., Diehn, M., & Alizadeh, A. A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* 2019 37:7, 37(7), 773–782. <https://doi.org/10.1038/s41587-019-0114-2>
- Newport, M. J., Huxley, C. M., Huston, S., Hawrylowicz, C. M., Oostra, B. A., Williamson, R., & Levin, M. (1996). A Mutation in the Interferon- γ –Receptor Gene and Susceptibility to Mycobacterial Infection. *New England Journal of Medicine*, 335(26), 518. <https://doi.org/10.1056/NEJM199612263352602>
- Newton, S. M., Mackie, S. L., Martineau, A. R., Wilkinson, K. A., Kampmann, B., Fisher, C., Dutta, S., Levin, M., Wilkinson, R. J., & Pasvol, G. (2008). Reduction of chemokine secretion in response to mycobacteria in infliximab-treated patients. *Clinical and Vaccine Immunology*, 15(3), 506–512. <https://doi.org/10.1128/CVI.00401-07>
- Newton, S., Martineau, A., & Kampmann, B. (2011). A Functional Whole Blood Assay to Measure Viability of Mycobacteria, using Reporter-Gene Tagged BCG or M.Tb (BCG lux/M.Tb lux). *JoVE (Journal of Visualized Experiments)*, 55, e3332. <https://doi.org/10.3791/3332>
- Neyrolles, O., Wolschendorf, F., Mitra, A., & Niederweis, M. (2015). Mycobacteria, metals, and the macrophage. *Immunological Reviews*, 264(1), 249–263. <https://doi.org/10.1111/IMR.12265>
- Ngabonziza, J. C. S., Loiseau, C., Marceau, M., Jouet, A., Menardo, F., Tzfadia, O., Antoine, R., Niyigena, E. B., Mulders, W., Fissette, K., Diels, M., Gaudin, C., Duthoy, S., Ssengooba, W., André, E., Kaswa, M. K., Habimana, Y. M., Brites, D., Affolabi, D., ... Supply, P. (2020). A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. *Nature Communications* 2020 11:1, 11(1), 1–11. <https://doi.org/10.1038/s41467-020-16626-6>

- Nicol, M. P., & Zar, H. J. (2020). Advances in the diagnosis of pulmonary tuberculosis in children. In *Paediatric Respiratory Reviews* (Vol. 36, pp. 52–56). W.B. Saunders Ltd.
<https://doi.org/10.1016/j.prrv.2020.05.003>
- Oh, T. S., Kang, H. Y., Nam, Y. S., Kim, Y. J., You, E. K., Lee, M. Y., Cho, S. Y., & Lee, H. J. (2016). An Effective Method of RNA Extraction from Mycobacterium tuberculosis. *Annals of Clinical Microbiology*, 19(1), 20–23. <https://doi.org/10.5145/ACM.2016.19.1.20>
- Olson, G. S., Murray, T. A., Jahn, A. N., Mai, D., Diercks, A. H., Gold, E. S., & Aderem, A. (2021). Type I interferon decreases macrophage energy metabolism during mycobacterial infection. *Cell Reports*, 35(9), 109195. <https://doi.org/10.1016/J.CELREP.2021.109195>
- Omotayo Togun, T., Maclean, E., Kampmann, B., & Pai, M. (2018). *Biomarkers for diagnosis of childhood tuberculosis: A systematic review*. <https://doi.org/10.1371/journal.pone.0204029>
- Oppong, Y. E. A., Phelan, J., Perdigão, J., Machado, D., Miranda, A., Portugal, I., Viveiros, M., Clark, T. G., & Hibberd, M. L. (2019). Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. *BMC Genomics*, 20(1), 252. <https://doi.org/10.1186/s12864-019-5615-3>
- Ordway, D., Henao-Tamayo, M., Harton, M., Palanisamy, G., Troudt, J., Shanley, C., Basaraba, R. J., & Orme, I. M. (2007). The Hypervirulent Mycobacterium tuberculosis Strain HN878 Induces a Potent TH1 Response followed by Rapid Down-Regulation. *The Journal of Immunology*, 179(1), 522–531. <https://doi.org/10.4049/JIMMUNOL.179.1.522>
- Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, 28(3), 245. <https://doi.org/10.1038/NBT.1614>
- O'Toole, R. F., & Gautam, S. S. (2017). Limitations of the Mycobacterium tuberculosis reference genome H37Rv in the detection of virulence-related loci. *Genomics*, 109(5–6), 471–474. <https://doi.org/10.1016/J.YGENO.2017.07.004>
- Oxford Genomics Centre. (2022). *Guides - Oxford Genomics Centre Oxford Genomics Centre*. <https://www.well.ox.ac.uk/ogc/guides/>
- Øyås, O., Borrell, S., Trauner, A., Zimmermann, M., Feldmann, J., Liphardt, T., Gagneux, S., Stelling, J., Sauer, U., & Zampieri, M. (2020). Model-based integration of genomics and metabolomics reveals SNP functionality in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 117(15), 8494–8502. <https://doi.org/10.1073/pnas.1915551117>
- Pai, M., Nicol, M. P., & Boehme, C. C. (2016). Tuberculosis Diagnostics: State of the Art and Future Directions. *Microbiology Spectrum*, 4(5). <https://doi.org/10.1128/MICROBIOLSPEC.TBTB2-0019-2016>
- Papavinasasundaram, K. G., Chan, B., Chung, J. H., Colston, M. J., Davis, E. O., & Av-Gay, Y. (2005). Deletion of the Mycobacterium tuberculosis pknH gene confers a higher bacillary load during the chronic phase of infection in BALB/c mice. *Journal of Bacteriology*, 187(16), 5751–5760. <https://doi.org/10.1128/JB.187.16.5751-5760.2005/ASSET/FAD00EC7-70DD-4ACE-84C7-607B2B757D07/ASSETS/GRAPHIC/ZJB0160549880005.JPEG>
- Park, J. H., Shim, D., Kim, K. E. S., Lee, W., & Shin, S. J. (2021). Understanding Metabolic Regulation Between Host and Pathogens: New Opportunities for the Development of Improved

- Therapeutic Strategies Against Mycobacterium tuberculosis Infection. *Frontiers in Cellular and Infection Microbiology*, 11, 212. <https://doi.org/10.3389/FCIMB.2021.635335/BIBTEX>
- Passemar, C., Arbués, A., Malaga, W., Mercier, I., Moreau, F., Lepourry, L., Neyrolles, O., Guilhot, C., & Astarie-Dequeker, C. (2014). Multiple deletions in the polyketide synthase gene repertoire of mycobacterium tuberculosis reveal functional overlap of cell envelope lipids in host-pathogen interactions. *Cellular Microbiology*, 16(2), 195–213. <https://doi.org/10.1111/CMI.12214/SUPPINFO>
- Phelan, J. E., Coll, F., Bergval, I., Anthony, R. M., Warren, R., Sampson, S. L., Gey Van Pittius, N. C., Glynn, J. R., Crampin, A. C., Alves, A., Barbosa Bessa, T., Campino, S., Dheda, K., Grandjean, L., Hasan, R., Hasan, Z., Miranda, A., Moore, D., Panaiotov, S., ... Clark, T. G. (2016). *Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages*. <https://doi.org/10.1186/s12864-016-2467-y>
- Pinto, S. M., Verma, R., Advani, J., Chatterjee, O., Patil, A. H., Kapoor, S., Subbannayya, Y., Raja, R., Gandotra, S., & Prasad, T. S. K. (2018). Integrated multi-omic analysis of Mycobacterium tuberculosis H37Ra redefines virulence attributes. *Frontiers in Microbiology*, 9(JUN), 1314. <https://doi.org/10.3389/FMICB.2018.01314/BIBTEX>
- Pisu, D., Huang, L., Grenier, J. K., & Russell, D. G. (2020). Dual RNA-Seq of Mtb-Infected Macrophages In Vivo Reveals Ontologically Distinct Host-Pathogen Interactions. *Cell Reports*, 30(2), 335–350.e4. <https://doi.org/10.1016/J.CELREP.2019.12.033>
- Portal-Celhay, C., Tufariello, J. M., Srivastava, S., Zahra, A., Klevorn, T., Grace, P. S., Mehra, A., Park, H. S., Ernst, J. D., Jacobs, W. R., & Philips, J. A. (2016). Mycobacterium tuberculosis EsxH inhibits ESCRT-dependent CD4+ T-cell activation. *Nature Microbiology* 2016 2:2, 2(2), 1–9. <https://doi.org/10.1038/nmicrobiol.2016.232>
- Portevin, D., Gagneux, S., Comas, I., & Young, D. (2011). Human Macrophage Responses to Clinical Isolates from the Mycobacterium tuberculosis Complex Discriminate between Ancient and Modern Lineages. *PLOS Pathogens*, 7(3), e1001307. <https://doi.org/10.1371/JOURNAL.PPAT.1001307>
- Pu, W., Zhao, C., Wazir, J., Su, Z., Niu, M., Song, S., Wei, L., Li, L., Zhang, X., Shi, X., & Wang, H. (2021). Comparative transcriptomic analysis of THP-1-derived macrophages infected with Mycobacterium tuberculosis H37Rv, H37Ra and BCG. *Journal of Cellular and Molecular Medicine*, 25(22), 10504–10520. <https://doi.org/10.1111/JCMM.16980>
- Puckett, S., Trujillo, C., Wang, Z., Eoh, H., Ioerger, T. R., Krieger, I., Sacchettini, J., Schnappinger, D., Rhee, K. Y., & Ehrt, S. (2017). Glyoxylate detoxification is an essential function of malate synthase required for carbon assimilation in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, 114(11), E2225–E2232. <https://doi.org/10.1073/PNAS.1617655114>
- Qiu, J., Ma, Y., Owusu, L., Jiang, T., & Xin, Y. (2014). Functional analysis of serine acetyltransferase from Mycobacterium smegmatis. *Journal of Basic Microbiology*, 54(7), 670–677. <https://doi.org/10.1002/JOBM.201300858>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. <https://doi.org/10.1093/bioinformatics/btq033>

- R Core Team. (2021). *R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.*
- Rachman, H., Strong, M., Ulrichs, T., Grode, L., Schuchhardt, J., Mollenkopf, H., Kosmiadi, G. A., Eisenberg, D., & Kaufmann, S. H. E. (2006). Unique transcriptome signature of Mycobacterium tuberculosis in pulmonary tuberculosis. *Infection and Immunity*, 74(2), 1233–1242. <https://doi.org/10.1128/IAI.74.2.1233-1242.2006>
- Rappoport, N., & Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20), 10546–10562. <https://doi.org/10.1093/NAR/GKY889>
- Rhee, K. Y., Carvalho, L. P. S. de, Bryk, R., Ehrt, S., Marrero, J., Park, S. W., Schnappinger, D., Venugopal, A., & Nathan, C. (2011). Central carbon metabolism in Mycobacterium tuberculosis: an unexpected frontier. *Trends in Microbiology*, 19(7), 307–314. <https://doi.org/10.1016/J.TIM.2011.03.008>
- Richard-Greenblatt, M., Bach, H., Adamson, J., Peña-Díaz, S., Li, W., C Steyn, A. J., & Av-Gay, Y. (2015). *Regulation of Ergothioneine Biosynthesis and Its Effect on Mycobacterium tuberculosis Growth and Infectivity* *. <https://doi.org/10.1074/jbc.M115.648642>
- Rienksma, R. A., Suarez-Díez, M., Mollenkopf, H. J., Dolganov, G. M., Dorhoi, A., Schoolnik, G. K., dos Santos, V. A. P. M., Kaufmann, S. H. E., Schaap, P. J., & Gengenbacher, M. (2015). Comprehensive insights into transcriptional adaptation of intracellular mycobacteria by microbe-enriched dual RNA sequencing. *BMC Genomics*, 16(1), 1–15. <https://doi.org/10.1186/s12864-014-1197-2>
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47–e47. <https://doi.org/10.1093/NAR/GKV007>
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1), 77. <https://doi.org/10.1186/1471-2105-12-77>
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *BIOINFORMATICS APPLICATIONS NOTE*, 26(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Rohart, F., Gautier, B., Singh, A., & Cao, K.-A. L. (2017). mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11), e1005752. <https://doi.org/10.1371/JOURNAL.PCBI.1005752>
- Rohde, K. H., Abramovitch, R. B., & Russell, D. G. (2007). Mycobacterium tuberculosis Invasion of Macrophages: Linking Bacterial Gene Expression to Environmental Cues. *Cell Host & Microbe*, 2(5), 352–364. <https://doi.org/10.1016/J.CHOM.2007.09.006>
- Rothchild, A. C., Olson, G. S., Nemeth, J., Amon, L. M., Mai, D., Gold, E. S., Diercks, A. H., & Aderem, A. (2019). Alveolar macrophages generate a noncanonical NRF2-driven transcriptional response to Mycobacterium tuberculosis in vivo. *Science Immunology*, 4(37). <https://doi.org/10.1126/SCIIMMUNOL.AAW6693>

- Roy, R. B., Sambou, B., Uhía, I., Roetynck, S., Robertson, B. D., & Kampmann, B. (2019). An auto-luminescent fluorescent BCG whole blood assay to enable evaluation of paediatric mycobacterial responses using minimal blood volumes. *Frontiers in Pediatrics*, 7(APR), 151. <https://doi.org/10.3389/FPED.2019.00151/BIBTEX>
- Roy, S., Guler, R., Parihar, S. P., Schmeier, S., Kaczowski, B., Nishimura, H., Shin, J. W., Negishi, Y., Ozturk, M., Hurdal, R., Kubosaki, A., Kimura, Y., Hoon, M. J. L. de, Hayashizaki, Y., Brombacher, F., & Suzuki, H. (2015). Batf2/Irf1 Induces Inflammatory Responses in Classically Activated Macrophages, Lipopolysaccharides, and Mycobacterial Infection. *The Journal of Immunology*, 194(12), 6035–6044. <https://doi.org/10.4049/JIMMUNOL.1402521>
- Roy, S., Schmeier, S., Kaczowski, B., Arner, E., Alam, T., Ozturk, M., Tamgue, O., Parihar, S. P., Kawaji, H., Itoh, M., Lassmann, T., Carninci, P., Hayashizaki, Y., Forrest, A. R. R., Guler, R., Bajic, V. B., Brombacher, F., & Suzuki, H. (2018). Transcriptional landscape of Mycobacterium tuberculosis infection in macrophages. *Scientific Reports* 2018 8:1, 8(1), 1–13. <https://doi.org/10.1038/s41598-018-24509-6>
- Ryndak, M. B., Wang, S., Smith, I., & Marcela Rodriguez, G. (2010). The Mycobacterium tuberculosis High-Affinity Iron Importer, IrtA, Contains an FAD-Binding Domain †. *JOURNAL OF BACTERIOLOGY*, 192(3), 861–869. <https://doi.org/10.1128/JB.00223-09>
- Sachdeva, G., Kaur, G., Bhutani, L. K., & Bamezai, R. (1997). Genetic variations at the T cell receptor γ locus in circulating peripheral blood mononuclear cells of clinically categorised leprosy patients. *Human Genetics* 1997 100:1, 100(1), 30–34. <https://doi.org/10.1007/S004390050461>
- Saini, V., Cumming, B. M., Guidry, L., Lamprecht, D. A., Adamson, J. H., Reddy, V. P., Chinta, K. C., Mazorodze, J. H., Glasgow, J. N., Richard-Greenblatt, M., Gomez-Velasco, A., Bach, H., Av-Gay, Y., Eoh, H., Rhee, K., & Steyn, A. J. C. (2016). Ergothioneine Maintains Redox and Bioenergetic Homeostasis Essential for Drug Susceptibility and Virulence of Mycobacterium tuberculosis. *Cell Reports*, 14(3), 572–585. <https://doi.org/10.1016/j.celrep.2015.12.056>
- Sathyanarayanan, A., Gupta, R., Thompson, E. W., Nyholt, D. R., Bauer, D. C., & Nagaraj, S. H. (2020). A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Briefings in Bioinformatics*, 21(6), 1920–1936. <https://doi.org/10.1093/BIB/BBZ121>
- Schnappinger, D., Ehrt, S., Voskuil, M. I., Liu, Y., Mangan, J. A., Monahan, I. M., Dolganov, G., Efron, B., Butcher, P. D., Nathan, C., & Schoolnik, G. K. (2003). Transcriptional Adaptation of Mycobacterium tuberculosis within Macrophages: Insights into the Phagosomal Environment. *The Journal of Experimental Medicine J. Exp. Med.* 198(5), 693–704. <https://doi.org/10.1084/jem.20030846>
- Schnappinger, D., Schoolnik, G. K., & Ehrt, S. (2006). Expression profiling of host pathogen interactions: how Mycobacterium tuberculosis and the macrophage adapt to one another. *Microbes and Infection*, 8(4), 1132–1140. <https://doi.org/10.1016/J.MICINF.2005.10.027>
- Sergi Sayols. (2020). *rrvgo: a Bioconductor package to reduce and visualize Gene Ontology terms*.
- Sharma, S., Ryndak, M. B., Aggarwal, A. N., Yadav, R., Sethi, S., Masih, S., Laal, S., & Verma, I. (2017). Transcriptome analysis of mycobacteria in sputum samples of pulmonary tuberculosis patients. *PLOS ONE*, 12(3), e0173508. <https://doi.org/10.1371/JOURNAL.PONE.0173508>

- Sia, J. K., & Rengarajan, J. (2019). Immunology of Mycobacterium tuberculosis Infections . *Microbiology Spectrum*, 7(4). <https://doi.org/10.1128/MICROBIOLSPEC.GPP3-0022-2018>
- Silva, D., Ponte, C. G. G., Hacker, M. A., & Antas, P. R. Z. (2013). A whole blood assay as a simple, broad assessment of cytokines and chemokines to evaluate human immune responses to Mycobacterium tuberculosis antigens. *Acta Tropica*, 127(2), 75–81. <https://doi.org/10.1016/J.ACTATROPICA.2013.04.002>
- Silva De Araujo, L., Ribeiro-Alves, M., Leal-Calvo, T., Leung, J., Durán, V., Samir, M., Talbot, S., Tallam, A., Carvalho De Queiroz Mello, F., Geffers, R., Helena, M., Saad, F., & Pessler, F. (2019). *Reprogramming of Small Noncoding RNA Populations in Peripheral Blood Reveals Host Biomarkers for Latent and Active Mycobacterium tuberculosis Infection*. <https://doi.org/10.1128/mBio.01037-19>
- Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, 40(Web Server issue), W452. <https://doi.org/10.1093/NAR/GKS539>
- Singer, S. N., Ndumego, O. C., Kim, R. S., Ndung'u, T., Anastos, K., French, A., Churchyard, G., Paramithiothis, E., Kasproicz, V. O., & Achkar, J. M. (2022). Plasma host protein biomarkers correlating with increasing Mycobacterium tuberculosis infection activity prior to tuberculosis diagnosis in people living with HIV. *EBioMedicine*, 75. <https://doi.org/10.1016/J.EBIOM.2021.103787/ATTACHMENT/15C1E6BE-D0F2-478A-B74E-0444F4142EF5/MMC3.XLSX>
- Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., & Lê Cao, K.-A. (2019). *DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays*. <https://doi.org/10.1093/bioinformatics/bty1054>
- Singhania, A., Wilkinson, R. J., Rodrigue, M., Haldar, P., & O'Garra, A. (2018). The value of transcriptomics in advancing knowledge of the immune response and diagnosis in tuberculosis. *Nature Immunology* 2018 19:11, 19(11), 1159–1168. <https://doi.org/10.1038/s41590-018-0225-9>
- Snewin, V. A., Gares, M. P., Ó Gaora, P., Hasan, Z., Brown, I. N., & Young, D. B. (1999). Assessment of immunity to mycobacterial infection with luciferase reporter constructs. *Infection and Immunity*, 67(9), 4586–4593. <https://doi.org/10.1128/IAI.67.9.4586-4593.1999>
- Sobota, R. S., Stein, C. M., Kodaman, N., Scheinfeldt, L. B., Maro, I., Wieland-Alter, W., Igo, R. P., Magohe, A., Malone, L. L., Chervenak, K., Hall, N. B., Modongo, C., Zetola, N., Matee, M., Joloba, M., Froment, A., Nyambo, T. B., Moore, J. H., Scott, W. K., ... Williams, S. M. (2016). A Locus at 5q33.3 Confers Resistance to Tuberculosis in Highly Susceptible Individuals. *The American Journal of Human Genetics*, 98(3), 514–524. <https://doi.org/10.1016/J.AJHG.2016.01.015>
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), 15545–15550. <https://doi.org/10.1073/PNAS.0506580102>

- Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, 14. <https://doi.org/10.1177/1177932219899051>
- Suliman, S., Thompson, E. G., Sutherland, J., Weiner, J., Ota, M. O. C., Shankar, S., Penn-Nicholson, A., Thiel, B., Erasmus, M., Maertzdorf, J., Duffy, F. J., Hill, P. C., Jane Hughes, E., Stanley, K., Downing, K., Fisher, M. L., Valvo, J., Parida, S. K., van der Spuy, G., ... Verver, S. (2018). Four-gene pan-African blood signature predicts progression to tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, 197(9), 1198–1208. https://doi.org/10.1164/RCCM.201711-2340OC/SUPPL_FILE/DISCLOSURES.PDF
- Sun, J. C., Lopez-Verges, S., Kim, C. C., DeRisi, J. L., & Lanier, L. L. (2011). NK Cells and Immune “Memory.” *The Journal of Immunology*, 186(4), 1891–1897. <https://doi.org/10.4049/JIMMUNOL.1003035>
- Sweeney, T. E., Braviak, L., Tato, C. M., & Khatiri, P. (2016). Genome-wide expression for diagnosis of pulmonary tuberculosis: A multicohort analysis. *The Lancet Respiratory Medicine*, 4(3), 213–224. [https://doi.org/10.1016/S2213-2600\(16\)00048-5](https://doi.org/10.1016/S2213-2600(16)00048-5)
- Tena, G. N., Young, D. B., Eley, B., Henderson, H., Nicol, M. P., Levin, M., & Kampmann, B. (2003). Failure to Control Growth of Mycobacteria in Blood from Children Infected with Human Immunodeficiency Virus and Its Relationship to T Cell Function. *The Journal of Infectious Diseases*, 187(10), 1544–1551. <https://doi.org/10.1086/374799>
- Tesmer, L. A., Lundy, S. K., Sarkar, S., & Fox, D. A. (2008). Th17 cells in human disease. *Immunological Reviews*, 223(1), 87. <https://doi.org/10.1111/J.1600-065X.2008.00628.X>
- Tini, G., Marchetti, L., Priami, C., & Scott-Boyer, M.-P. (2019). Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings in Bioinformatics*, 20(4), 1269–1279. <https://doi.org/10.1093/BIB/BBX167>
- Tounta, V., Liu, Y., Cheyne, A., & Larrouy-Maumus, G. (2021). Metabolomics in infectious diseases and drug discovery. *Molecular Omics*, 17(3), 376–393. <https://doi.org/10.1039/D1MO00017A>
- Tretina, K., Park, E.-S., Maminska, A., & Macmicking, J. D. (2019). *Interferon-induced guanylate-binding proteins: Guardians of host defense in health and disease*. <https://doi.org/10.1084/jem.20182031>
- Turner, C. T., Gupta, R. K., Tsaliki, E., Roe, J. K., Mondal, P., Nyawo, G. R., Palmer, Z., Miller, R. F., Reeve, B. W., Theron, G., & Noursadeghi, M. (2020). Blood transcriptional biomarkers for active pulmonary tuberculosis in a high-burden setting: a prospective, observational, diagnostic accuracy study. *The Lancet Respiratory Medicine*, 8(4), 407–419. [https://doi.org/10.1016/S2213-2600\(19\)30469-2](https://doi.org/10.1016/S2213-2600(19)30469-2)
- van Crevel, R., Avila-Pacheco, J., Thuong, N. T. T., Ganiem, A. R., Imran, D., Hamers, R. L., Alisjahbana, B., Dian, S., Estiasari, R., Tram, T. T. B., Vinh, D. N., Hai, H. T., Donovan, J., Ardiansyah, E., Koeken, V., Kumar, V., Netea, M. G., van Laarhoven, A., Clish, C., & Thwaites, G. (2021). Improving host-directed therapy for tuberculous meningitis by linking clinical and multi-omics data. *Tuberculosis*, 128, 102085. <https://doi.org/10.1016/J.TUBE.2021.102085>
- Verhagen, L. M., Zomer, A., Maes, M., Villalba, J. A., del Nogal, B., Eleveld, M., Aft Van Hijum, S., de Waard, J. H., & Hermans, P. W. (2013). *A predictive signature gene set for discriminating active*

- from latent tuberculosis in Warao Amerindian children. <https://doi.org/10.1186/1471-2164-14-74>
- version 8.0.0 for Windows, GraphPad Software, San Diego, California USA, www.graphpad.com. (n.d.).
- Veyrier, F., Pletzer, D., Turenne, C., & Behr, M. A. (2009). Phylogenetic detection of horizontal gene transfer during the step-wise genesis of *Mycobacterium tuberculosis*. *BMC Evolutionary Biology*, 9(1), 1–14. <https://doi.org/10.1186/1471-2148-9-196/FIGURES/5>
- Volpe, E., Cappelli, G., Grassi, M., Martino, A., Serafino, A., Colizzi, V., Sanarico, N., & Mariani, F. (2006). Gene expression profiling of human macrophages at late time of infection with *Mycobacterium tuberculosis*. *Immunology*, 118(4), 449–460. <https://doi.org/10.1111/J.1365-2567.2006.02378.X>
- von Both, U., Berk, M., Agapow, P. M., Wright, J. D., Git, A., Hamilton, M. S., Goldgof, G., Siddiqui, N., Bellos, E., Wright, V. J., Coin, L. J., Newton, S. M., & Levin, M. (2018). *Mycobacterium tuberculosis* Exploits a Molecular off Switch of the Immune System for Intracellular Survival. *Scientific Reports*, 8(1), 1–17. <https://doi.org/10.1038/s41598-017-18528-y>
- Walzl, G., McNerney, R., du Plessis, N., Bates, M., McHugh, T. D., Chegou, N. N., & Zumla, A. (2018). Tuberculosis: advances and challenges in development of new diagnostics and biomarkers. *The Lancet Infectious Diseases*, 18(7), e199–e210. [https://doi.org/10.1016/S1473-3099\(18\)30111-7](https://doi.org/10.1016/S1473-3099(18)30111-7)
- Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* 11:3, 11(3), 333–337. <https://doi.org/10.1038/nmeth.2810>
- Wang, S., He, L., Wu, J., Zhou, Z., Gao, Y., Chen, J., Shao, L., Zhang, Y., & Zhang, W. (2019). Transcriptional Profiling of Human Peripheral Blood Mononuclear Cells Identifies Diagnostic Biomarkers That Distinguish Active and Latent Tuberculosis. *Frontiers in Immunology*, 10, 2948. <https://doi.org/10.3389/FIMMU.2019.02948/BIBTEX>
- Wei, W., Yan, H., Zhao, J., Li, H., Li, Z., Guo, H., Wang, X., Zhou, Y., Zhang, X., Zeng, J., Chen, T., & Zhou, L. (2019). Multi-omics comparisons of p-aminosalicylic acid (PAS) resistance in folC mutated and un-mutated *Mycobacterium tuberculosis* strains. <https://doi.org/10.1080/22221751.2019.1568179>, 8(1), 248–261. <https://doi.org/10.1080/22221751.2019.1568179>
- Westermann, A. J., Barquist, L., & Vogel, J. (2017). Resolving host–pathogen interactions by dual RNA-seq. *PLoS Pathogens*, 13(2), 1–19. <https://doi.org/10.1371/journal.ppat.1006033>
- Westermann, A. J., Förstner, K. U., Amman, F., Barquist, L., Chao, Y., Schulte, L. N., Müller, L., Reinhardt, R., Stadler, P. F., & Vogel, J. (2016). Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions. *Nature* 2016 529:7587, 529(7587), 496–501. <https://doi.org/10.1038/nature16547>
- Westermann, A. J., Gorski, S. A., & Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology*, 10(9), 618–630. <https://doi.org/10.1038/nrmicro2852>
- WHO. (2014). *High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting*. www.who.int

- WHO. (2014). *The end TB Strategy. Global strategy and targets for tuberculosis prevention, care and control after 2015*.
- WHO. (2020). WHO GLOBAL TUBERCULOSIS REPORT 2020. *GLOBAL TUBERCULOSIS REPORT*. <http://apps.who.int/bookorders>.
- WHO. (2021). *WHO consolidated guidelines on tuberculosis. Module 3: Diagnosis - Rapid diagnostics for tuberculosis detection 2021 update*. <https://www.who.int/publications/i/item/9789240029415>
- Wilkinson, M. E., Charenton, C., & Nagai, K. (2020). RNA Splicing by the Spliceosome. <https://doi.org/10.1146/Annurev-Biochem-091719-064225>, 89, 359–388. <https://doi.org/10.1146/ANNUREV-BIOCHEM-091719-064225>
- Witten, D. M., Tibshirani, R., Hastie, T., & Witten, D. M. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3), 515–534. <https://doi.org/10.1093/biostatistics/kxp008>
- Wu, R., Chen, F., Wang, N., Tang, D., & Kang, R. (2020). ACOD1 in immunometabolism and disease. *Cellular & Molecular Immunology* 2020 17:8, 17(8), 822–833. <https://doi.org/10.1038/s41423-020-0489-5>
- Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., ... Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, 46(D1), D754–D761. <https://doi.org/10.1093/nar/gkx1098>
- Zhang, H., Li, D., Zhao, L., Fleming, J., Lin, N., Wang, T., Liu, Z., Li, C., Galwey, N., Deng, J., Zhou, Y., Zhu, Y., Gao, Y., Wang, T., Wang, S., Huang, Y., Wang, M., Zhong, Q., Zhou, L., ... Bi, L. (2013). Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. *Nature Genetics* 2013 45:10, 45(10), 1255–1260. <https://doi.org/10.1038/ng.2735>
- Zhang, R., Ou, H. Y., & Zhang, C. T. (2004). DEG: a database of essential genes. *Nucleic Acids Research*, 32(suppl_1), D271–D272. <https://doi.org/10.1093/NAR/GKH024>
- Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, 2(3). <https://doi.org/10.1093/nargab/lqaa078>
- Zhao, J., Wei, W., Yan, H., Zhou, Y., Li, Z., Chen, Y., Zhang, C., Zeng, J., Chen, T., & Zhou, L. (2019). Assessing capreomycin resistance on tlyA deficient and point mutation (G695A) Mycobacterium tuberculosis strains using multi-omics analysis. *International Journal of Medical Microbiology*, 309(7), 151323. <https://doi.org/10.1016/J.IJMM.2019.06.003>
- Zhou, M., Varol, A., & Efferth, T. (2021). Multi-omics approaches to improve malaria therapy. *Pharmacological Research*, 167, 105570. <https://doi.org/10.1016/J.PHRS.2021.105570>
- Zhou, M., Yu, G., Yang, X., Zhu, C., Zhang, Z., & Zhan, X. (2016). Circulating microRNAs as biomarkers for the early diagnosis of childhood tuberculosis infection. *Molecular Medicine Reports*, 13(6), 4620–4626. <https://doi.org/10.3892/mmr.2016.5097>
- Zimmermann, M., Kogadeeva, M., Gengenbacher, M., McEwen, G., Mollenkopf, H.-J., Zamboni, N., Kaufmann, S. H. E., & Sauer, U. (2017). Integration of Metabolomics and Transcriptomics

Reveals a Complex Diet of *Mycobacterium tuberculosis* during Early Macrophage Infection.
MSystems, 2(4). <https://doi.org/10.1128/MSYSTEMS.00057-17>