University of Cape Town

Department of Statistical Sciences

In collaboration with the Department of Pathology and the South African Tuberculosis Vaccine Initiative

---

MULTIVARIATE ANALYSIS OF THE IMMUNE RESPONSE UPON RECENT ACQUISITION OF *MYCOBACTERIUM TUBERCULOSIS* INFECTION

Author:
Tessa Lloyd

Supervisors:
Associate Prof. Francesca Little
Associate Prof. Elisa Nemes

Co-supervisor:
Dr. Pia Steigler

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Abstract

Tuberculosis (TB), caused by the pathogen *Mycobacterium tuberculosis* (*M.tb*), is the leading cause of mortality due to an infectious agent worldwide. Based on data from an adolescent cohort study carried out from May 2005 to February 2009, we studied and compared the immune responses of individuals from four cohorts that were defined based on their longitudinal QFT results: the recent QFT converters, the QFT reverters, the persistent QFT positives and negatives. Analysis was based on the integration of different arms of the immune response, including adaptive and "innaptive" responses, measured on the cohorts. COMPASS was used to filter the adaptive dataset and identify bioligically meaningful subsets, while, for the innaptive dataset, we came up with a novel filtering method. Once the datasets were integrated, they were standardized using variance stabilizing (vast) standardization and missing values were imputed using a multiple factor analysis (MFA)-based approach. We first set out to define a set of immune features that changed during recent *M.tb* infection. This was achieved by employing the kmlShape clustering algorithm to the recent QFT converters. We identified 55 cell subsets to either increase or decrease post-infection. When we assessed how the associations between these changed pre- and post-infection using correlation networks, we found no notable differences. By comparing the recent QFT converters and the persistent QFT positives, a blood-based biomarker to distinguish between recent and established infection, namely ESAT6/CFP10-specific expression of HLA-DR on total Th1 cells, was identified using elastic net (EN) models (average AUROC = 0.87). The discriminatory ability of this variable was confirmed using two tree-based models. Lastly, to assess whether the QFT reverters are a biologically distinct group of individuals, we compared them to the persistent QFT positive and QFT negative individuals using a Projection to Latent Space Discriminant Analysis (PLS-DA) model. The results indicated that reverters appeared more similar to QFT negative individuals rather than QFT positive. Hence, QFT reversion may be associated with clearance of *M.tb* infection. Immune signatures associated with recent infection could be used to refine end-points of clinical trials testing vaccine efficacy against acquisition of *M.tb* infection, while immune signatures associated with QFT reversion could be tested as correlates of protection from *M.tb* infection.

Acknowledgments

I thank my supervisors Francesca Little, Elisa Nemes and Pia Steigler for their support, guidance, time and effort that they spent on making this project a success. In particular I thank Francesca for her practical advice and increasing my statistical knowledge and sparking my interest in biostatistics, Elisa for all her time and expertise to ensure the quality of this project, and for all she has taught me about research, and Pia for providing the innaptive dataset and all her guidance and support throughout the project. I would also like to acknowledge Cheleka Mpande, who provided the adaptive dataset and all her advice throughout the project. I further acknowledge the efforts of the Adolescent Cohort Study (ACS) team. I thank the excellent team of researchers at the South African Tuberculosis Vaccine Initiative (SATVI) for their funding, support and their investment in me as a researcher. Lastly, I gratefully acknowledge the National Research Foundation (NRF) and the Statistical Association of South Africa (SASA) for funding my degree and for their financial support otherwise.

Plagiarism declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used an appropriate convention for citation and referencing (APA). Each significant contribution to, and quotation in, this project report from the work of other people has been attributed, and has been cited and referenced.
3. This proposal is my own work.
4. I have not allowed, and will not allow, anyone to copy our work with the intention of passing it off as his or her own work.

Name: **Signature** Signed by candidate

Tessa Lloyd _____

Table of Contents

List of Algorithms

List of Figures

List of Tables

Chapter 1

Introduction and background of study

1.1 Introduction

This project aimed to define immune signatures associated with the acquisition and potential clearance of *Mycobacterium tuberculosis* (*M.tb*) infection. Tuberculosis (TB) is an airborne bacterial disease that is the leading cause of mortality due to an infectious agent worldwide [98]. It is estimated that about a quarter of the world's population is infected with *M.tb*, the pathogen that causes TB [66]. In the case of pulmonary disease, *M.tb* can be expelled while coughing and transmitted to exposed individuals (Figure 1.1). After exposure some individuals are able to eliminate *M.tb* through innate or adaptive memory immune responses. Other individuals progress rapidly to active disease, this is known as primary TB and tends to be more common in children. Most individuals, however, contain the infection after *M.tb* exposure and are classified as having latent TB infection (LTBI). LTBI is measured by the immune sensitization to *M.tb*-specific antigens such as present in the QuantiFERON TB (QFT) test (see below). These individuals are asymptomatic, show no clinical signs of TB and are generally considered to not be contagious. Within the first two years after exposure, latently infected individuals are at the highest risk of progressing to active or post-primary TB. Active TB patients will exhibit TB-specific symptoms such as coughing or fever and *M.tb* infection is confirmed by microbiological tests performed in sputum (for pulmonary TB) or other biological fluids (for extra-pulmonary TB). There are large variations among the severity of active TB disease, which are best explained by the host's response to the pathogen. A process known as reactivation may occur much later and generally is the result of a compromised immune system, such as in the case of HIV co-infection, diabetes or from old age. TB can disseminate to any part of the body, but in about 70% of cases, TB mainly manifests as a pulmonary disease [97]. It is further suggested that only 5-10% of latently infected individuals will progress to a state of active TB disease in their lifetime [12].

Acquisition of *M.tb* infection is generally asymptomatic and tends to remain undiagnosed unless serial diagnostic tests are performed in exposed populations. As a result very little is known about the immune response induced during primary *M.tb* infection in humans. This project aims to focus on the immune responses involved in the acquisition and the possible event of clearance of *M.tb* infection, detected by QFT conversion and reversion respectively (Figure 1.1). LTBI is classified by the absence of clinical evidences of active disease in the presence of an immune response against antigens specific to *M.tb*. Currently no blood-based tests to distinguish LTBI and active TB are available and disease diagnosis is drawn from combining microbiological tests with symptoms. The drawback of existing tests to identify asymptomatic *M.tb* infected individuals is

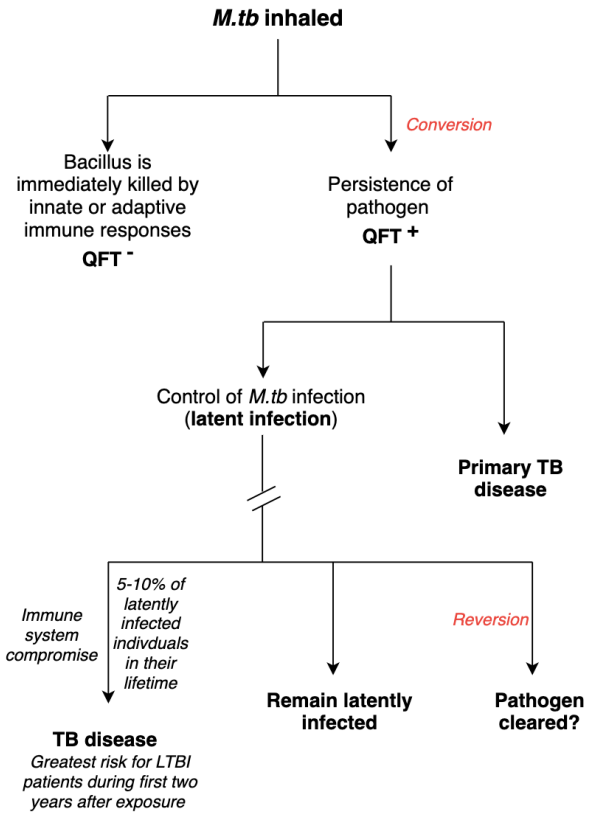

Figure 1.1: Outcomes of *M.tb* infection.

the use of indirect measures of immune responses to *M.tb* antigens, since *M.tb* can not be readily detected in sputum of healthy individuals.

Diagnostic tests for the detection of *M.tb* infection measure a memory T cell response to *M.tb*-specific antigens, which should be present only if the person is (or had been) infected. One such test is the Tuberculin Skin Test (TST), which is an intradermal injection of a purified protein derivative (PPD). If a person has been previously exposed to *M.tb*, cell-mediated immunity will cause a visible skin reaction. The reaction is measured in millimeters (mm) 48-72 hours after the injection is administered. A positive TST is confirmed when the reaction is greater than a pre-defined threshold value that will be different depending on the specific endemic setting. PPD contains antigens that are not specific to *M.tb* but also present in other mycobacteria. Infants in high burden TB settings are vaccinated at birth with Bacille Calmette-Guerin (BCG, which is an attenuated form of *Mycobacterium bovis*), which can trigger a false TST positive result even if babies have never been exposed to *M.tb*. This is a limitation of TST in high burden settings.

Although the TST is still commonly used, more specific tests have been developed; collectively termed as Interferon Gamma Release Assays (IGRA). These tests are *in vitro* blood tests measuring *M.tb*-specific cell mediated immune responses. Two commercial IGRA tests are currently available, the T-spot TB assay and the QuantiFERON-TB (QFT) test. For the scope of this project we will only focus on the QFT test. QFT measures the level of interferon-γ (IFN-γ), a cytokine which is released by *M.tb* pre-sensitised T cells upon stimulation by early secretory antigen target-6

(ESAT-6) and culture filtrate protein-10 (CFP-10), collectively termed E6C10. QFT results are reported in international units (IU) per mL and the cut-off for a positive test is 0.35 IU/mL. This test has a higher specificity than the TST because E6C10 are antigens only present in *M.tb*. Hence a person will not react if they have been previously vaccinated by BCG or exposed to naturally occurring environmental mycobacteria [99].

Both the TST and IGRA tests can indicate *M.tb* infection, but neither of them is able to distinguish

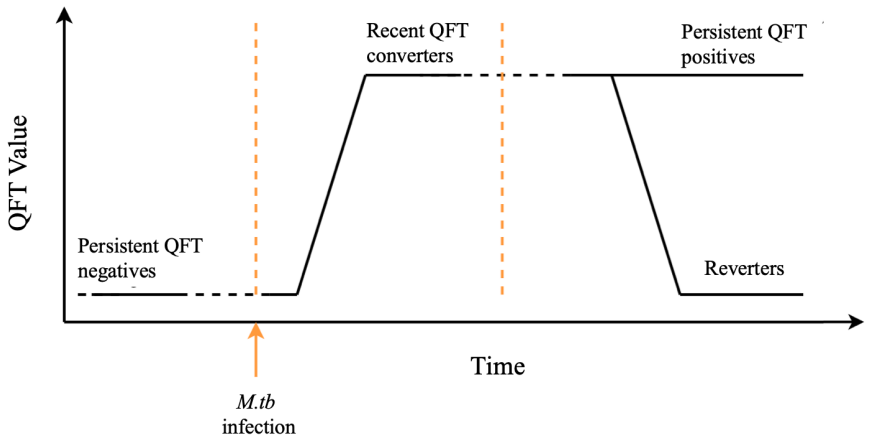

Figure 1.2: Definition of the cohorts based on their QFT results.

recent infection from established infection, unless serial testing is performed showing conversion from a negative to a positive test. Recent acquisition of infection, in the absence of other factors, is the greatest risk factor for the progression to TB [12], therefore being able to distinguish between the two states would allow us to target recently infected individuals for preventative TB treatment.

The phenomenon of reversion of immune-diagnostic tests has become a point of interest in recent studies. Reversion is described as a positive QFT result that reverts to a negative QFT test in a longitudinal study (Figure 1.2) [6]. Whether reversion has any clinical significance, or if it is simply due to technical or biological variability, is still unknown. In studies done on guinea pigs, it was found that TST reversion was associated with sterilizing immunity [104][88][38]. In humans, TST reverters had evidence of self-healed TB prior to availability of antibiotics or had a lower risk of TB compared to persistent TST positives [62]. QFT reversion could therefore be associated with antigen clearance.

In order to assess QFT conversion and reversion dynamics in a TB endemic setting, an Adolescent Cohort Study (ACS) was carried out from May 2005 through February 2009 [85]. Students between the ages of 12 and 18 years were recruited from 11 local schools in the town of Worcester in the Western Cape and QFT tests were performed every 6 months during a two year follow-up period. Four cohorts were defined based on their QFT results (Figure 1.2): persistent negative individuals (four negative QFT tests), recent QFT converters (two negative QFT tests followed by two positive QFT tests), persistent QFT positive (four positive QFT tests) and reverters (two positive QFT tests followed by two negative QFT tests).

Analysis was performed on different components of the immune system in order to define the immunological determinants of recent QFT conversion and reversion. Specifically, we built on existing datasets that measured adaptive immunity, donor unrestricted T (DURT) cell and innate

immunity. Adaptive immunity consists of memory-driven cellular responses, including conventional T cells, while innate immunity consists of non-specific defence mechanisms including B cells (in their antigen presenting capacity), monocytes and natural killer (NK) cells. DURT cell immunity shares features from both adaptive and innate immunity and bridges between both arms. The DURT cells studied included gamma delta ($\gamma\delta$) T cells, NKT-like cells and mucosal associated invariant T (MAIT) cells. In this project I will refer to the integrated datasets as the adaptive and "innaptive" datasets, where the innaptive data contains the innate and DURT cell responses.

In the context of TB, classical helper T cells that recognize peptide antigens measured by QFT tests are necessary but not sufficient for the control of *M.tb* [5]. Other immune cells such as DURT cells may complement classical T cell-mediated immunity to TB as they recognize a variety of mycobacterial non-peptide antigens and host-derived stress signals, which classical T cells do not recognize. Moreover, innaptive cells can respond more rapidly to stimuli than conventional T cells and may be required for early containment of *M.tb* infection. Therefore, being able to integrate the data from all these cell types will increase our understanding of the immune response as a whole and could provide insights into synergies and relationships between different arms of the immune response that contribute to the different outcomes in QFT conversion and reversion.

Immune signatures associated with recent QFT conversion could be used to understand the immune response behind the establishment of *M.tb* infection and to refine end-points of clinical trials testing vaccine efficacy against acquisition of *M.tb* infection. Ultimately this will help to produce better preventative interventions. On the other hand, immune signatures associated with QFT reversion could be tested as correlates of protection from *M.tb* infection, for which we have samples from a vaccine efficacy trial [92] available.

The proposed study will therefore enhance our understanding of the dynamics of host-pathogen interactions, which are inherently complex but crucial to the design of better immunodiagnostics and interventions, such as vaccines and host directed therapies.

1.2 Study population

About half of the participants (n=3,236) from the ACS (n = 6,363) were tested for *M.tb* sensitization by QFT every six months, at month 0, 6, 12 and 18, and had cyro-preserved peripheral blood mononuclear cells (PBMC) available at the same time points. QFT qualitative results were defined by a cut-off of 0.35 IU/mL of IFN-γ produced in response to *M.tb* specific antigens after subtraction of background responses in unstimulated samples (a negative control). Since quantitative values around the assay cut-off (0.2-0.7 IU/mL) may lead to misclassification of qualitative results, likely due to technical variability, a more stringent QFT cut-off was applied to define the study groups [93]. The study population was separated into four groups based on their QFT results (Figure 1.3):

- Persistent QFT negatives (n = 30): QFT negative ($<$ 0.2 IU/mL) at 4 consecutive visits (negative control group).
- Recent QFT converters (n = 29): QFT negative (of which at least one $<$ 0.2 IU/mL) at 2 consecutive visits, followed by 2 QFT positive (of which at least one $>$ 0.7 IU/mL) at 2 consecutive visits.
- Persistent QFT positive (n = 30): QFT positive ($\geq$ 0.35 IU/mL) at 4 consecutive visits (positive control group).
- QFT reverters (n = 30): QFT positive at 2 consecutive visits (at least one $\geq$ 0.7 IU/mL), followed by QFT negative at 2 consecutive visits (at least one $<$ 0.2 IU/mL).

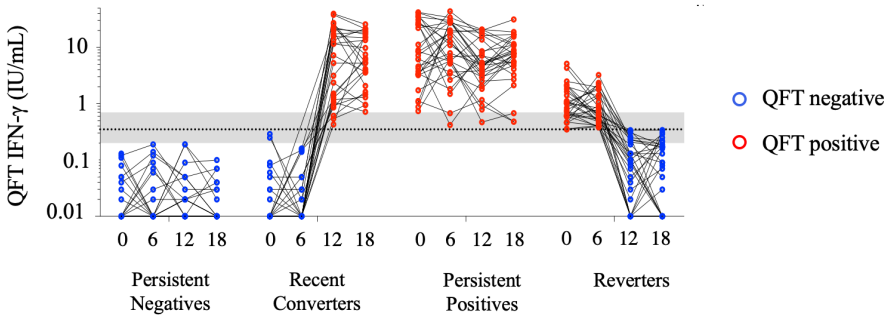

Figure 1.3: QFT values in the study groups.

1.3 Aims and objectives

The aim of the project is to define immune signatures associated with acquisition and potential clearance of *M.tb* infection.

We applied sophisticated statistical methods to analyse individual sets of data as well as analyzing a combined dataset measured on the same N individuals to determine whether the combined analysis will outperform, or add to, the individual analyses. Such an integration of datasets helped us to gain a better understanding of the interplay between the different levels of data. In addition, we hypothesized that immune interactions between different immune cells will occur upon *M.tb* acquisition or that the immune interactions are present but may change according to phenotype after infection.

- **AIM 1**: To define features of the immune response that change upon QFT conversion.
  We hypothesize that upon QFT conversion frequencies of IFN-γ+ TNF+ IL-2+ M.tb-specific CD4+ T cells are higher compared to pre-conversion time-points.
  Upon stimulation with *M.tb*, monocytes release IL-12, which promotes Th1 cell differentiation, and hence we expect the polyfunctional CD4+ count to increase upon conversion.
  In addition to testing a small set of pre-defined hypotheses, integration of different datasets will contribute to the understanding of the biology underlying QFT conversion and how the overall immune response changes upon *M.tb* infection (Figure 1.4).

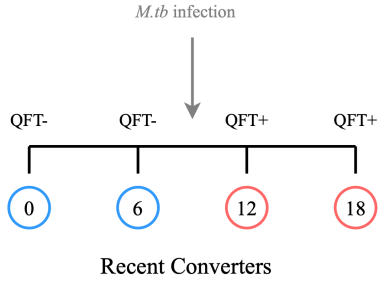

Figure 1.4: The recent converter cohort used in Aim 1.

- **AIM 2**: To investigate which features of the immune response differ between recent and established *M.tb* infection

We hypothesize that proportions of TNF-only producing CD4+T cells with a T_{EFF} phenotype and HLA-DR expression on M.tb-specific CD4+ T cells are higher in individuals with recent M.tb infection compared to those with established infection.
Halliday et al. [59] found TNF-only producing CD4+T cells were a marker of recent infection. In addition, there is significant evidence that *M.tb*-specific T cell activation is correlated with bacterial load, and therefore we expect proportions of HLA-DR+IFN-γ+CD4+ T cells to be higher in recent compared to established infection.
We will compare immune responses in individuals with recent (QFT conversion from negative to positive within 6 months) and established *M.tb* infection (QFT positive for more than 1 year) (Figure 1.5) to identify the best discriminating biomarkers across different datasets.

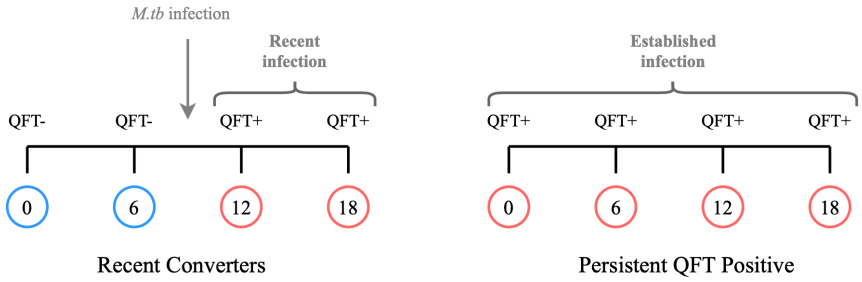

Figure 1.5: Definition of recent and established *M.tb* infection for Aim 2.

- **AIM 3**: To identify which features of the immune response are associated with QFT reversion.
  Under the assumption that QFT reversion is associated with clearance of M.tb infection, we hypothesize that immune features induced by M.tb infection will have opposite trends upon QFT reversion.
  We investigated how immune responses change upon QFT reversion, and compare immune features to QFT converters (Aim 1), persistent QFT positives (Aim 2) and QFT non-converters (Figure 1.6).

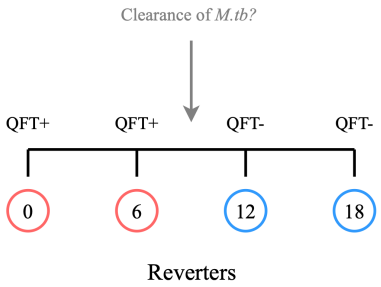

Figure 1.6: The reverters group explored in Aim 3.

The final objective of this project is to interrogate the derived integrated model for novel and testable biological hypotheses.

Chapter 2

Literature review

2.1 Immune response to *M.tb*

The immune response to *M.tb* is composite and not fully understood due to the complexity of human-*M.tb* interactions and the heterogeneity of the global epidemic. After entering the host, *M.tb* can either be eliminated by cells of the innate or adaptive immune system. Both arms of the immune response play an important role in determining the host response of the immune response to the pathogen [44]. Innate immunity refers to non-specific defense mechanisms against an invading pathogen that become functional soon after an antigen's appearance in the body. Adaptive immunity, on the other hand, refers to an antigen-specific immune response and consists of a collection of highly specialized cells that require days to mature and expand before they can effectively respond to a pathogen. Lastly, DURT cell immunity includes immune cells that share features from both adaptive and innate immunity and bridge both arms.

Heterogeneity of the host immune response and the bacterial metabolism within individuals makes it difficult to identify biomarkers that can be applied and tested on a population level. An ideal biomarker should be able to differentiate between exposed individuals who do or do not become infected, differentiate between LTBI and active disease states, return to normal levels post or during successful TB treatment and should reproduce any clinical outcomes. Expansion beyond the narrow focus on IFN-γ producing T cells is necessary to identify novel biomarkers, which could be immediately tested as correlates of protection (CoP) in ongoing clinical trials. Identifying such biomarkers will help design vaccines that are ideally able to prevent infection, prevent development to active TB disease or to prevent the recurrence of disease in individuals with clinical TB [5].

2.1.1 The innate immune response

The panel of innate cells that we will focus on for this project consists of monocytes, B cells and natural killer (NK) cells. Alveolar macrophages are the main target cells infected by *M.tb*, which can inhibit phagolysosome formation, and play a key role in the coordination of innate responses and initiation of adaptive T cell immunity. We will not study resident lung alveolar macrophages specifically in this project, but it is important to understand the critical role they play during the initiation of the immune response against *M.tb*.

Monocytes are types of leukocytes (white blood cells) that can differentiate to macrophages or dendritic cells (DCs). Macrophages and DCs are phagocytes, release cytokines and present antigens to adaptive immune cells. Antigen presentation is the process by which antigens are presented to T cells in the form of short peptide fragments that activate specific T cells. The principal antigen-presenting cells are DCs, monocytes and B cells, and they express major histocompatibility

complex (MHC) class I and MHC class II molecules on their cell surfaces. Classical CD4+ T helper cells recognize peptides presented on MHC class II, while CD8+ T cells typically recognize peptides presented on MHC class I molecules.

Monocytes play a key role in the primary stages of the innate immune response as critical defense mechanisms. *M.tb*-infected monocytes produce IL-12 , IL-10, TNF and IL-6, and are classified based on relative expression levels of CD14 and CD16 surface proteins. IL-12 is a pro-inflammatory cytokine that promotes the development of Th1 cells and induces the production of IFN-γ in T cells and NK cells [16]. IL-10 on the other hand is an anti-inflammatory cytokine which, if over-produced, can contribute to chronic infection, but is essential for reducing host immunopathology. IL-6 is also a pro-inflamatory cytokine, while TNF plays a critical role in the formation of the granuloma in immune competent individuals [11][28][100]. The granuloma is a collection of cells that enclose the bacilli. TB granulomas consist of a mass of infected macrophages, stimulated macrophages, and other myeloid cells that are surrounded by CD4+ and CD8+ T cells and B cells [111][102]. If the granuloma contains *M.tb* then the person is defined to be latently infected and could be a candidate for therapeutic treatment, but this is virtually impossible to measure in humans. Therefore, there is evidence that monocytes play an important role during *M.tb* infection and, together with other immune cells, may be able to clear *M.tb*.

NK cells are granular lymphocytes that are also known to play an important role against viral infections. NK cells and macrophages are the main contributing cells of the innate immune system and are involved in inhibiting the growth of *M.tb* in the absence of T cells [4]. NK cells can kill an invading pathogen either through direct or indirect mechanisms. NK cells can "recognize" pathogens or infected cells through the direct binding of antibodies specific for the microbial antigens expressed on the surface of infected cells. NK cells release granzymes and perforin, which form pores in the cell membrane of the pathogen or infected cell causing cell lysis or apoptosis. Otherwise, NK cells secrete IFN-γ and TNF, which activate macrophages to perform the killing. Apoptosis of monocytes has also been found to be induced by NK cells [145]. NK cells express CD16 and CD56 and respond to the cytokines IL-2 and IL-12. IL-2 increases the lytic activity of NK cells [18] while IL-12 promotes the secretion of IFN-γ by NK cells, and these IL-12-stimulated NK cells have been shown to inhibit the growth of *M.tb* [16]. Only recently have NK cells been considered important in the context of TB, as either a diagnostic marker or taking on a more protective role in both early and later stages of *M.tb* infection [43].

Currently, the role of B cells and humoral immunity is not fully understood in the context of TB. B cells share features of both adaptive (antibody production) and innate (antigen presentation) immunity and are present in large quantities in granuloma. They can either be found circulating the blood in the form of long-lived plasma cells, or as memory B cells. Memory B cells mainly reside in the bone marrow but they may also be present at the site of infection in lung tissue. It was found that the depletion of B cells in non-human primates increased bacterial burden [103], which suggests a role for B cells beyond their antibody-mediated effector functions as antigen presenting cells. Otherwise, B cells could be controlling the immune response and susceptibility to infection via by secreting IL-10 [84].

Presently, knowledge about immune markers associated with TB disease in the cells described above is limited.

2.1.2 DURT immune response

Classical T cells respond to peptide antigens presented by MCH class I or II molecules. Non-peptide responsive T cell subsets have been found in humans and respond to various stimuli in a way that is not restricted to classical MHC-dependent antigen presenting systems. The resulting T

cell responses can be shared among a mixed population, and are responding to highly evolutionary conserved antigens, creating the concept of donor-unrestricted (DURT) T cells. The unrestricted nature of antigen presentation in DURTs holds great potential for vaccine strategies as one vaccine could target an entire population without respect to host genetic factors [73]. For this study, the DURT cells consisted of gamma delta ($\gamma\delta$) T cells, NKT-like cells and mucosal associated invariant T (MAIT) cells.

$\gamma\delta$ T cells are a small subset of T cells that are abundantly present in the human skin and large intestine [130][37]. These cells are distinguished from classical T cells, such as CD4+ and CD8+ lineages, by their expression of T cell receptors (TCRs) that are composed of γ and δ chains, instead of $\alpha\beta$ TCRs, expressed by conventional T cells. $\gamma\delta$ T cells recognize the phospho-antigens on the surface of *M.tb* infected cells presented by the BTN3A1 molecule (Figure 2.1) [73] .

$\gamma\delta$ T cells are activated by BCG-infected cells and evidence has shown their ability to lyse BCG-infected cells [121]. $\gamma\delta$ T cells could also have a protective role during *M.tb* infection. Specifically, a subset of $\gamma\delta$ T cells that is a major *M.tb* reactive, Vγ2Vδ2, was found to reduce *M.tb* pathology and infection after moderate to high *M.tb* challenge in rhesus macaques non-human primates. The subset co-produced IFN-γ and perforin and was established as a concept for incorporating immunogens stimulating this subset for a human TB vaccine [117]. This is in line with other studies that have also suggested that *M.tb*-specific $\gamma\delta$ T may have a protective role during *M.tb* infection [144][39][30]. Further studies need to be done to determine whether the $\gamma\delta$ cell populations are involved in the protection against *M.tb* infection, or whether they prevent LTBI progression in individuals who cannot produce IFN-γ. $\gamma\delta$ T cells could therefore either have a direct lytic effects on mycobacteria or could be activating monocytes to produce TNF, which will activate pathways to kill the mycobacteria [73]. Generally, these cells may play a protective role during infection, and will be examined in this study.

MAIT cells are innate T cells that have the capacity to secrete IFN-γ and TNF in the thymus

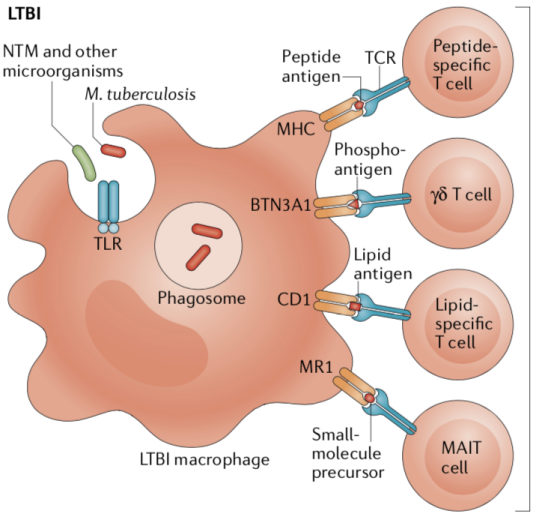

Figure 2.1: T cell responses to *M.tb* infected cells. The figure depicts the antigens detected by the different T cell populations. Credit: Simons et al. *Nature Reviews*, 2018.

and peripheral blood [56][76][55]. They also secrete IL-17 and Granzyme B, and respond to environmental signals, most notably IL-12 [57]. MAIT cells recognize a variety of bacterial metabolites presented by MR1 molecules (Figure 2.1), which are conserved MHC class I molecules. Evidence

of MAIT cells in the context of TB thus far is that patients with active TB show an activated phenotype, suggesting that MAIT cells are activated *in vivo* [112].

2.1.3 The immune response mediated by "conventional" T cells

T cells play a key protective role during *M.tb* infection and belong to the adaptive immune response. CD4+ T cells are T helper (Th) cells that recognize peptides presented on MHC class II molecules presented by antigen presenting cells. For protection against TB, CD4+ T cells are a particular subset of interest.

CD4+ T cell differentiation is typically modelled as a linear process (Figure 2.2). Once naive T cells are primed after interaction with antigen-MHC complex in the lymph nodes, CD4+ T cells can differentiate to central memory T (T_{CM}) and re-circulate within lymphoid tissues. These cells can progressively gain functionality until they reach an optimized stage (such as polyfunctional CD4+ T cells). T cells can then further differentiate into effector memory (T_{EM}) cells, which are found in peripheral blood and have specific effector functions. However, proloned antigenic stimulation can cause progressive loss of memory potential as well as cytokine production, resulting in terminally differentiated CD4+ T cells. These cells only produce IFN-γ and are short-lived. Therefore, initial anigen exposure or innate immune factors will control the degree of T cell differentiation [115]. Lastly, effector T (T_{EFF}) cells contribute to pathogen removal by moving into infected tissues [5].

In the context of TB, CD4+ T cells secrete cytokines that activate *M.tb*-infected macrophages to

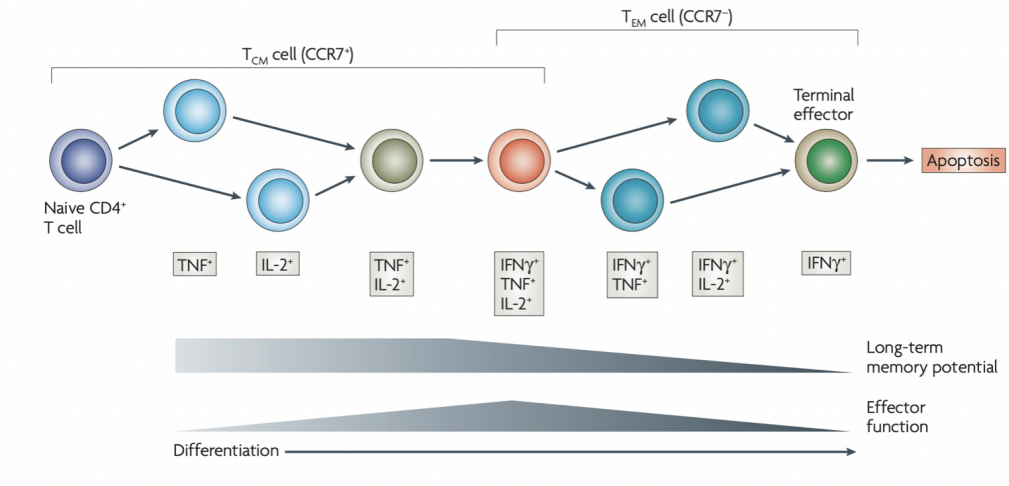

Figure 2.2: CD4+ T cell differentiation. Credit: Seder et al., 2008. *Nat. Review Immunol.*

contain the infection, to delay bacterial growth, and to bring immune cells to the granuloma. The most relevant cytokines that CD4+ T cells produce in the context of TB are termed "Th1" and include IFN-γ, TNF and IL-2. TNF, as previously mentioned, plays a key role in the formation of the granuloma, and individuals that are TNF deficient are highly susceptible to *M.tb* infection. IL-2, on the other hand, induces proliferation, assists the survival of TCR-activated T cells and is involved in the development of memory T cells during the initial stages of infection [140]. There is sufficient literature to support that CD4+ T cells that express IFN-γ play an important role in controlling *M.tb* replication and containing the bacteria within the granuloma [27][46][74]. This was validated by humans with abnormal IFN-γ receptors who were more susceptibility to disease [119].

However, the presence of IFN-γ+CD4+ T cells is not sufficient for immune control of *M.tb* infection. Further, CD154 is a functional marker for helper cells that produce Th1 cytokines, but also other functions [29]. Expansion beyond the narrow focus on IFN-γ producing T cells is therefore necessary to identify new biomarkers for novel vaccine strategies.

CD8+ T cells, like CD4+ T cells, are generated in the thymus but recognize peptides presented on MHC class I molecules found on all cells. CD8+ T cells are known as cytotoxic T cells and they release granules, which are similar to those in NK cells, and kill infected cells. Naive CD8+ T cells fully differentiate into activated effector CD8+ T cells. These cells are able to secrete cytokines and exhibit cytolytic activity on cells infected by intra-cellular pathogens. Similar to CD4+ T cells, prolonged antigenic stimulation of CD8+ T cells can result in terminal differentiation and apoptosis [115]. CD8+ T cells are less studied in the context of TB and found in lower magnitude during *M.tb* infection compared to CD4+ T cells [136]. There are also either too low, or undetectable frequencies of IFN-γ+CD8+ T cells in LTBI and active TB individuals [2]. Because of this, CD8+ T cells have not been studied as extensively as CD4+ T cells as biomarkers in TB.

Like CD4+ T cells, CD8+ T cells also have the ability to produce IL-2, IFN-γ, and TNF during

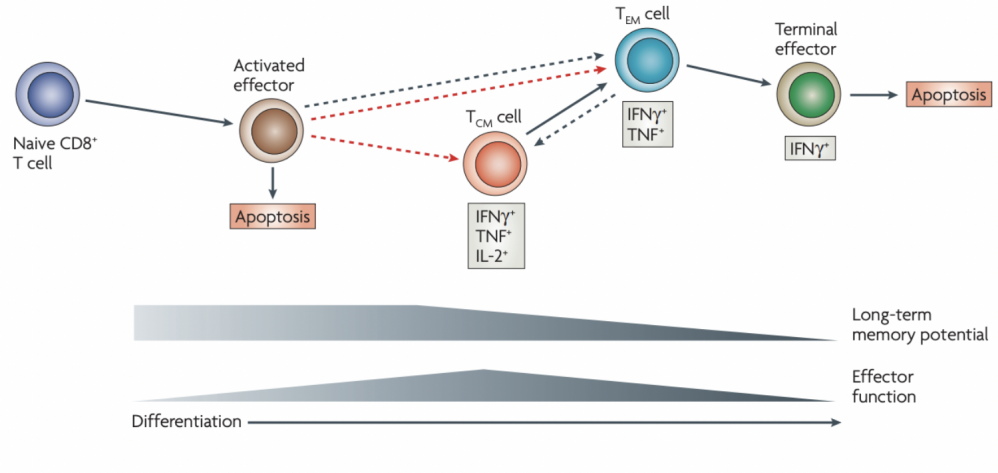

Figure 2.3: CD8+ T cell differentiation. Credit: Seder et al., 2006. *Nat. Review Immunol.*

M.tb infection (Figure 2.3). CD8+ T cells kill *M.tb*-infected cells via perforin and granzymes, to induce lysis. In humans, CD8+ T cell can produce granulysin, which directly kills *M.tb* [124]. In fact, antigen-specifici CD8+ T cells were found to lyse *M.tb* infected macrophages and kill the intracellular bacilli [31]. Therefore CD8+ T cells could play a role as a cytokine producing cell or have more of a cytolytic function, which can be measured by detecting CD107 expression on the surface of cells that degranulated [17].

It is widely known that there is a significant delay in the onset of detectable T cell responses during *M.tb* infection. Such delays can be explained by the inhibition of apoptosis in neutrophils and macrophages, which in turn inhibits the presentation of antigens by DCs, the main antigen presenting cells; and IL-10 production that inhibits the production of IFN-γ in the lungs [97].

2.1.4 TB biomarkers

A combination of phenotypic markers, namely CD45RA, CC-chemokine receptor 7 (CCR7), CD27, killer cell lectin-like receptor G1 (KLRG1), human leukocyte antigen-DR (HLA-DR) and CXCR3

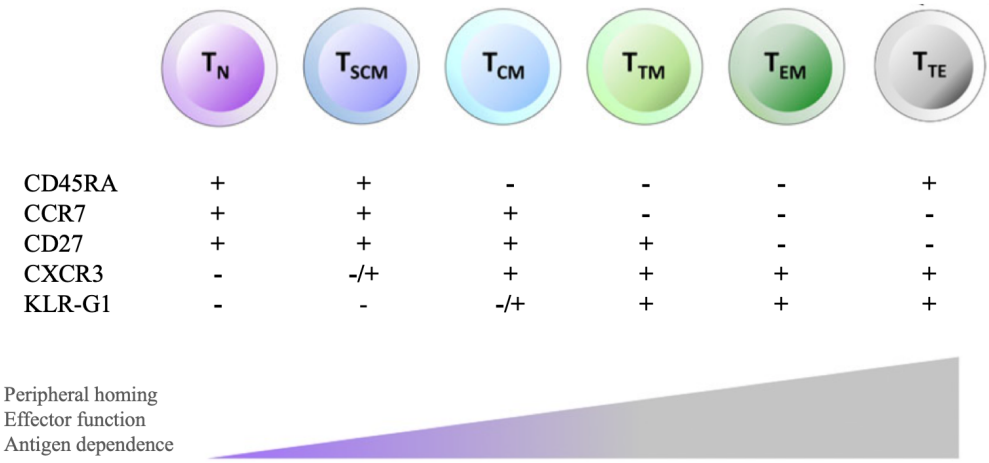

Figure 2.4: Phenotypic markers present on specific T cell subsets during T cell differentiation. T cells are categorized based on their effector functions and phenotypic markers into naive T cells (T_N), stem cell memory T cells (T_{SCM}), central memory T cells (T_{CM}), transitional memory T cells (T_{TM}), effector memory T cells (T_{EM}) or terminal effector T cells (T_{TE}). A "+" symbol means that the phenotypic marker is present on that specific T cell and "-" means it is absent. Figure adapted from Mahnke et al., 2013. *Eur. J. Immunol.*

were measured on CD3+ IFN-γ+, IL-2+ or TNF+ producing T cells (total Th1) for this project. HLA-DR is a MHC class II cell-surface receptor that plays a role during antigen presentation and is expressed highly on activated antigen-specific T cells. It was found to be an early response immune marker that indicates T cell activation in response to bacteria. Many studies [91][2][139][105][106][118] hypothesized that HLA-DR would be expressed on *M.tb* specific CD4+ T cells *in vivo* and the frequency would increase proportionally to bacterial load and systemic inflammation. In all studies, they confirmed that the frequencies of HLA-DR+ IFN-γ+CD4+ T cells in response to E6C10 stimulation were significantly higher in individuals with active TB than LTBI, and one study demonstrated that the frequency was even higher in individuals with extra pulmonary TB [118]. These relationships observed were regardless of HIV status. Adekambi and colleagues (2015) [2] were also able to use this marker to distinguish untreated active TB from successful anti-TB treatment, which correlated with decreased *M.tb* loads during treatment. In addition they found HLA-DR+ *M.tb*-specific CD4+ T cells correlate with *M.tb* burden *in vivo*. Lastly, Vickers and colleagues [134] studied sputum culture positivity in response to TB treatment in HIV negative, active TB patients from The Gambia. They found that slow treatment responders (culture positive at 2 months but negative by 6 months) had higher frequencies of PPD-specific CD4+CD27+HLA-DR+CD38+ compared to fast responders (culture negative by 2 months) at baseline. The authors concluded that the expression of this T cell subset could be used to predict speed of response to TB treatment [134].

A general trend seen in literature is that differentiated effector T cells, that is decreased expression of CD27 or CCR7, or the expression of KLRG1 (Figure 2.4) on CD4+ T cells in both mice and humans, was found to be a biomarker of active TB [63][67][125][72][96][51][114][78][94]. It's hypothesized that this is due to persistent antigenic stimulation [36]. Another study found that KLRG1-deficient mice had improved control of, and increased survival after *M.tb* infection compared to wild type mice [34]. The same study also showed that CD4+ T cells were affected by the presence of KLRG1 during infection, which further suggests a destructive role of KLRG1 expression

on T cells during *M.tb* infection.

Very few CD8+ T cell-based biomarkers for TB diagnosis have been described. Some studies have found that differentiated CD8+ T cells with the phenotype CD45RA+CCR7- was characteristic of LTBI individuals and was lower in active TB patients compared to individuals with LTBI [26][108][35]. Active TB patients in a South African study were mostly CD45RA+CCR7+CD8+ [9], however, Rozot et al. [108] and Day et al. [35] found that active TB was characterized by effector or central memory CD8+ T cells (CD45RA-CCR7-).

This project aimed to compare recent versus established *M.tb* infection defined by QFT tests. Besides documented QFT or TST conversion on serial testing, no validated biomarkers exist that can successfully distinguish recent from established infection in a cross-sectional fashion, which are clinically indistinguishable. Most T cell biomarkers that exist are ones that distinguish latent from active TB disease, as summarized above. To our knowledge, three such studies exist that have successfully identified candidate biomarkers of recent TB infection (TBI), of which two are presented below and the other is presented in Chapter 5.

A study by Halliday et al. (2017) [59] was able to define markers that are linked directly to time since infection, a primary risk factor for active TB [12]. The study was based on a longitudinal design and consisted of active TB patients, individuals with recent latent infection (patients that had been exposed to TB within 6 months prior to the study) and patients with remote latent infection (patients that were born in a high TB burden country but who have moved and haven't been exposed to TB since). LTBI was confirmed by a positive T-spot, QFT or TST test. Similar to the marker that distinguishes latent from active TB, the frequencies of TNF+ T_{EFF} cells were significantly different between the three groups. *M.tb*-specific T cells from active cells of active TB patient had the largest proportion of TNF+ T_{EFF} cells, followed by the recently infected LTBI patients and then the remotely infected LTBI patients. This biomarker was able to distinguish individuals with recent versus remote infection with a sensitivity of 89% (95% CI: [43% – 93.95%]).

In Sweden, a study was carried out where individuals in contact with persons with pulmonary TB were recruited and tested repeatedly with IGRA. The findings from this study indicated that recent infection corresponded to an increased proliferative response in CD4+ cells in response to C10 and purified protein derivative (PPD), and a low response to E6. These responses after early exposure (less than one month) were able to predict recent TBI with an AUC of 0.69 [19]. Another study aimed to identify gene expression patterns in blood RNA that correlated with time since *M.tb* exposure or infection [8]. Firstly, in mice and cynomolges macaques, they found RNA signatures that could discriminate between early and later stages of infection [19]. In humans, they found a 250-gene and 6-gene RNA signature for recent infection [8]. Using an independent cohort consisting of longitudinal samples of adolescents that acquired *M.tb* during a 6-month follow up, the 6-gene signature was able to discriminate pre- and post-IGRA converted time points with an AUC of 0.68.

In summary, many combinations of *M.tb*-specific cytokine production and phenotypic markers show promise to distinguish LTBI from active TB disease. Very few studies have looked into biomarkers of recent TBI in comparison. This project will further look into whether any of these could also be used to distinguish early from established LTBI.

2.2 QFT dynamics and clinical outcomes

No reliable test exists that can successfully detect the presence or absence of *M.tb* in asymptomatic individuals. The IGRA and TST are classically used to support diagnosis of *M.tb* infection. Neither test can differentiate between active and latent TB or reactivation from reinfection, nor can they

indicate the time since *M.tb* infection [116][87].

IGRA is ideal for serial testing because they are unaffected by BCG, only require one visit, the tests are repeatable and have standardized interpretations. The downfall of the IGRA tests lies with inherent variability of T cell assays (Figure 2.5). Functional assays are susceptible to variability, which makes it difficult to use a single cut-off value to determine whether the test is positive or negative with once-off testing, and to define conversion and reversion in longitudinal studies. Such variability can arise from manufacturing issues (temperatures during shipping), preanalytical sources (time when blood was collected, the volume of blood collected, test tube shaking), analytical sources (imprecision of pipetting, centrifugation, decanting or washing) and immunological sources (immune boosting). Sources of variability in the IGRA can be eliminated or reduced through standardization by assay manufacturers. If the sources of variability are random and unavoidable they should be accounted for when interpreting results. Since the study groups were defined based on their QFT results, this is a limitation in this particular project.

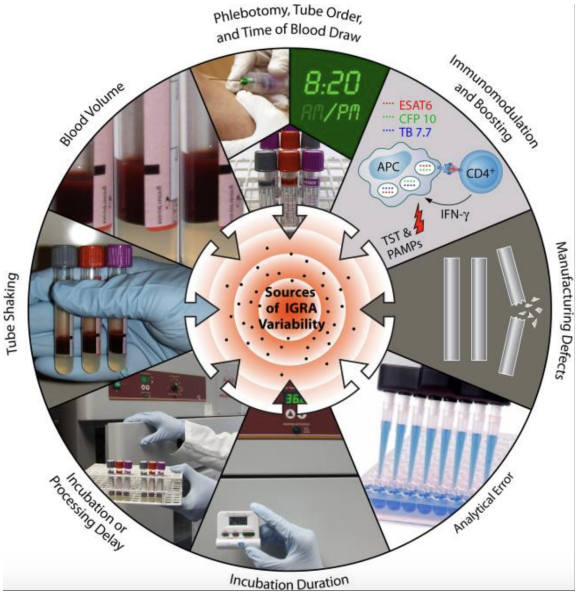

Figure 2.5: Sources of variability in IGRA tests. Credit: Pai et al., 2014. *Clinical Microbiology Reviews.*

In order to assess QFT conversion and reversion dynamics in a high TB endemic setting, an epidemiological study was carried out from May 2005 through February 2009. Students between the ages of 12 and 18 years were recruited from 11 local schools in the town of Worcester in the Western Cape and QFT and TST tests were performed every 6 months during a two year follow-up period. Among a total of 5,357 participants, 2,751 (51.4%) and 2,987 (55.8%) had positive QFT and TST results at baseline respectively. Annual QFT and TST conversion risks were calculated and found to be 14.0 and 13.0% respectively. The annual QFT and TST reversion risks were 5.1 and 4.1% respectively. Concordance between the TST and QFT tests was good for conversions (κ = 0.74), but poor for reversions (κ = 0.12). Additional analyses were performed by Machingaidze et al. (2012) [82] who found that in recent QFT converters, the TB incidence rate was 1.46 cases per 100 person-years, and the cumulative incidence was 2.8%. A significantly lower TB incidence rate (0.17 cases per 100 person-year) and cumulative incidence (0.32%) was observed for persistent

QFT non-converters. Therefore recent QFT conversion is associated with an increase of TB disease incidence when compared to persistent QFT negatives. Subsequently, when looking at the recent QFT converters, it was found that the risk of reversion was inversely correlated with the magnitude of the QFT test value (p = 0.0001). Incident tuberculosis was 8-fold higher in the QFT reverters than in participants that were persistently QFT negative (1.47 vs. 0.18 cases/100 person-years, p = 0.011) [6].

Due to the immunological and analytical variability potentially resulting in inaccurate results, stricter rules than the manufacturer's recommendations were applied to define QFT conversion, with the aim of enhancing the consistency of longitudinal QFT to interpret test conversions. This was achieved by introducing an uncertainty zone between QFT values of 0.2-0.7 IU/mL [93]. In the same adolescent cohort study from which study groups for this project were selected, a stringent converter was defined as QFT conversion from < 0.2 IU/mL to > 0.7 IU/mL. When compared to stringent non-converters (all tests < 0.2 IU/mL), stringent converters had a 10-fold higher risk of developing TB. "Uncertain" QFT converters (values falling within the uncertainty zone) were not different from the stringent non-converters, and conversion values to less than 0.7 IU/mL were associated with a 50% probability of reversion. Nemes and colleagues (2017) therefore showed the significance of redefining QFT conversions, which was applied to this study.

The clinical implication of reversions is still unclear. It is possible that QFT reversion is associated with clearance of *M.tb*, or it could be as a result of technical assay variability. The QFT value at conversion, on the other hand, was found in a separate ACS study to be strongly inversely associated with the risk of reversion [141][7]. Lastly, higher QFT conversion values were found to be predictors for the increased risk TB disease [141][7].

Very little research has investigated immune markers associated with QFT conversion and reversion. An adolescent study was conducted by Jenum et al. (2014) [71] in India, which, similar to South Africa, has one of the highest incidences of TB in the world. PBMCs were collected every 6 months and individuals that were classified as reverters and persistent positives were compared and used for the analysis. They hypothesized that the reverters and persistent positives would differ with respect to the *M.tb*-specific T cell responses. Analysis found that PPD-specific polyfunctional (IFN-γ, TNF, IL-2) CD4+ T cells frequencies were high in both groups with no significant difference. This study did not include a follow-up period but suggested that reversion was associated with a reduced risk of progression, which they were not able to show in the study. We aim to get a better understanding of reversion, and the immune response associated with reversion in our study.

Chapter 3

Datasets and data manipulations

3.1 Introduction

Every system in our body relies upon the interaction between system components at a cellular level and fine integration and regulation at the system level. Therefore, by its nature, such complex systems cannot be predicted from the behaviour of any of its individual parts separately but rather as a whole functional network. Most immunological knowledge in the context of TB thus far has come from deconstructing the immune system into its various parts, despite the immune system being a complex, multi-level interaction network. Data integration allows for a more comprehensive analysis of how such complex systems works, considering many biological and immunological applications.

Data integration presents several challenges, typically

1. the high dimensionality of the data after integration,
2. different scales of different data types, and
3. missing values that arise in the dataset due to some individuals or time points not being available in each data table.

This chapter will briefly introduce the datasets used in this dissertation and how they were generated, and discuss how the three challenges above were overcome.

3.2 The datasets

Figure 3.1 summarizes the features of the immune response that this project is focused on, and provides an overview of the variables and their quantities. The two datasets contained cytokine expressions and phenotypic markers measured on features from both the adaptive and innaptive immune responses.

3.2.1 Biological assays

The laboratory methods described below were provided by members of the South African Tuberculosis Vaccine Initiative (SATVI) who generated the data.

Classical T cell responses

Most proteins expressed by *M.tb* are also present in the *M. bovis* BCG vaccine (administered at birth in an endemic setting such as South Africa) and non-tuberculous mycobacteria (NTM), often

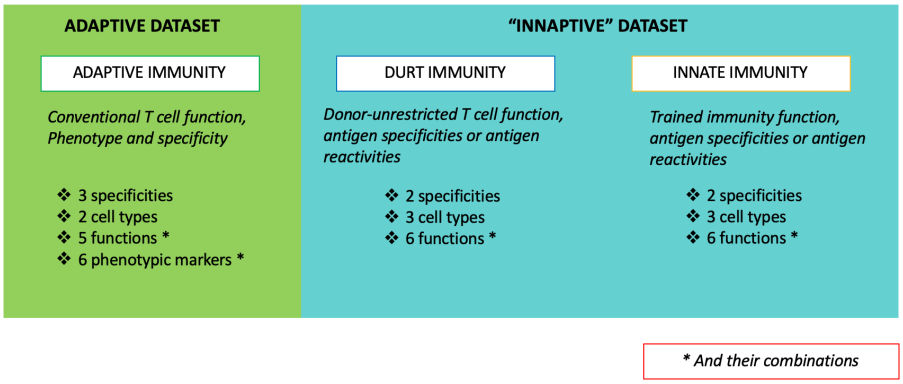

Figure 3.1: The datasets. A breakdown of the cell types, functions and phenotypic markers investigated in the adaptive and "innaptive" datasets.

found in water and soil. To detect *M.tb*-specific T cell responses, peptides specific for *M.tb* were used to stimulate cryopreserved PBMCs for 18hrs. As a negative control cells were left unstimulated to determine background responses.

- ESAT-6 and CFP-10, which are two immunodominant antigens (structure that most T cells tend to recognize) expressed by *M.tb* but deleted from the genome of BCG and absent in NTM. These are the same antigens used in the QFT assay.
- EspC, EspF and Rv2348c are three highly recognized *M.tb*-specific antigens [110]. Including these antigens will show whether T cell responses are restricted to antigens used in the QFT.
- *M.tb*-lysate contains a mixture of all antigens present in *M.tb*, some of which are present in BCG or NTMs. In addition, *M.tb*-lysate contains non-peptide antigens including lipids, phospho-antigens and metabolites that are recognized by DURTs. We expect to see a response in all individuals as they are BCG vaccinated and have likely been exposed to NTMs.
- Staphylococcus Enterotoxin B (SEB), a super-antigen that stimulates all T cells, served as a positive control. The responses to this stimulus were not included in the analysis.

Figure 3.2 summarizes the cell type, effector function and phenotypic markers of T cells. This dataset consisted of a total of 226 variables including a combination of 5 effector functions (IL-2, CD107, CD154, IFN-γ and TNF) produced by CD4+ and CD8+ T cells under the three stimulations. A combination of phenotypic markers (CD45RA, CCR7, CD27, KLRG1, HLA-DR and CXCR3) were then measured on IFN-γ, IL-2 or TNF producing T cells (total Th1), when stimulated with E6C10 or *M.tb*-lysate.

Effector responses were background subtracted (subtracting the frequencies detected in corresponding unstimulated samples from frequencies in stimulated samples), while the phenotypic markers were expressed as proportions of Th1 cells. Further, phenotypes were only measured in "responding" samples (see below). Background subtraction is performed to ensure that cytokines produced by cells in response to a stimulation are purely a response induced by that antigen (antigen-specific) and not due to other factors such as processing of samples.

Innate and DURT cell responses

M.tb-lysate, which contains non-peptide antigens including lipids, phospho-antigens and metabolites was used to stimulate the DURT and innate cells. Thawed PBMCs were left either unstimulated

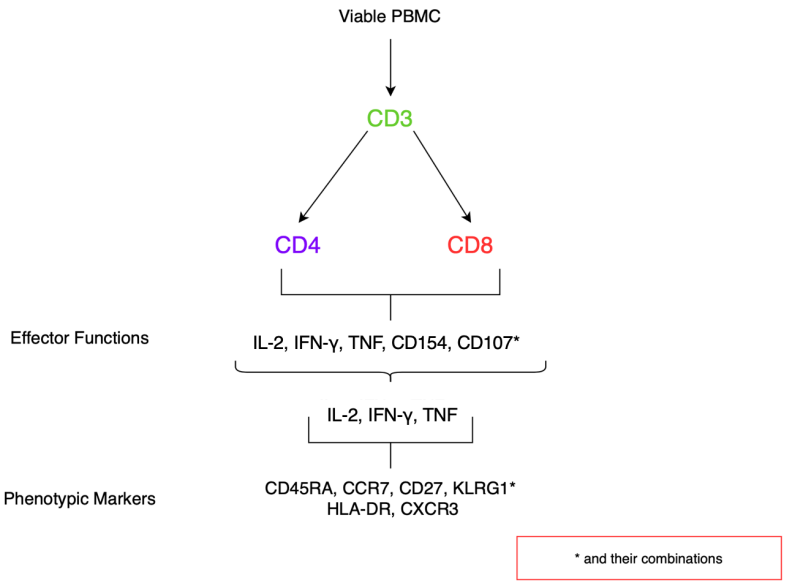

Figure 3.2: Cell types, effector functions and phenotypic markers for classical T cells.

or were incubated for 6 hours with:

- *M.tb* lysate, which contains many antigens that are also expressed by BCG and NTM.
- *E. coli* was used to assess whether differences in potentially cross-reactive immune responses are restricted to mycobacteria or also displayed to other bacteria. This stimulus serves as a positive control as it is well known from literature that innate and DURTs recognize *E. coli*. The responses of cells to this stimulus were not included in the analysis.

The phenotypes and effector functions for this panel of cells are summarized in Figure 3.3. The innaptive dataset consisted of 283 variables including a combination of 6 functions (Granzyme-B (GB), IL-6, IL-10, IL-12, IFN-γ and TNF) produced by non-T cells (NK cells, B cells, monocytes) and DURTs (MAITs, $\gamma\delta$ T cells, NKT cells).

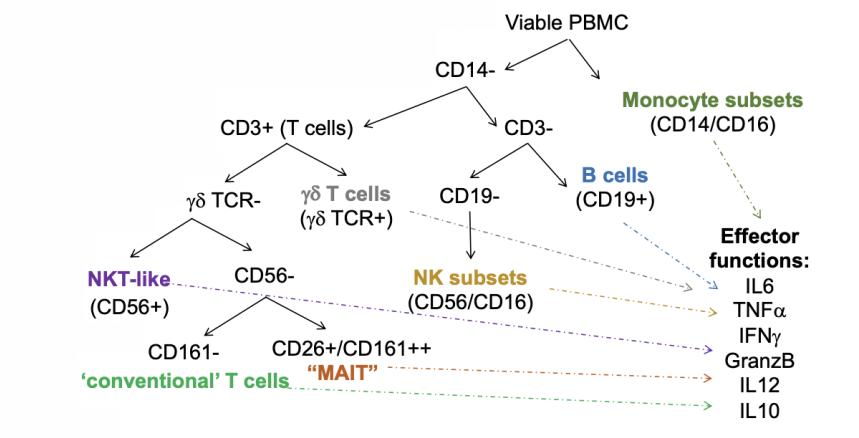

Figure 3.3: The phenotypes and effector functions measured on innate and DURT cells.

3.3 Data integration

Integration was performed by aligning each dataset according to participant ID (PID), QFT status (positive or negative) and month of sample collection (0, 6, 12 and 18). However, several nuances in each dataset made the integration, and hence the analysis, more complicated.

In the innaptive dataset, either month 0 and month 6, *or* month 12 and month 18 were chosen at random for each participant in the persistent QFT negative and positive groups to generate this dataset. Figure 3.4A illustrates this. Therefore, participants in these two groups had a maximum of two time points each, and their full longitudinal profiles were not available. For the adaptive dataset, specifically the data containing the phenotypic markers on total Th1 cells, I received the data with all four time points in the persistent QFT negative cohort averaged for each individual and variable. Similarly, for the individuals in the reverter cohort, I received the data where an average of the two pre- and post- reversion time points were taken separately for each individual and variable (Figure 3.4B).

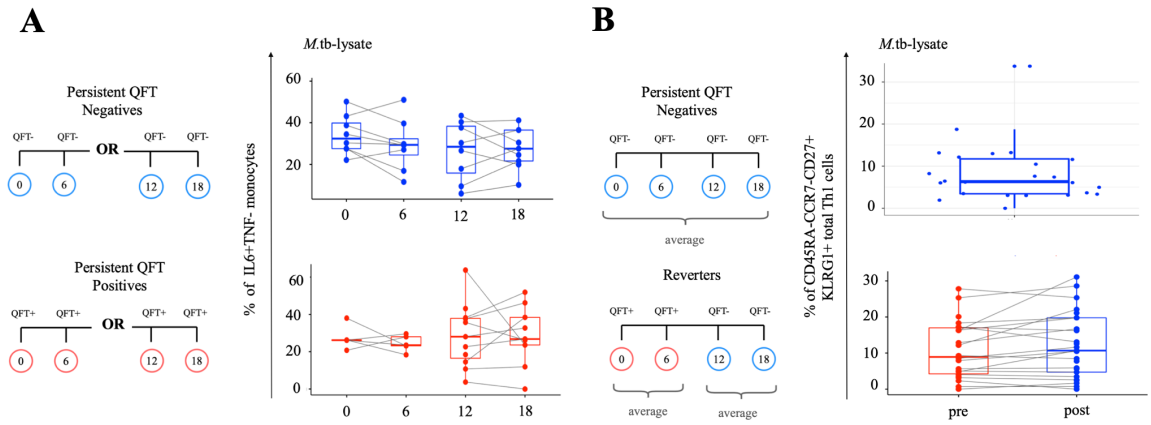

Figure 3.4: Nuances in the two datasets that made integration and analysis a challenge. (**A**) In the innaptive dataset, either month 0 and month 6, or month 12 and month 18 were chosen at random to generate the dataset. The paired boxplots of *M.tb*-lysate-specific IL6+TNF- monocytes illustrates this and therefore the full longitudinal profiles in these two cohorts for this dataset were not available. (**B**) An average was taken for the four time points available in the persistent negative cohort for the data on phenotypic marker expressions on total Th1 cells. Further, for the QFT reverters, an average of the two pre- and post-reversion time points were taken separately. This is illustrated in the expression of CD45RA-CCR7-CD27+KLRG1+ *M.tb*-lysate-specific total Th1 cells. For the persistent negatives, a single time point per individual summarized the group, while for the QFT reverters, two time points per individual were available.

At the time when the datasets were generated, it was not known that they would be used in a data integration project, hence the different datasets were not generated in a consistent matter. This was a major limitation in this project as we could not use all the time and data points available. Median values had to be taken for each individual in each of the four cohorts stratified by QFT status (Figure 3.5). For example, an individual in the recent QFT converter cohort would have a maximum of two samples for each variable, a pre-conversion time point and a post-conversion time point, while an individual in the persistent QFT negative time point would only have one. This created some consistency across the two data types.

Friedman's [50] non-parametric test was used to compare the four paired sampling occasions

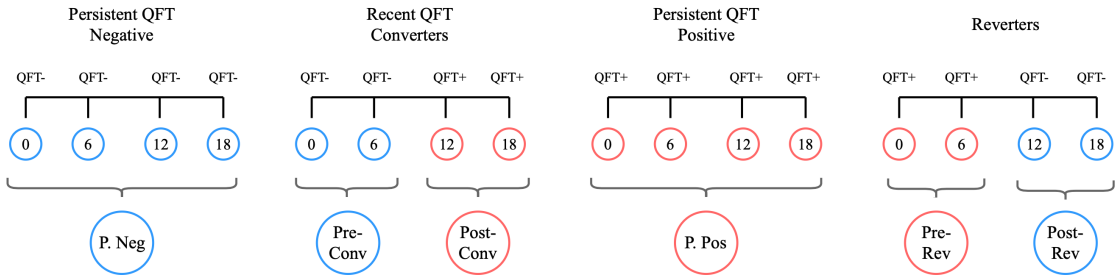

Figure 3.5: How the multiple time points were dealt with. The curly braces indicate where the median values were taken, stratified according to group and QFT status, for each individual and variable separately.

in the two control cohorts to assess whether there were any differences between the four time points. For the recent QFT converters and reverters, Wicoxon's signed rank [138] test was used to compare months 0 and 6, and months 12 and 18 (Figure 3.6). These tests were repeated for each variable in the integrated dataset and p-values were corrected using the Benjamini-Hochberg (BH) [14] protocol. The results indicated that there were no significant differences between the time points compared and therefore we were able to justify taking their median values.

3.4 High dimensionality

The first challenge to overcome was the high dimension of the data post-integration, where the number of predictors was larger than the number of samples. The problem related to high data dimensionality is typically solved by employing dimension reduction techniques. The most common dimension reduction technique is principal component analysis (PCA) [101], which replaces the observed variables with fewer underlying latent variables. An alternative approach is to pre-filter the data, thereby excluding a set of the less important variables. We opted to pre-filter the data to identify and remove biologically meaningless subsets in the data. In addition, PCA was difficult to implement due to the large degree of missingness in the data (see below).

3.4.1 Filtering the adaptive dataset

The adaptive data was generated by flow cytometry using intracellular cytokine staining (ICS). This type of assay is specifically advantageous as it allows multiple phenotypic, differentiation and functional parameters related to antigen-specific T-cells, most notably, the expression of multiple effector cytokines (polyfunctional) in individual cells, to be assessed simultaneously.

Combinations of five effector functions were measured for CD4+ and CD8+ T cells, namely IFN-γ, TNF, IL-2, CD154 and CD107 (Figure 3.2). This resulted in $2^5 = 32$ possible combinations of co-expressed functions, not all of which are biologically meaningful. We therefore employed COMPASS (Combinatorial Polyfunctionality analysis of Antigen-Specific T cell Subsets) [80], a method that employs a Bayesian hierarchical model in order to identify biologically relevant cell subsets.

The COMPASS method is as follows. Assume there are I subjects with both antigen-specific T cell responses and unstimulated T cell response counts. Let M be the number of measured markers, with $K_M = 2^M$ Boolean combinations defining the functional subsets. K, where $K \leq K_M$, is used to denote the true number of cell subsets considered for statistical analysis, as some cell subsets may be sparse. Further, define n_{sik} to be the observed counts for the k cell subsets in the stimulated

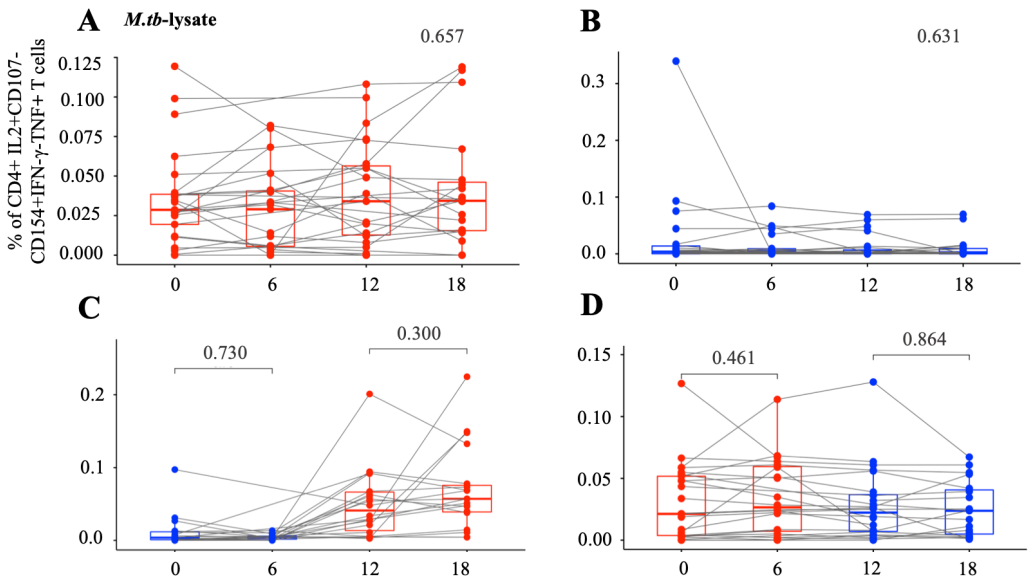

Figure 3.6: Justification for taking median values of the indicated time points. Paired boxplots for *M.tb*-specific frequencies of CD4+IL-2+CD107-CD154+IFN-γ-TNF+ T cells in (**A**) the persistent QFT positive cohort, (**B**) the persistent QFT negative cohort, (**C**) the recent QFT converter cohort and (**D**) the reverter cohort. Friedman's test was used to compare the four time points in the two control groups (**A** and **B**) and their respective p-values are superimposed onto the plots. Both p-values indicated that there were no differences between the four time points in the control groups. Paired Wilcoxon tests were then used to compare the two pre- and post- conversion and reversion time points in (**C**) and (**D**) respectively. The resulting p-values are superimposed onto the plots and indicated that there were no differences.

samples, and, similarly, n_{uik} be the observed counts for the unstimulated samples. For each subject $i = 1, \ldots, I$, the counts are jointly modeled using the following multinomial distributions:

$$(n_{si}|p_{si}) \sim \text{MN}(N_{si}, p_{si})$$

$$(n_{ui}|p_{ui}) \sim \text{MN}(N_{ui}, p_{ui})$$

where p_{si} and p_{ui} are unknown proportion vectors of the paired samples for the stimulated and unstimulated counts respectively and $N_{si} = \sum_k n_{sik}$ and $N_{ui} = \sum_k n_{uik}$. In order to detect antigen-specific subsets within a subject, the following hypothesis is then considered and tested: $H_0 : p_{ui} = p_{si}$ against $H_a : \exists k \in \{1, \ldots, K-1\}$ such that $p_{sik} > p_{uik}$. Essentially, under the null hypothesis, the assumption is that there is no difference between the proportion of stimulated and unstimulated cells. Under the alternative hypothesis, cell subsets that express at least one function and are different are considered antigen-specific. Therefore, for each subject, COMPASS has the ability to quickly identify antigen-specific cell subsets.

A binary indicator, $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{iK})$, is introduced such that $\gamma_{ik} = 1$ if $p_{uik} \neq p_{sik}$ and 0 otherwise, for $k = 1, \ldots, K$. The joint model then becomes

$$(n_{si}|p_{si}, \gamma_i) \sim \text{MN}(N_{si}, p_{si}\gamma_i + p_{ui}(1 - \gamma_i))$$

$$(n_{ui}|p_{ui}, \gamma_i) \sim \text{MN}(N_{ui}, p_{ui}).$$

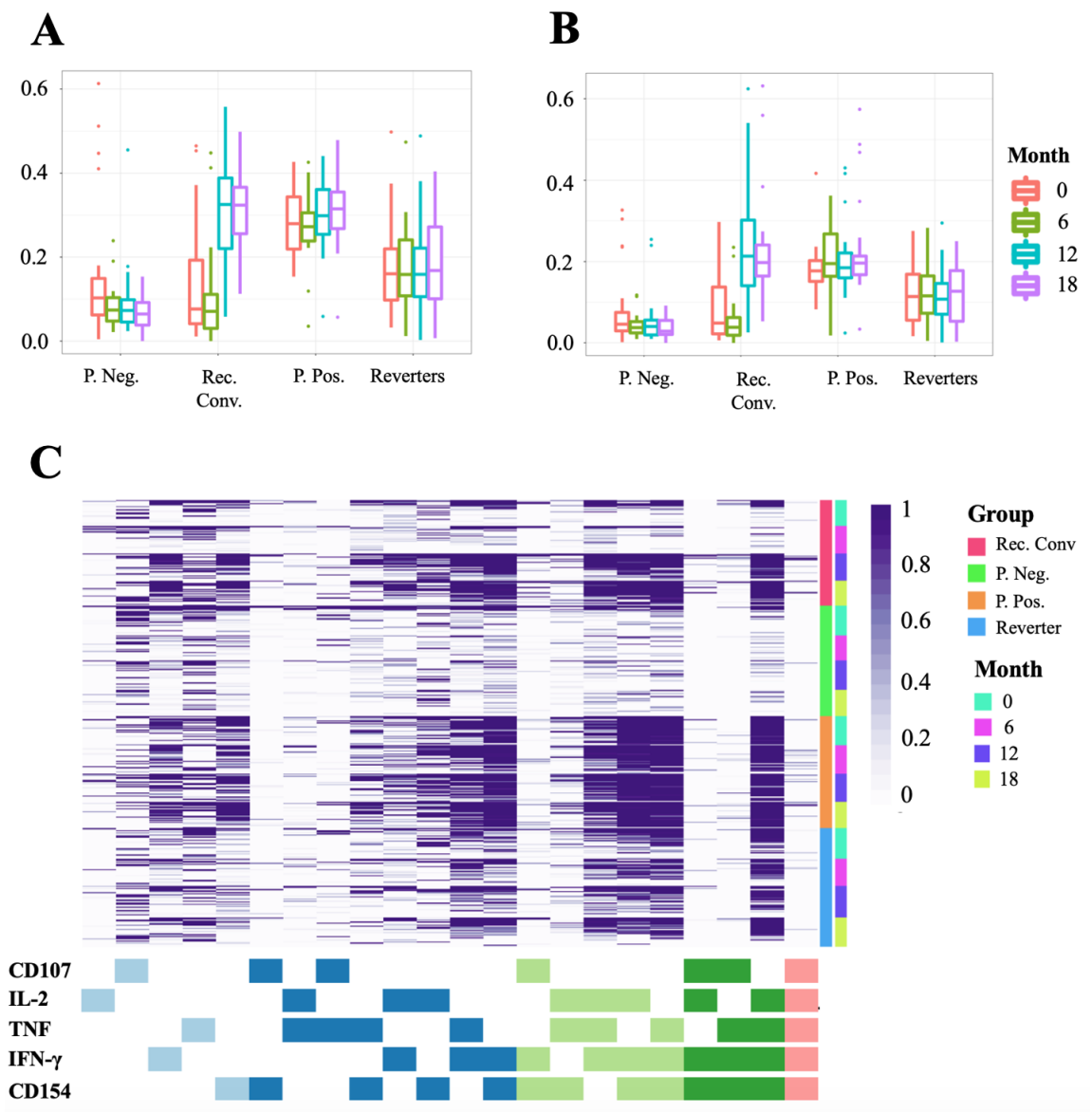

Figure 3.7: Output from COMPASS. COMPASS provides two scores, (**A**) a functional score and (**B**) a polyfunctional score, which both can be directly correlated with any clinical outcome of interest. To identify biological relevant subsets, COMPASS also outputs a heatmap (**C**) of the posterior probabilities of an antigen-specific response for each binary combination of cytokines. This is an example of CD4+ T cells stimulated with E6C10, where a darker colour indicated a higher posterior probability, while white equals zero. Subsets with low posterior probabilities corresponding to low antigen-specific responses are identified as biologically irrelevant.

Therefore, p_{si}, p_{ui} and γ_i are considered the prior parameters in the model and a Bayesian Markov chain Monte Carlo (MCMC) algorithm is employed to extensively explore the joint posterior distribution (see [80] manuscript for full details). Antigen-specific responses for any cell subset of interest are available as the posterior mean of the latent indicators. These posterior probabilities can then be used to quantify cell subset responses, which are available in the form of a heatmap (Figure 3.7C).

In addition to the above, COMPASS calculates two scores that can be used to summarize the different responses for the cell subsets for each subject. The two scores are the functional score (FS) and the polyfunctional score (PFS) (Figure 3.7A and 3.7B respectively). The FS is defined as the proportion of antigen-specific cell-subsets detected among all possible ones.

$$\mathrm{FS}_i = \frac{\sum_{k=1}^{K-1} \hat{\gamma}_{ik}}{(K_M - 1)}$$

where $\hat{\gamma}_{ik}$ is the posterior mean of γ_{ik} estimated using MCMC. The FS will be a score between 0 and 1 and will be higher for antigen-specificity in more cell subsets regardless of the degree of functionality for that subset.
The PFS is an extension of the FS and weights different subsets by their degree of functionality.

$$\mathrm{PFS}_i = \frac{\sum_{k=1}^{K-1} \hat{\gamma}_{ik} \text{ x } d(k)/\binom{M}{d(k)}}{M \text{ x } (M+1)/2}$$

where $d(k)$ is the degree of functionality for the k-th cell subset (the number of cytokines the cell subset is positive for). The PFS will be higher for antigen-specific cell subsets that have a higher degree of functionality. Both the FS and PFS scores can be directly correlated with a clinical outcome of interest.

In summary, COMPASS was used to filter the adaptive dataset and identify biologically meaningless cell subsets. Any subset that had more than 10 observations (one third of the number of samples in one cohort) at one of month 0, 6, 12 or 18 with a posterior probability greater than 0.1 was retained. Otherwise the subset was identified as biologically irrelevant and was removed. For example, in Figure 3.8, CD4+ T cells stimulated with E6C10 with a joint expression of all five cytokines (all positive) would be omitted because the number of observations with posterior probabilities > 0.1, at any of the sampling occasions, were less than 10. This filtering protocol was applied to both the CD4+ and CD8+ T cells for each stimulation separately.

Since COMPASS could not be used to filter the phenotypic markers expressed on T cells, and because background expression cannot be subtracted on these markers, the markers were only measured in stimulated samples from responder individuals. This was done using MIMOSA [45], which identified which individuals had a significant antigen-specific T cell response over background (unstimulated condition). MIMOSA identifies responding subjects by testing whether proportions of cytokine-producing cells in stimulated and unstimulated samples are different from each other using an MCMC algorithm. In this project, using MIMOSA, responders were defined as individuals with a total Th1 response in the stimulated samples that had a 3-fold change over unstimulated samples, with a corresponding MIMOSA false discovery rate p-value less than or equal to 0.01.

3.4.2 Filtering innaptive dataset

Cell types that are classified as part of the innate immune response or DURT cells have high background (unstimulated) values. Consequently, COMPASS could not be used to filter these cell types, which was designed for T cell and aims to identify cells that have significantly higher responses than background (antigen-specific). Therefore, we had to design a filtering approach for this dataset. The method is outlined in Figure 3.9.

Six functions were measured on cells from the innaptive dataset, namely GB, IFN-γ, TNF, IL-10, IL-12 and IL-6 (Figure 2.3). Both total frequencies of cells expressing each function as well as their binary combinations were generated. These functions were measured in all NK cells, NKT

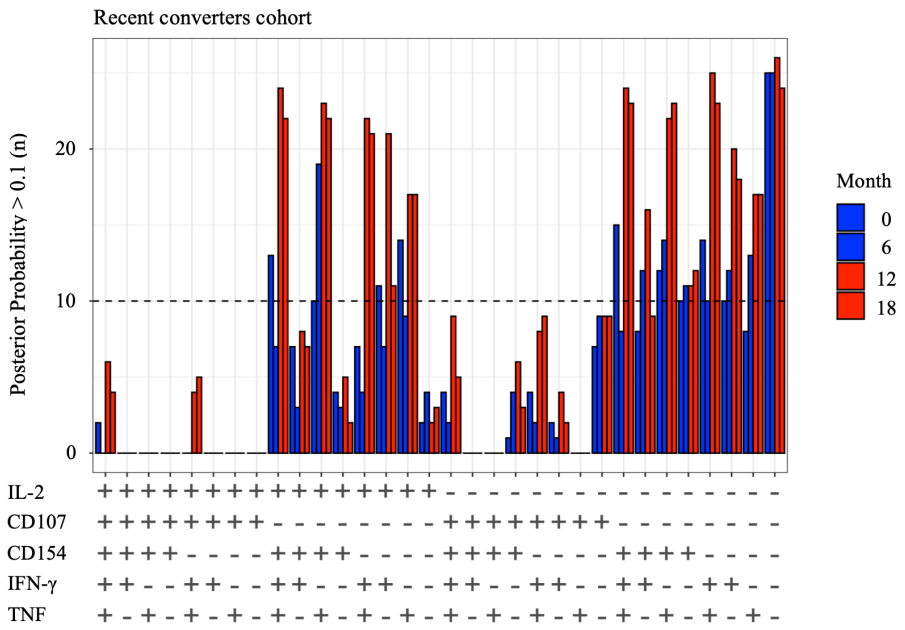

Figure 3.8: The criteria for filtering the adaptive dataset. Figure shows the number of observations for CD4+ T cell counts stimulated with E6C10 in the recent converters that had posterior probabilities greater than 0.1 for each binary combination, stratified according to month. A subset was classified as biologically meaningful if the number of observations with posterior probability values greater than 0.1, at any of month 0, 6, 12 or 18, was greater than 10.

cells, MAITs, monocytes, $\gamma\delta$ T cells and B cells, regardless of whether the cell is biologically capable of producing these cytokines or not. Hence, many subsets were biologically irrelevant.

In order to filter variables of the innaptive dataset, we began by defining a threshold value that would identify whether a cell subset was detectable or not. A number of different thresholds were tried and tested, but we decided on the thresholds below by observing which cell subsets were being removed. Based on literature and knowledge of innaptive cell subsets, we could assess whether the subsets removed were biologically meaningful or not. The majority of the cell subsets removed by this threshold, to our knowledge, were bioligcally irrelevent. The threshold was defined as followed: responses were considered as detectable if

- the upper bound of the 95% confidence interval (CI) for the median (found by boostrapped methods) was non-zero, and
- one third of all samples have values that were greater than zero.

As a first step, this detection criterion was applied to the total cytokine variables. The goal was to identify which cytokines each cell type was able to produce in response to either *M.tb*-lysate or *E.coli* (the positive control for this dataset) stimulations. If the total cytokine variable was considered undetectable for either of the two stimulations, then the variable was removed from the analysis. In addition, that cytokine would be removed from the binary combination. Assume that a specific cell type had a total IL-6 count that was considered undetectable, but the total counts for GB and IFN-γ were detectable. All subsets that are IL-6+ would be summed over, and as a result 'removed' from the binary combinations. Mathematically, this would be

$$(\text{GB+}\mathbf{\text{IL-6+}}\text{IFN-}\gamma\text{+}) \;+\; (\text{GB+}\mathbf{\text{IL-6-}}\text{IFN-}\gamma\text{+}) \;=\; \text{GB+IFN-}\gamma\text{+}.$$

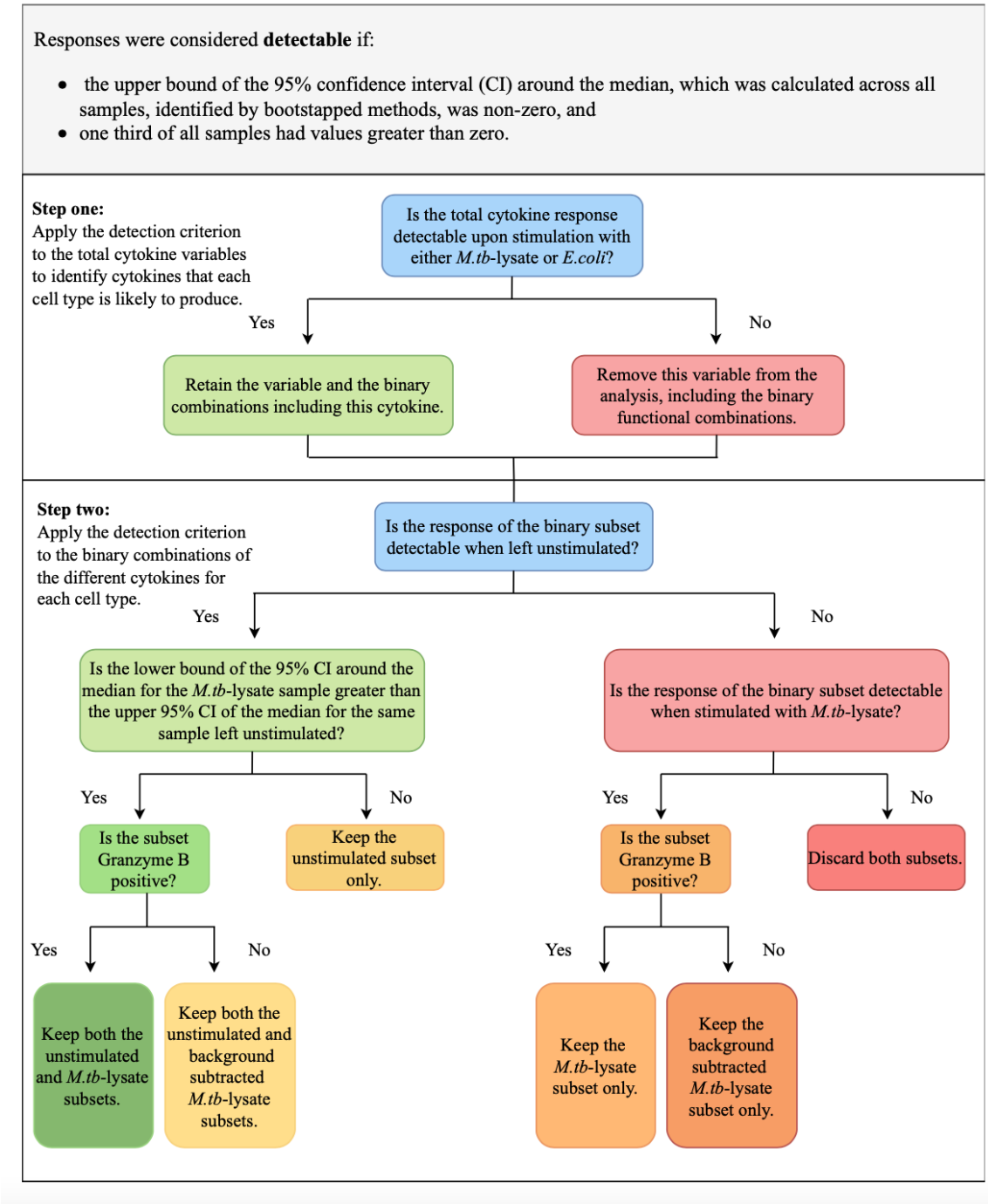

Figure 3.9: Flow chart for filtering the innaptive data.

Step two in the filtering process was focused on removing the biologically irrelevant binary combinations of the different cytokines for each cell type. We first tested whether the binary subset was detectable when it was left unstimulated. If it was detectable, we further tested whether the *M.tb*-lysate stimulated samples of this subset was significantly higher than when left unstimulated. If the lower bound of the 95% CI around the median for the *M.tb*-lysate samples was greater than

the upper 95% CI of the median for the unstimulated samples, the *M.tb*-lysate stimulated samples were kept. If the lower bound of the 95% CI around the median of the *M.tb*-lysate stimulated sample overlapped with the upper 95% CI of the unstimulated sample, the *M.tb*-lysate stimulated samples were removed and only values of the unstimulated samples were kept. Our rationale was that if the *M.tb*-lysate and unstimulated values were the same, then the biological responses were not different and hence it was unnecessary to keep both versions. An example of a response that was different for *M.tb*-lysate and unstimulated samples is shown in Figure 3.10.

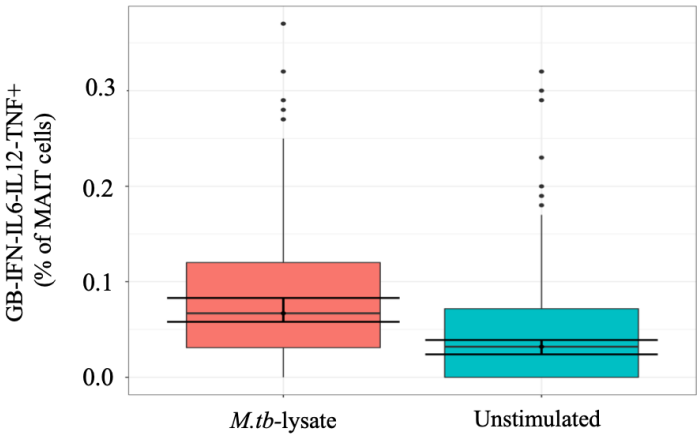

Figure 3.10: Illustrating the innaptive filtering protocol. Figure shows a boxplot of the raw values for GB-IFN-γ-IL-6-IL-12-TNF+ MAIT cells under two stimulations. The black bars represent the 95% CI around the median (the dark grey line). The lower 95% CI for the *M.tb* subset is larger than the upper 95% for the unstimulated subset. According to our filtering method, both versions of this cell subset would be kept in the final model. In addition, we would perform background subtraction on the *M.tb* stimulated version (GB- subset).

If the unstimulated version of a cell subset was found to be undetectable, we tested whether the *M.tb* -lysate stimulated version was considered detectable. If detectable, then the *M.tb*-lysate version was kept, otherwise we discarded both versions.

Lastly, for *M.tb*-lysate-specific subsets that were preserved post-filtering, we performed background subtraction, as done for T cells (Section 3.2.1). The exception for this was when the cell subset was GB positive. As GB is a cytokine that is constantly present in cells and not only expressed after stimulation, it is not meaningful to perform background subtraction.

3.5 Data standardization

Due to the intrinsic biological variability of the measurements in the datasets, another pre-processing step included standardizing the data to a common scale. The risk with some scaling methods, however, is that values may become inflated. Let X be a measured variable where x_i is the i^{th} $(i = 1, \ldots, n)$ observation in X and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean of the n observations in X. Further, let $\sigma_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}^2)$ be the sample variance of X. An example of a scaling method that has the potential to inflate results would be the Z-score function $(\frac{x_i - \bar{x}}{\sigma_X})$. X will now have a mean of zero and a standard deviation of one, however, some values in X may have been increased in order to achieve this variance of one. The values in our dataset were already noticeably small and could easily be inflated, which could have lead to false positive results in the predictive models.

A number of standardization techniques, data transformation methods, and a combination of

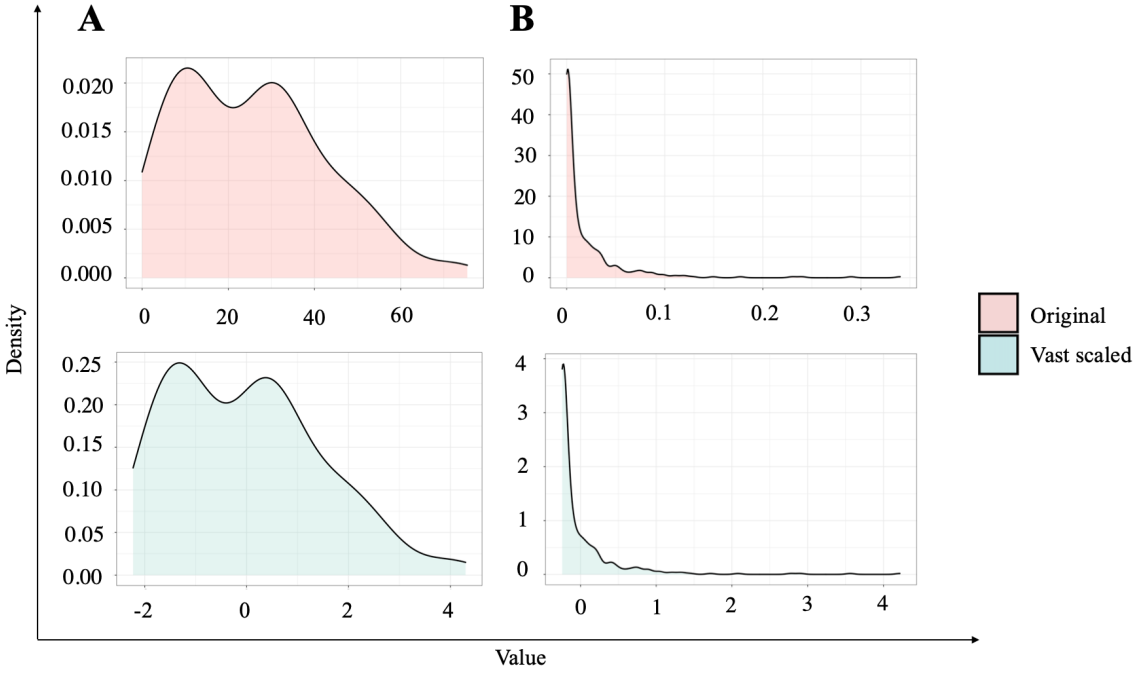

Figure 3.11: Vast scaling preserved the densities of the raw data. The figure compares the densities of the original and vast scaled data values for (**A**) a variable from the innaptive dataset, *M.tb*-specific IL-6+TNF+ monocyte, and (**B**) a variable from the adaptive dataset, E6C10-specific CD4+IL-2+CD107-CD154+IFN-γ+TNF-. The data ranges of the two variables are more comparable after vast standardization (x-axis of blue plots) compared to the raw values (x-axis of red plots), but the densities of the original variables are still preserved.

the two, were tested on the data. A scaling method known as vast (variance stabilizing) standardization [75] performed the best and scaled all variables to a similar, more comparable range, while still preserving the original density of the raw data (Figure 3.11). Vast standardization involves dividing the Z-score by a coefficient of variation (cv) as scaling factor

$$\tilde{x}_i = \frac{(x_i - \bar{x})}{\sigma_X} \frac{\bar{x}}{\sigma_X}.$$

Division by this scaling factor, which is the sample standard deviation divided by the sample mean, gives higher importance to those variables with small relative standard deviations. Vast scaling aims to be robust and is commonly used on metabolites that show small fluctuations [133].

3.6 Missing data imputation

Each dataset was aligned according to PID, QFT status and month of sample collection. Several participants had to be removed from the innaptive dataset as they did not have enough viable cells available, therefore only $n = 16$ of the $n = 29$ converters, $n = 13$ of the $n = 30$ reverters and $n = 34$ out of the $n = 60$ participants from the two control groups were analyzed. As not all variables were measured for each individual at each time point, row-wise missingness (missing values) arose in the final dataset (Figure 3.12) as a consequence of this integration step. Since the missingness

arose through this data integration step, we have good reason to believe that the data was missing completely at random (MCAR).

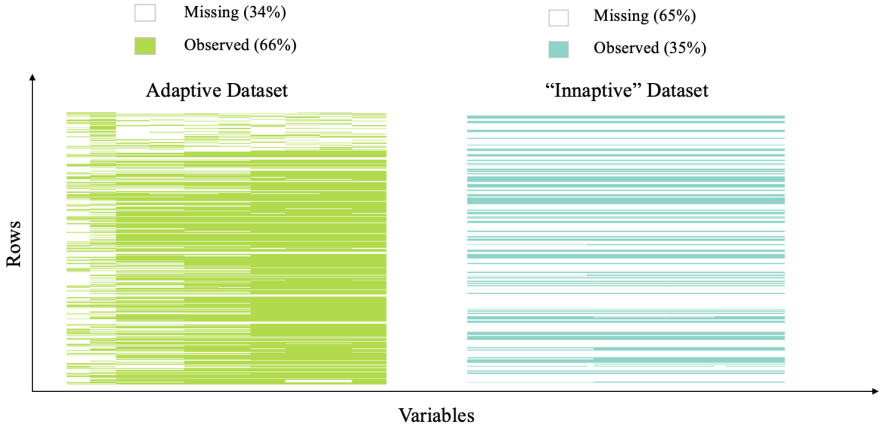

Figure 3.12: Missingness patterns in the two datasets. Figure was generated using the R package Amelia II [65] where a white block indicates a missing value.

Most statistical methods require the dataset to be complete, so missing values had to be meaningfully replaced in a process known as imputation. In a similar manner to the data standardization step, a number of imputation methods were implemented and their performances compared. A good imputation method is one where the imputed data (replacement of missing values) can successfully replicate the density of the raw, incomplete data, and not introduce new structure into the data. The specific imputation methods that were tested are summarized in Table 3.1.

Table 3.1: The various imputation methods that were tested.

Method	Explanation	Author(s)	R package
missForest	A non-parametric imputation method that uses Random Forest models to predict the missing values and can handle mixed data	Stekhoven et al. (2012) [123]	missForest [122]
k nearest neighbours (KNN)	An imputation method that employs the KNN algorithm to group variables with similar profiles. A weighted average of the K nearest complete variables is taken and used to impute the missing values in the incomplete variables	Troyanskaya et al. (2001) [132]	DMwR [129]
Column median	Missing values were simply imputed with the column median of the incomplete variable		
Multiple Factor Analysis (MFA)	Builds an MFA model on the incomplete dataset and uses the model to predict the missing values.	Husson et al. (2013) [68]	missMDA [69]

Each imputation method tested can cope with multivariate data, and missingness that arises

for any reason, but because we believe the data is MCAR, we had no reason to have any other aim but to see how well the imputation method replicated the raw data. We found the Multiple Factor Analysis (MFA) imputation method to outperform all other methods (Figure 3.13). MFA imputation was recommended in a paper by Voillet and colleagues in 2016 [135] for effectively imputing data with row-wise missingness.

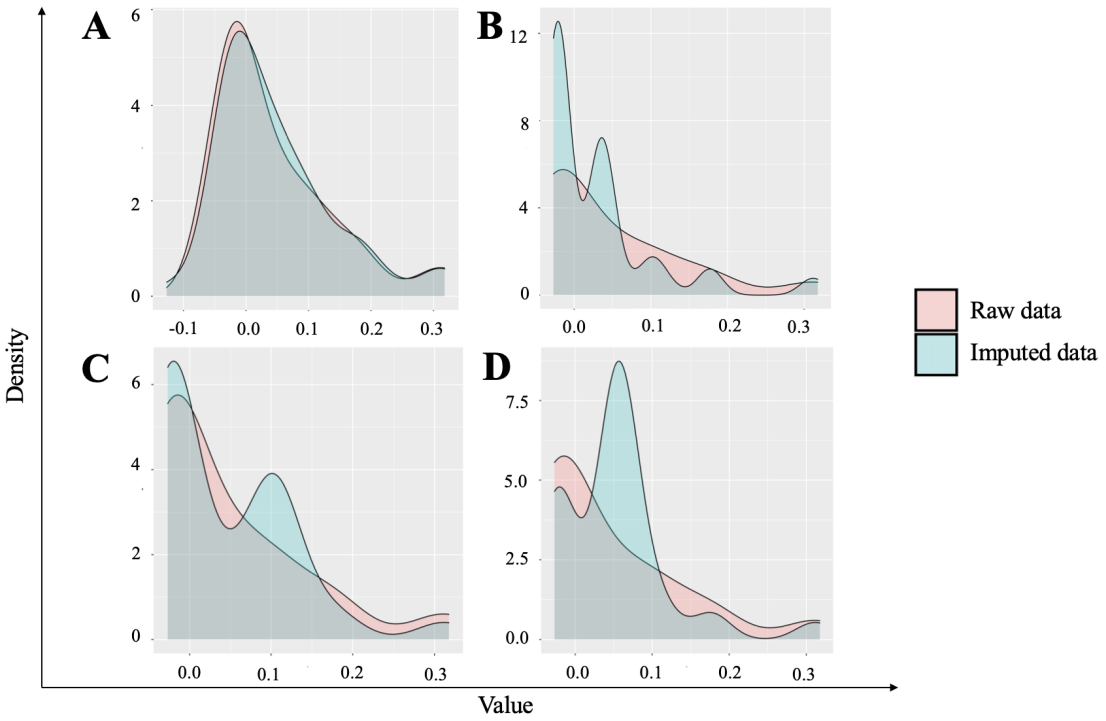

Figure 3.13: The efficacy of each imputation method to capture the distribution of the raw data frequencies. Figure shows total Granzyme B production in CD19+ B cell stimulated with *M.tb*. The red lines are the raw data in each plot and the blue lines are (**A**) MFA, (**B**) column median, (**C**) k-nearest neighbours and (**D**) missForest imputed values. The MFA imputation method managed to capture the distribution of the raw data almost perfectly.

MFA is one of the many data reduction techniques available and is typically used for data structured in groups of variables [25]. Similar to PCA, MFA aims to provide a subspace in a lower dimension that best represents the data in order to maximize the variability of the projected points. Therefore MFA is a useful approach as it sets out to study the similarities between rows, from a multidimensional point of view, as well as the correlation between variables and highlights the similarities and differences between groups of variables. One of the aims of MFA is to ensure the fair contribution of each group of variables in the analysis, such that no single group influences the first dimension of variability.

Let $X = [X_1, \ldots, X_K]$ be a data matrix that is made up of K data tables measured on the same I observations, and D_Σ (KxK) to be a diagonal matrix with each element equal to one. We can then construct $XD_\Sigma^{-\frac{1}{2}}$, which is centered to give $Z = XD_\Sigma^{-\frac{1}{2}} - M$, where M (IxK) $= 1m'$ is the matrix of average values of X. The first step of MFA involves performing PCA on each group of variables separately (PCA method outlined in Chapter 6 Section 6.2.3). The K data tables are then weighted according to the first eigenvalue, λ_1^k of the PCA. This weighting step balances the influence of the groups of variables. The second step of MFA then involves taking a global PCA

Algorithm 1: MFA imputation algorithm adapted from [68]

Result: An imputed data matrix X with all missing values meaningfully replaced using the MFA algorithm.

$l \leftarrow 0$;

Replace missing values in X^0 with the mean of the observed variable and calculate D_Σ^0 and $\hat{M}^0$;

The first eigenvalue of each group of variables is computed and $\Lambda^0 = (\lambda_1, \ldots, \lambda_k)^0$;

repeat

$l \leftarrow l + 1$;

1. Construct the weighted matrix $Z^{l-1}(\Lambda^{-\frac{1}{2}})^{l-1} = (X^{l-1}(D_\Sigma^{-\frac{1}{2}})^{l-1} - \hat{M}^{l-1})(\Lambda^{-\frac{1}{2}})^{l-1}$;
2. Perform PCA and use results to estimate $\hat{F}$ and $\hat{U}$; S dimensions are kept;
3. Estimate the fitted values $\hat{Z}^l = \hat{F}^l(\hat{U}^l)'(\Lambda^{\frac{1}{2}})^{l-1}$;
4. Impute missing values with $\hat{X}^l = (\hat{M}^{l-1} + \hat{Z}^l)(D_\Sigma^{-\frac{1}{2}})^{l-1}$;
5. The l-th imputed dataset is then $X^l = W * X + (1 - W) * \hat{X}$;
6. Using X^l, calculate D_Σ^l, $\hat{M}^l$ and Λ^l.

until;

Convergence.

across this now weighted data matrix $Z\Lambda^{-\frac{1}{2}} = [Z_1/\sqrt{\lambda_1^1}, \ldots, Z_k/\sqrt{\lambda_1^k}]$

In 2013, Husson and colleagues [68] developed a regularized, iterative MFA algorithm that used the MFA algorithm to perform imputation. In terms of notations, let $\hat{F}$ be the estimated principal score after performing PCA, and $\hat{U}$ be a matrix of loadings. Further define a matrix W with $W_{ik} = 0$ if X_{ik} is missing and $W_{ik} = 1$ otherwise. For the l^{th} iteration of the algorithm (Algorithm 1) the missing values are essentially replaced with the sum of the estimated fitted values after performing PCA ($\hat{Z}^l$) and the average values of X from the previous iteration ($\hat{M}^{l-1}$).

The imputation method was performed within the missMDA package in R [69] and we defined the groups as the variables that came from the different datasets. Most of the analysis within this dissertation was on the form of multivariate plots, and, when modeling approaches were used, we were less focused on estimating the coefficients but rather on identifying the predictive markers. Therefore, we did not use Rubin's rules [109] to take into account between and within-imputation variability. For Chapter 5 and 6 in particular, the MFA imputation was rather repeated for every cross validation (CV; Section 5.2.2) run during parameter tuning or model building to ensure that any results found were not a consequence of the imputation method used.

3.7 Discussion

This chapter outlined the many pre-processing steps that were required to successfully integrate the datasets from the adaptive and innaptive immune responses. In summary, the two datasets were not generated consistently, which made integration and analysis a challenge. Median values of the longitudinal time points therefore had to be taken for each for each individual, stratified according to cohort and QFT result. Non-parametric tests indicated that there were no significant

differences between the time points where median values were taken. However, the inconsistency in the datasets was a major limitation in this project, as we could not make use of all the data points available or explore the longitudinal nature of the data.

To overcome the high dimensionality of the dataset post-integration, pre-filtering methods were applied to each dataset separately. For the adaptive dataset, COMPASS was applied to identify antigen-specific responses that were higher than background. Binary combinations of cytokines that did not satisfy at least 10 observations with posterior probabilities greater than 0.1 at either month 0, 6, 12 or 18 were omitted. MIMOSA was also used to identify responders, where phenotypic markers on total Th1 cells were only measured on responding individuals. For the innaptive dataset, due to the high unstimulated values, we could not employ COMPASS and hence designed a filtering method that would identify the biologically meaningful cell subsets in this dataset. After these pre-filtering steps were applied in both the datasets, we were satisfied that all biologically irrelevant cell subsets had been identified and then removed from the final dataset. The final dataset included 123 variables from the adaptive dataset and 70 variables from the innaptive dataset. The intrinsic biological variability between the two datasets was then accounted for by standardizing the datasets to a common scale using vast scaling, and missing values were meaningfully imputed using an MFA-based imputation method.

3.8 Conclusion

In order to successfully integrate the data, several data pre-processing steps were required. Processing of data risks influencing and potentially biasing any results found in the final integrated model. This chapter emphasizes the importance of testing different approaches to identify the best suited method for a given dataset such that unbiased and valid results are yielded.

Chapter 4

Defining the features of the immune response that change during recent acquisition of *Mycobacterium tuberculosis* infection

AIM 1: To identify features of the immune response that change upon QFT conversion. *We hypothesize that upon QFT conversion, frequencies of IFN-γ+ TNF+ IL-2+ M.tb-specific CD4+ T cells are higher compared to pre-conversion time-points.*

4.1 Introduction

Acquisition of *M.tb* infection is generally asymptomatic and tends to remain undiagnosed unless serial diagnostic tests are performed in exposed populations. As a result, very little is known about the immune response induced during primary *M.tb* infection in humans. This chapter describes exploratory analyses that were used to identify and define immune features that change upon infection with *M.tb*. The following chapter, Chapter 5, aimed to identify a small set of biomarkers of recent TBI, which could help targeting individuals requiring preventative TB treatment.

The high-dimensionality of the data ruled out the use of multiple hypothesis tests to compare pre- and post conversions values as the preferred method. Therefore, we explored alternative methods to cluster longitudinal data. Clustering algorithms, unlike classification algorithms, are unsupervised methods that aim to split a large dataset into a plurality of clusters, which share some traits. Clustering algorithms typically involve calculating a similarity or proximity matrix based on the distance between data points, and can be broadly divided into hierarchical and non-hierarchical clustering in the data. Hierarchical clustering can be agglomerative (bottom-up), where all data points initially belong to their own cluster, which are then merged in an iterative fashion, or divisive (top-down), where all data points initially belong to the same cluster, which is then split in an iterative fashion. Non-hierarchical clustering algorithms, on the other hand, do not follow a tree-like structure. A popular non-hierarchical clustering algorithm is the k-means algorithm [83]. K-means is an iterative algorithm that aims to partition a group of observations into k pre-defined, non-overlapping groups or clusters. The clusters are formed such that observations in the same

cluster are similar to each other and observations in different clusters are different from each other.

The k-means algorithm, however, cannot easily be extended to longitudinal data. This is intuitive as the metrics used by k-means, which are a distance metric, typically taken to be the Euclidean distance, and the arithmetic mean, are not shape-respecting tools. Genolini and colleagues [52] hence introduced the kmlShape algorithm. kmlShape aims to cluster longitudinal trajectories based on their shapes and extends the k-means algorithm by employing metrics that are suitable for longitudinal data. Specifically, the Fréchet distance and Fréchet mean [48]. The Fréchet distance is a shape-respecting distance that is small if two trajectories have similar shapes, and the Fréchet mean is, informally, the middle of two trajectories. This algorithm was applied to cluster the longitudinal trajectories of all four time points available in the recent converter cohort based on their shapes. The assumption was that all variables that had an increasing trend over time would cluster together, and, similarly, those with a decreasing trend would cluster. We could then extract which variables fell into these clusters. Based on literature and knowledge of T cell differentiation, we hypothesized that, upon *M.tb* infection, levels of polyfunctional (IL-2+IFN-γ+TNF+) *M.tb*-specific CD4+ T cells would increase.

We further aimed to explore how the cell subsets from the different datasets may interact with each other, and how those interactions changed as a result of infection with *M.tb*. Using the variables that changed during *M.tb* infection, whether they increased or decreased, correlation networks were constructed for pre- and post- conversion time points separately. Correlation networks fall under a larger field of statistics known as network analysis. Network analysis focuses on modeling complex systems as a network of graphs and places a greater emphasis on the relationships between individual entities, rather than analyzing them as separate, isolated events. A correlation network therefore is a network where the relationships between individual entities are based on their correlations. Using a correlation network to model the immune response to *M.tb* infection, which is, by its nature, a complex, multi-level interaction network, would enable numerous connections and relationships between the cell types to be visualized in a clear and structured way. We could then draw conclusions about the relationships and interactions that formed between the cells based on the network topology. A separate correlation network was also built and used as a statistical validation of, or an alternative to, the kmlShape algorithm to identify the cell subsets that changed as a result of TBI.

4.2 Methods

4.2.1 Study design

All analyses in this chapter were performed on the recent QFT converter cohort (Figure 4.1) to infer the immune response change upon recent *M.tb* infection. With the exception of phenotypic marker expression on E6C10-specific total Th1 cells, all variables that passed the various filtering criterion were included in the analyses. We did not include the phenotypic data because very few participants had detectable E6C10 responses prior to QFT conversion. Vast scaling was then applied to standardize the dataset and the missing values were not accounted for, as no missing values were present after calculating the median "variable trajectories" (Section 4.2.2), and only pairwise complete correlations were taken when generating the correlation networks.

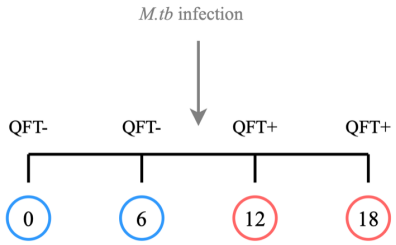

Figure 4.1: The study cohort for Aim 1. The recent converter cohort was defined by two negative QFT time points followed by two positive QFT time points, assuming that *M.tb* infection happened between month 6 and 12.

4.2.2 Variable trajectories

The kmlShape algorithm aims to cluster longitudinal trajectories based on their shapes. As we were interested in which variables changed over time, rather than which individuals changed over time, median "variable trajectories" were used as an input instead. A variable's trajectory was found by taking the median value of the 29 individual samples in the recent converters, at each time point separately (Figure 4.2). This was done on the vast scaled but un-imputed integrated dataset.

Define the median of an ordered sequence x of length n to be

$$\text{median}(x) = \begin{cases} x[\frac{n+1}{2}], & \text{if } n \text{ is odd} \\ \frac{x[\frac{n}{2}]+x[\frac{n+2}{2}]}{2}, & \text{if } n \text{ is even.} \end{cases}$$

If x_{ikj} is the value for participant i at time $k \in \{0, 6, 12, 18\}$ for variable j, then $\mathbf{x}_{kj}$ is a vector containing all participant's values at time k for variable j. The variable trajectory (VT) for variable j is then defined by

$$VT_j = \left(\begin{pmatrix} 0 \\ \text{median}(\mathbf{x}_{0j}) \end{pmatrix}, \begin{pmatrix} 6 \\ \text{median}(\mathbf{x}_{6j}) \end{pmatrix}, \begin{pmatrix} 12 \\ \text{median}(\mathbf{x}_{12j}) \end{pmatrix}, \begin{pmatrix} 18 \\ \text{median}(\mathbf{x}_{18j}) \end{pmatrix} \right).$$

Therefore, each variable was summarized by four longitudinal time points.

4.2.3 K-means to cluster longitudinal trajectories

The standard k-means algorithm selects k distinct data points at random, which form the initial clusters. The distances between every data point and the initial clusters are measured and the points are assigned to their nearest cluster. Therefore, a suitable distance metric is required. For two points $(x_1, y_1)^T$ and $(x_2, y_2)^T$, the Minkowski distance between these two points is defined as

$$d\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \right) = [|x_1 - x_2|^m + |y_1 - y_2|^m]^{1/m} \tag{4.1}$$

for $m > 0$. When $m = 2$, this becomes the Euclidean distance, and when $m = 1$, the Manhattan distance. Typically, the Euclidean distance is used as the distance metric

$$d\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \right) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{4.2}$$

The next step of the k-means algorithm involves computing the centroid of each cluster using the arithmetic mean. The centroid is now used to represent each cluster and the distances between

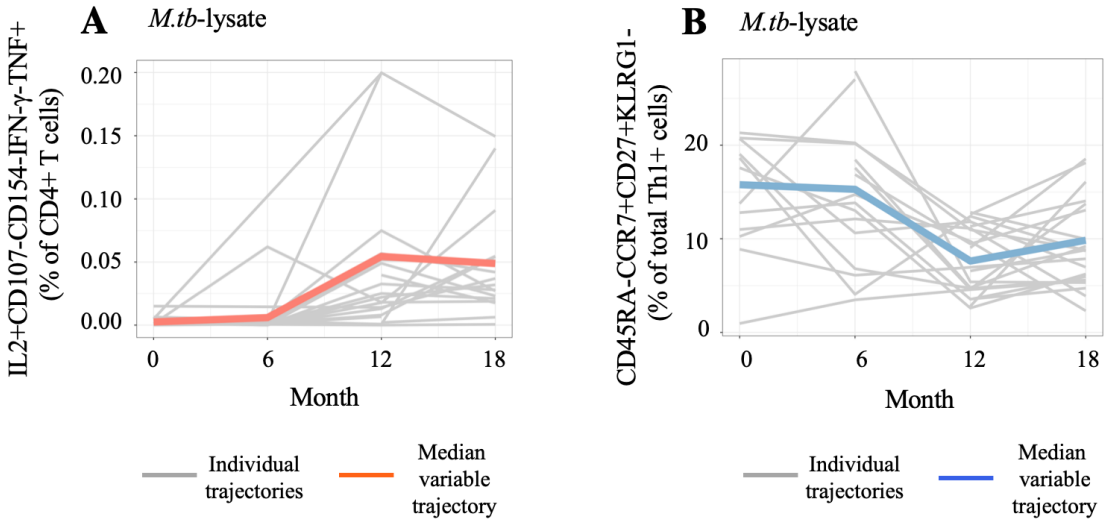

Figure 4.2: Median "variable trajectories". These were defined by taking the median of the individual values in the recent converters at each sampling occasion to determine the general trend over time for each variable. For example, (**A**) *M.tb*-lysate-specific CD4+IL-2+CD107-CD154-IFN-γ+TNF+ T cells have an increasing trend over time (red) and (**B**) *M.tb*-lysate-specific CD45RA-CCR7+CD27+KLRG1- expression on total Th1 cells have a decreasing trend over time (blue).

every data point and the centroids are compared. The two steps are repeated until there is no change to the centroids. The entire k-means algorithm is also repeated several times to account for the initial randomness of choosing which data points make up the first clusters. The clustering that results in the smallest total variation within the clusters is then chosen. The total variation within clusters can also be plotted as a scree plot to determine the optimal value for k in a given dataset.

kmlShape extends the k-means algorithm by employing metrics that are suitable for longitudinal data, specifically the Fréchet distance and Fréchet mean [48]. Note that the kmlShape algorithm has an additional parameterization that can account for different time-scales when comparing trajectories. Since each variable trajectory was measured on the same four time points, this parameterization was not applied and will not be discussed.

The formula to calculate the Fréchet distance between two trajectories P and Q is

$$D_{\text{Frechet}}(P,Q) = \inf_{f,g} \max_{t\in[0,1]} \{d(P(f(t)), Q(g(t))\}$$

where $d()$ is the Euclidean distance in (4.2). To develop the intuition behind this method, I draw your attention to Figure 4.3A. Suppose we have two trajectories, P and Q, measured on the same x values, $0, \ldots, x_n$, and we want to find the distance between these two trajectories. One way of doing this would be to measure the Euclidean distance between the two curves at each shared x-value. For example, the maximum distance between P and Q at x_1 would be the black arrow in Figure 4.3A. However, we notice that we can find another distance between P and Q, represented by the red arrow in the figure, by re-parameterizing the x-values. The Fréchet distance method therefore introduces two continuous, monotonically non-decreasing functions, $f(t)$ and $g(t)$, which map values from $[0,1]$ to $[0, x_n]$, such that $f(0) = g(0) = 0$, and $f(1) = g(1) = x_n$. Hence, returning to the example in Figure 4.3A, for some $t^* \in [0,1]$, $x_2 = f(t^*)$ and $x_1 = g(t^*)$, we can find a different maximum Euclidean distance between P and Q at t^*.

In practice, several maximum distances will exist after re-parameterization (Figure 4.3B), and the Fréchet distance is therefore defined as the smallest possible (infimum) maximum distance between P and Q, after re-parameterization.

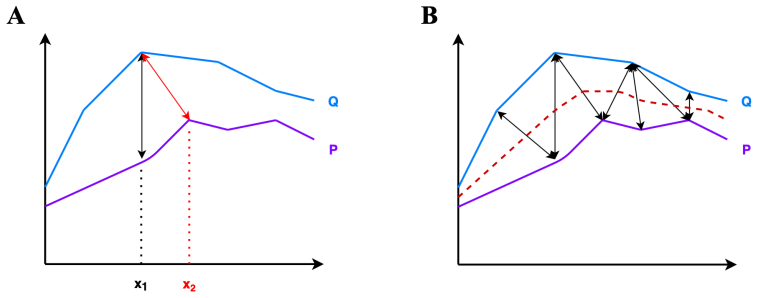

Figure 4.3: Definition of the Fréchet distance and Fréchet mean. Two trajectories, labelled P and Q, are plotted in purple and blue respectively. (**A**) This figure is used to develop the intuition behind the re-parameterization in calculating the Fréchet distance. The black arrow represents the maximum Euclidean distance between P and Q at x_1, while the red arrow represents the maximum distance after re-parameterization. (**B**) The black arrows represent the numerous maximum distances between P and Q after reparameterization, and the Fréchet distance is defined as the smallest maximum distance between P and Q. The Fréchet mean trajectory (orange dotted line) is found by taking the mean of each line segment linking P and Q for the re-parameterizations that minimized the Fréchet distance.

The other shape respecting measure that is employed by kmlShape is the Fréchet mean. Informally, the Fréchet mean is middle of the segments linking P to Q (Figure 4.3B). Given that f and g are the re-parameterizations that minimize the Fréchet distance, the Fréchet mean is the average distances between the points of the trajectories when run at a speed defined by f and g:

$$\mu_{\text{Frechet}}(P, Q) = \left(\frac{P(f) + Q(g)}{2} \right).$$

The algorithm then follows the same protocol as the k-means algorithm, using the Fréchet distance and Fréchet mean to iteratively cluster trajectories that have similar shapes until convergence. The full algorithm is outlined below in "Algorithm 2".

We applied the kmlShape algorithm to the variable trajectories defined in the previous section. The value of k was chosen to be three based on the following expected longitudinal trends: increasing, decreasing, or remaining constant over time. The method was run on the median variable trajectories and implemented in R package kmlShape [53]. We could then identify the variables that fell into each cluster.

Algorithm 2: kmlShape algorithm adapted from [52]

Population: n variables $V_1, \ldots, V_n$.
Result: Partition trajectories into k clusters: Cluster vector of size n taking values in $[1, \ldots, k]$.
k variables, $C_1, \ldots, C_k$ are chosen at random from $V_1, \ldots, V_n$;
$s \leftarrow 0$;
$\text{Cluster}_0 \leftarrow (0, 0, \ldots, 0)$ ▷ a vector of size n;
repeat
  $s \leftarrow s + 1$;
  for i *in* $1, \ldots, n$ **do**
    for j *in* $1, \ldots, k$ **do**
      Compute Fréchet distance between V_i and C_j ($\text{Dist}F_{i,j}$);
      $\text{Cluster}_s(\text{i}) \leftarrow j$ such that $\text{Dist}F_{i,j} < \text{Dist}F_{i,j'}$, for $j' \neq j$;
    end
  end
  for j *in* $1, \ldots, k$ **do**
    Compute the Fréchet mean, M_j, of cluster j (all V_i such that $\text{Cluster}_s(\text{i})==j$);
  end
until;
$\text{Cluster}_s == \text{Cluster}_{s-1}$ or $s >$ maximum iteration

4.2.4 Correlation networks

A network is made up of a series of individual entities and the connections between them. These entities are referred to as vertices or nodes of the graph and the connections are called edges (Figure 4.4A). A finite network can be summarized in a square matrix, known as the adjacency matrix where the rows and column names of the matrix are the nodes, and the elements of the matrix indicate whether pairs of nodes are adjacent in the graph or or not. Adjacency matrices are typically used as an input for network analyses.

An alternative to the adjacency matrix is an edge list. An edge list is a data frame made up of two columns, one containing the source nodes, and another with the target nodes. If the distinction between source and target is meaningful, then the network is said to be directed (Figure 4.4B), otherwise it is undirected. An example of a directed network would be a network built that tracks postcards sent between cities. This is directed because the distinction between source and target is relevant. An additional column can be appended to an edge list, which assigns weights to edges in the network, creating a weighted graph. Weighted networks can loosely be defined as soft-thresholding the pairwise associations. If no weights, or equal weights, are assigned to the edges in the graph, the network is unweighted.

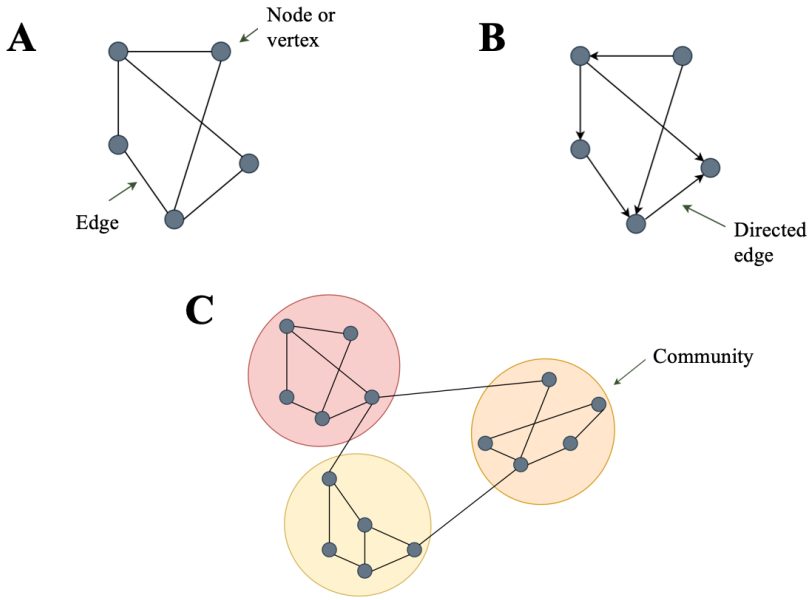

Figure 4.4: Network analysis notation. (**A**) shows a simple, undirected network made up of entities (nodes) and their relationships (edges), (**B**) shows a directed network, where the distinction between the source and target node is important, and (**C**) shows community structures, made up of dense sub-graphs within a network.

A correlation graph is therefore a network where the adjacency matrix is predicted from pair-wise correlations between entities [10]. We specifically used Spearman's rank, or rho (ρ), correlation coefficient [120] to generate the correlation matrix. Spearman's correlation is a non-parametric test that measures the strength and direction of the association between two ranked variables. Each pair of variables is ranked separately, such that the largest value is assigned a rank of n and the smallest a rank of one. Similar to Wilcoxon's signed rank test, tied scores are given the average rank. The difference, d_i, is calculated between the ranks of the two variables and the correlation coefficient is given by

$$\rho = 1 - \left(\frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \right)$$

where $-1 \leq \rho \leq +1$. Intuitively, the correlation coefficient will be equal to positive one if the ranks of two variables are identical, and negative one if they are completely opposite. Therefore, values of ρ closer to one indicate a stronger association between two variables.

Let X be an n x p matrix, with n rows and p variables. A correlation matrix is then a square and symmetric p x p matrix where the elements of the matrix hold the pair-wise correlations between each of the p variables:

$$\begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \dots & \rho_{1,p} \\ \rho_{2,1} & 1 & \rho_{2,3} & \dots & \rho_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{p,1} & \rho_{p,2} & \rho_{p,3} & \dots & 1 \end{bmatrix}$$

where $\rho_{i,j} = \rho_{j,i}$ is the Spearman's correlation coefficient between variable i and j and $\rho_{i,i} = 1$.

The kmlShape algorithm was used to identify all the variables that changed during infection with *M.tb*, *i.e.* the variables that either had an increasing or decreasing trend over time. Of the variables that changed, we only included cells stimulated with *M.tb*-lysate or unstimulated cells from

the innaptive dataset. We restricted the network to include these cells only because *M.tb*-lysate was the common stimulation between the two datasets. Spearman's correlation coefficient was used to calculate the correlations between these variables and generate the correlation matrices for pre- and post-conversion time points separately. By doing this we could infer how the relationships or interactions between cell subsets change as a result of infection with *M.tb*. The p-values were then adjusted using the BH [14] correction.

The R package igraph [33] was used to build two undirected and unweighted networks with the pre- and post-conversion correlation adjacency correlation matrices as inputs. Associations between cells with correlations greater than 0.5 and BH adjusted p-values less than 0.05 were included in the network. The edges connecting the nodes in the graph were shaded according to strength of the correlation between nodes, and the nodes were coloured according to cell type.

A correlation network was also used as a statistical validation of the kmlShape algorithm. Using the same median variable trajectories that were used for the kmlShape algorithm from the integrated dataset, a correlation network was constructed once again using Spearman's correlation to generate the correlation matrix. The multiple time points of the same variable were treated as independent observations. The idea was that variables that had an increasing trend would have higher correlations and cluster together, and these variables would have negative correlations with those variables that have a decreasing trend. Only associations with an absolute correlation coefficient greater than 0.5 ($\rho > |0.5|$) were included in the network. A community detection algorithm was then applied to identify sub-graphs within the network. Specifically, the optimal community detection algorithm was used, which is a function built-into the igraph package. The algorithm calculates the optimal community structure of a specific network by maximizing a measure known as the modularity across all possible partitions. The modularity of a graph measures how separated different vertex types are from each other. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules (Figure 4.4C).

Define $G = (V, E)$ to be a simple, undirected network with $n = |V|$ vertices and $m = |E|$ edges. Let A_{ij} be the weight of the edge between vertices i and j (from the adjacency matrix), k_i be the sum of weights of the vertex attached to vertex i, and c_i be the community to which vertex i is assigned. Modularity Q is then defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

where $\delta(c_i, c_j) = 1$ if $c_i = c_j$ and 0 otherwise. Suppose we begin by dividing the network into two parts. Define the scalar s_i that will take the value 1 if vertex i belongs to cluster 1 and -1 if it belongs to cluster 2. Then $\frac{1}{2}(s_i s_j + 1)$ will take the value 1 if i and j are in the same group, and 0 otherwise. By letting $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$, we can then update the above equation to give

$$Q = \frac{1}{4m} \sum_{i,j} B_{ij} s_i s_j = \frac{1}{4m} \mathbf{s}^T \mathbf{B} \mathbf{s}$$

The goal is to then find a membership vector $\mathbf{s}$ that maximizes Q for a given $\mathbf{B}$, which can be cast as an integer linear program [20].

The optimal community detection algorithm was used to identify subgraphs within the network. Variables that belonged to each subgraph or community were identified and their associated variable trajectories were plotted. From this we could decipher whether the algorithm successfully grouped variables with similar trends over time, and whether it could be an alternative method to cluster longitudinal data.

4.2.5 Wilcoxon's signed rank test

To further explore the findings from kmlShape and the community detection algorithm in the correlation network, non-parametric tests were used to compare pre- and post-conversion time points (Figure 4.5) in the variables that were identified to have either increased or decreased over time. Because there were no significant differences between the two QFT- time points, nor between the two QFT+ time points, we were able to justify taking the median values of these time points without risking any loss of information (Section 3.3; Figure 3.6). Specifically, Wilcoxon's signed rank test [138] was used to compare these two time points, as it has no distributional assumptions about the data. The null hypothesis of the test is that the median difference between two paired samples is zero. Hence there are no differences between the paired samples.

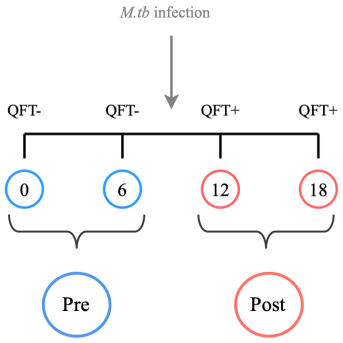

Figure 4.5: Pre- and post-conversion time points were aggregated by taking the median of the two negative QFT time points and the two positive QFT time points, respectively.

Wilcoxon's signed rank test begins by calculating the difference between each sample pair, D_i, and temporarily ignoring the sign by taking its absolute value, $|D_i|$. All observations such that $|D_i| = 0$ are omitted and the sample size reduces to $n \leq n'$, where n is the total number of non-zero absolute differences. Ranks are assigned to the remaining paired observations such that $\min(|D_i|) = 1$ and $\max(|D_i|) = n$. For instances when there are two or more equal absolute differences, the rank assigned is the average rank value across the tied data. The original signs are then re-introduced and W^+ is defined as the sum of the positive ranks and W^- the sum of the negative ranks, such that $W^+ + W^- = \frac{n(n+1)}{2}$. The test statistic, W, is taken to be the maximum of W^+ and W^-.

For a large sample size, the test can be approximated by a normal distribution, where $\mu_W = \frac{n(n+1)}{4}$ and $\sigma_W^2 = \frac{n(n+1)(2n+1)}{24}$. If ties are present in the data, σ_W is reduced by $r(t) = \frac{t^3-t}{48}$, where t is the number of tied ranks. Then

$$Z = \frac{W - \mu_W}{\sqrt{\sigma_W - r(t)}}$$

and the null hypothesis is rejected at a 5% level of significance if $Z \geq 1.96$.

4.3 Results

4.3.1 kmlShape clustered variables according to three longitudinal trends over time

With the number of clusters to identify in the data, k, set to three, the kmlShape algorithm was applied to the 193 vast scaled median variables trajectories, and the output from the algorithm is

shown in Figure 4.6. The majority of the variables (68%) fell into the first cluster (red), which grouped the variables that did not have any specific trend, or did not change over time. Variables with an increasing trend were then grouped into the second cluster (green), which included 47 variables (27% of the total number of variables). The last cluster (blue) grouped variables with a decreasing trend over time. Eight variables fell into this cluster (5% of the total number of variables). The algorithm therefore identified a total of 55 variables that changed over time.

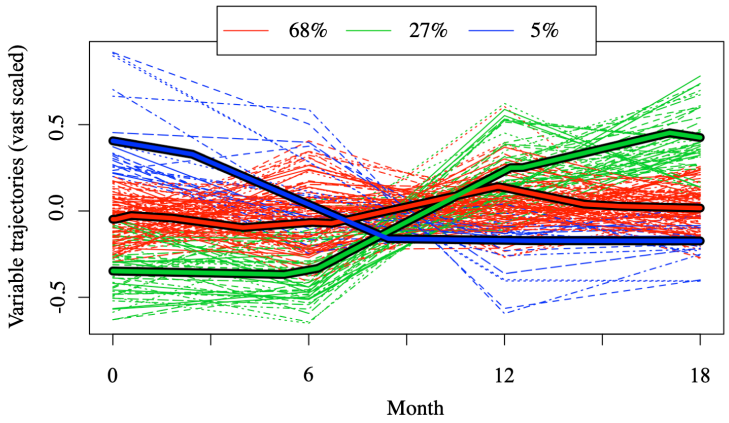

Figure 4.6: The three clusters identified by the kmlShape algorithm. Variables are colour-coded depending on their trajectory: no change over time (red), an increasing trend (green), and a decreasing trend over time (blue).

The variables that either increased or decreased over time are summarized in Table 4.1 and 4.2, respectively. The majority of the cell subsets that increased were *M.tb*-specific CD4+ T cells, and included polyfunctional CD4+ T cells (Figure 4.7). All cells subsets that decreased were proportions of *M.tb*-lysate-specific total Th1 cells that expressed phenotypic markers combinations mostly representative of early differentiated T cells (Figure 2.4).

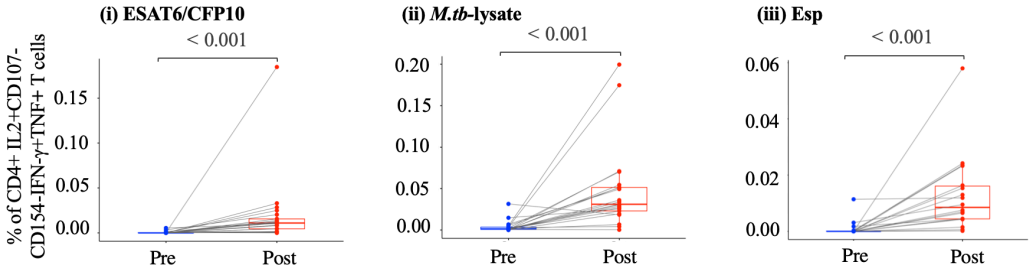

Figure 4.7: Differences between pre- and post-conversion time points in polyfunctional CD4+ T cells. The plots shows significant differences ($p < 0.05$) in polyfunctional CD4+ T cell responses when stimulated with (**i**) E6C10, (**ii**) *M.tb*-lysate or (**iii**) Esp. Wilcoxon paired tests were used to compare the two time points and the resulting p-values are superimposed onto the plots.

Table 4.1: A summary of the cells that had an increasing trend identified by kmlShape. The variables coloured in red are CD4+IL-2+CD107-CD154-IFN-γ+TNF+ T cells stimulated with Esp, *M.tb*-lysate and E6C10; the variables that were hypothesized to have increased post-infection.

Cell Type	**Stimulation**	**Effector Function/ Phenotypic Markers**
CD3+ T cells	*M.tb*-lysate	Total IFN-γ
		Total TNF
		GB-IFN-γ-IL-6-TNF+
		GB-IFN-γ+IL-6-TNF+
CD4+ T cells	Esp	IL-2-CD107-CD154-IFN-γ+TNF-
		IL-2-CD107-CD154-IFN-γ+TNF+
		IL-2-CD107-CD154+IFN-γ+TNF+
		IL-2+CD107-CD154-IFN-γ-TNF+
		IL-2+CD107-CD154-IFN-γ+TNF-
		IL-2+CD107-CD154-IFN-γ+TNF+
		IL-2+CD107-CD154+IFN-γ-TNF+
		IL-2+CD107-CD154+IFN-γ+TNF+
	M.tb-lysate	IL-2-CD107-CD154-IFN-γ-TNF+
		IL-2-CD107-CD154-IFN-γ+TNF-
		IL-2-CD107-CD154-IFN-γ+TNF+
		IL-2-CD107-CD154+IFN-γ-TNF+
		IL-2-CD107-CD154+IFN-γ+TNF-
		IL-2-CD107-CD154+IFN-γ+TNF+
		IL-2-CD107+CD154-IFN-γ-TNF-
		IL-2-CD107+CD154-IFN-γ+TNF+
		IL-2-CD107+CD154+IFN-γ+TNF+
		IL-2+CD107-CD154-IFN-γ-TNF+
		IL-2+CD107-CD154-IFN-γ+TNF-
		IL-2+CD107-CD154-IFN-γ+TNF+
		IL-2+CD107-CD154+IFN-γ-TNF-
		IL-2+CD107-CD154+IFN-γ-TNF+
		IL-2+CD107-CD154+IFN-γ+TNF-
		IL-2+CD107-CD154+IFN-γ+TNF+
	E6C10	IL-2-CD107-CD154-IFN-γ+TNF-
		IL-2-CD107-CD154-IFN-γ+TNF+
		IL-2-CD107-CD154+IFN-γ+TNF+
		IL-2+CD107-CD154-IFN-γ-TNF+
		IL-2+CD107-CD154-IFN-γ+TNF-
		IL-2+CD107-CD154-IFN-γ+TNF+
		IL-2+CD107-CD154+IFN-γ+TNF-
		IL-2+CD107-CD154+IFN-γ+TNF+
CD8+ T cells	*M.tb*-lysate	IL-2-CD107+CD154-IFN-γ-TNF-
		IL-2+CD107-CD154-IFN-γ+TNF+
	E6C10	IL-2+CD107-CD154-IFN-γ+TNF+
Total Th1+ cells	*M.tb*-lysate	CD45RA-CCR7-CD27+KLRG1-
		Total HLA-DR+
B cells	Unstim	Total IL-10
		GB-IL-6-IL-10+IL-12-TNF-
MAIT cells	*M.tb*-lysate	Total IFN-γ
		Total TNF
		GB-IFN-γ+IL-6-IL-12-TNF+
NKT cells	*M.tb*-lysate	Total IFN-γ

Table 4.2: A summary of the cells that had a decreasing trend identified by kmlShape.

Cell Type	Stimulation	Effector Function/ Phenotypic Markers
Total Th1+ cells	*M.tb*-lysate	Total CCR7
		Total CD45RA
		Total CD27
		CD45RA-CCR7+CD27-KLRG1-
		CD45RA-CCR7+CD27+KLRG1-
		CD45RA-CCR7+CD27+KLRG1+
		CD45RA+CCR7+CD27+KLRG1-
		CD45RA+CCR7+CD27+KLRG1+

Wilcoxon's signed rank test was then applied to compare pre- and post-conversion time points in the 55 variables that were identified to have changed. Of those 55, 49 had significant differences ($p < 0.05$) between these two time points. We further explored those variables that had p-values greater than 0.05 by plotting their variable trajectories and raw data values. Figure 4.8 shows total IFN-γ production by *M.tb*-lysate-specific NKT cells, which had a p-value greater than 0.05 when comparing pre- and post-conversion time points (Figure 4.8A), but the median variable trajectory did have an increasing trend (Figure 4.8C). Therefore, the algorithm did correctly cluster this trajectory, however, kmlShape may just be more sensitive to changes over time than Wilcoxon's test.

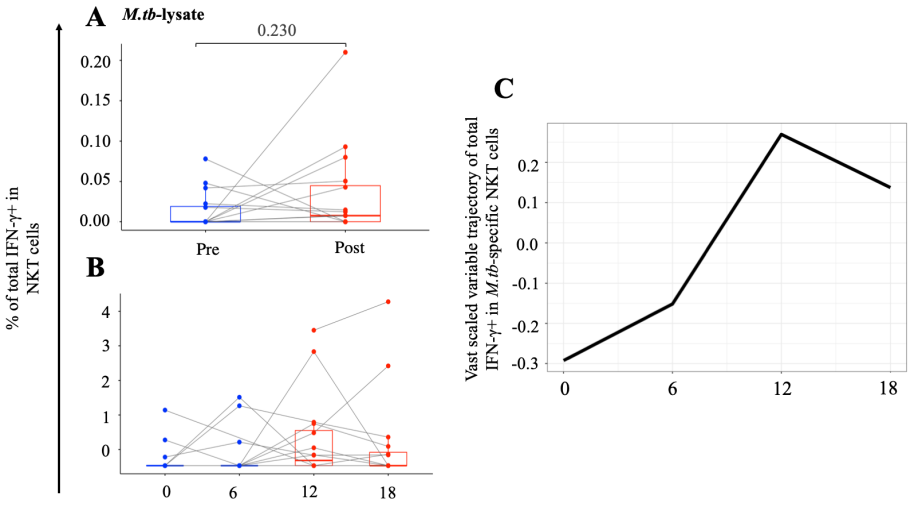

Figure 4.8: A representative variable identified by kmlShape. IFN-γ production by *M.tb*-lysate-specific NKT T cells at (**A**) pre- and post-conversion time points, (**B**) all four longitudinal time points, and (**C**) the vast scaled median variable trajectory for this variable.

4.3.2 Associations between the variables that changed upon QFT conversion

After we had identified the variables that changed using the kmlShape method, we explored how associations between distinct immune responses may change before and after *M.tb*-infection. The networks built to the pre- and post-conversion time points are shown in Figures 4.9A and B respectively. The networks consisted of *M.tb*-specific CD4+ T cells, total Th1 cells, CD3+ T cells, MAIT cells and unstimulated B cells with phenotypic markers or cytokine expressions that either increased

or decreased post-conversion. The associations between the nodes correspond to correlation coefficients that were greater than or equal to 0.5 with significant BH adjusted p-values. Therefore, if nodes are not connected, they either have a weak positive correlation less than 0.5, or a negative correlation. Negative correlations between nodes, although not seen in the network by means of a connection, can be found by looking at the data output from the network.

The less cellular differentiated total Th1 cells (purple), those variables that decreased over time, were negatively correlated with the CD4+ T cells (light blue) prior to *M.tb* infection, which increased over time. Total CCR7, CD54RA and CCR7+CD45RA+CD27+KLRG1- expression on total Th1 cells were strongly correlated with each other, and formed weak correlations with CCR7+CD45RA-CD27+KLRG1-, HLA-DR and CCR7+CD45RA+CD27+KLRG1+ pre-conversion. Post-infection, total CCR7 and CCR7+CD45RA-CD27+ KLRG1-, and CD45RA and CCR7+CD45RA+CD27+ KLRG1- had strong correlations with each other, and total CD45RA and CCR7 were now weakly correlated. In addition, HLA-DR no longer correlated with the other total Th1 cells, as this cell subset increased post-conversion.

In the pre-conversion network, CD4+ T cells loosely clustered according to the cytokines they produced, with CD154+ T cells clustering to the left in the network, TNF+ T cells in the middle and IFN-γ+ T cells to the right. Post-conversion, however, the CD4+ T cells formed an exclusive and tight sub-network regardless of cytokine production. A similar pattern was seen in the CD3+ T cells (red), however, pre-conversion, TNF+ CD3+ T cells were not correlated with any CD4+ T cells, or with IFN-γ+ CD3+ T cells.

The correlations observed between the cell subsets from the adaptive dataset were expected and unsurprising. Therefore, we were more interested in associations that occurred between cell subsets of the adaptive and innaptive dataset. Pre-conversion, CD4+IL-2+CD107-CD154-IFN-γ-TNF+ T cells and total IL-10 production in unstimulated B cells (orange) had a weak positive correlation (Figure 4.9C(i)). Further, total IFN-γ production and the co-production of IFN-γ and TNF in MAIT cells (grey) both positively correlated with total IFN-γ producing CD3+ T cells, and CD4+ T cells that co-expressed TNF and IFN-γ (Figure 4.9C(ii)). Hence, unstimulated B cells producing IL-10 correlated with *M.tb*-specific CD4+ T cells that co-expressed IL-2 and TNF, while MAIT cells expressing IFN-γ correlated with CD4+ and CD3+ T cells that also expressed IFN-γ or TNF. The cell subsets from the innaptive dataset had no negative correlations with adaptive variables pre-conversion.

After *M.tb* infection, total IL-10 production in unstimulated B cells no longer correlated with CD4+IL-2+CD107-CD154-IFN-γ-TNF+ T cells, and had no positive correlations with any cell subset. GB-IL-6-IL-10+IL-12-TNF- B cells did, however, have a negative correlation with CCR7+ CD45RA+CD27+KLRG1+ total Th1 cells (Figure 4.9D(i)). The MAIT cell subsets also no longer formed any positive correlations with CD4+ T cells, but formed a tight subnetwork with CD3+ T cells expressing IFN-γ and TNF, and total IFN-γ production (Figure 4.9D(ii)). Cell subsets post-conversion, therefore, formed tight sub-networks according to cell type, with the MAIT and CD3+ T cell subsets as an exception.

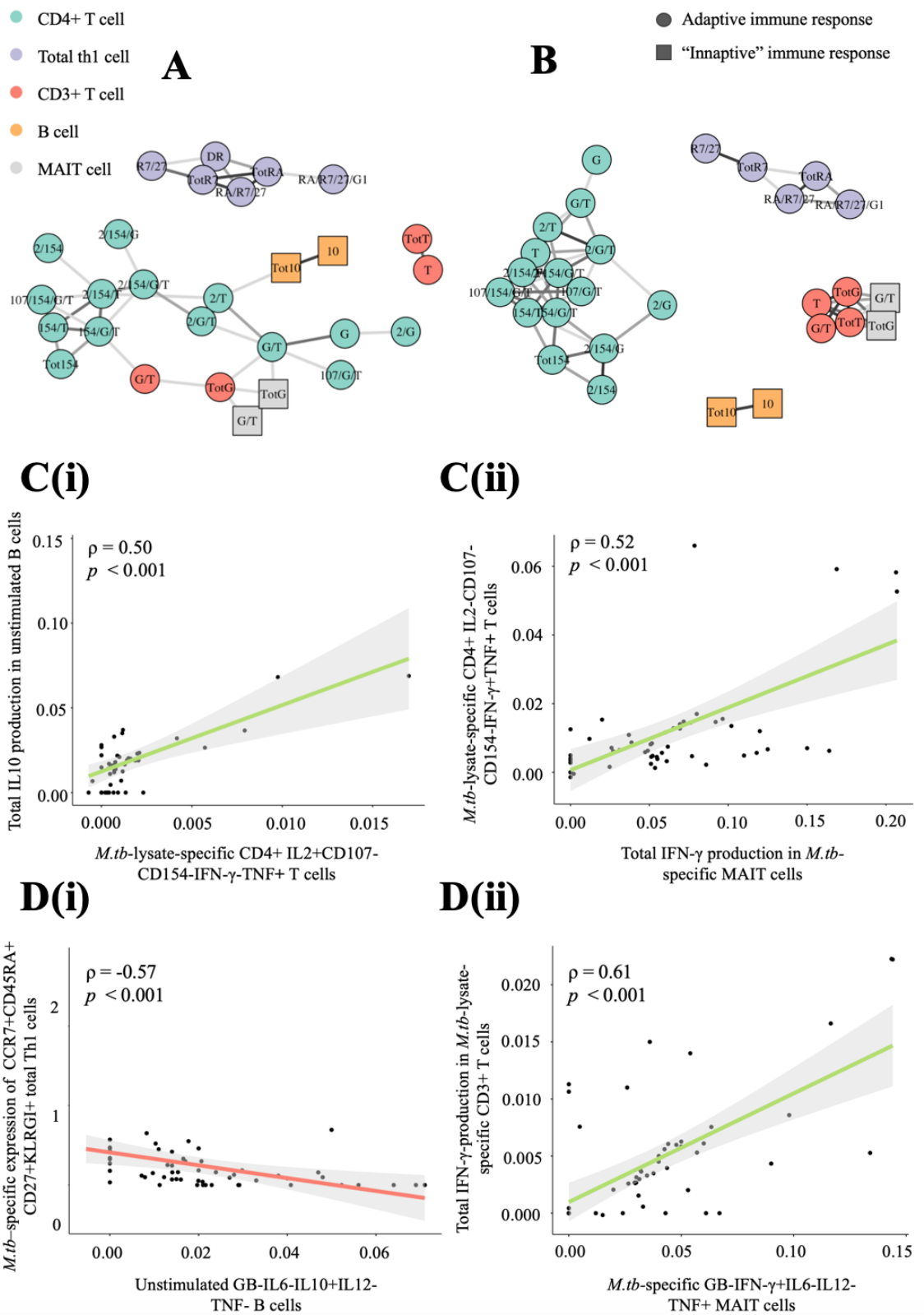
CD4+ T cell
Total th1 cell
CD3+ T cell
B cell
MAIT cell
A
Adaptive immune response
"Innaptive" immune response
B
C(i)
ρ = 0.50
p < 0.001
Total IL10 production in unstimulated B cells
M.tb-lysate-specific CD4+ IL2+CD107-CD154-IFN-γ-TNF+ T cells
C(ii)
ρ = 0.52
p < 0.001
M.tb-lysate-specific CD4+ IL2-CD107-CD154-IFN-γ+TNF+ T cells
Total IFN-γ production in M.tb-specific MAIT cells
D(i)
ρ = -0.57
p < 0.001
M.tb–specific expression of CCR7+CD45RA+ CD27+KLRGI+ total Th1 cells
Unstimulated GB-IL6-IL10+IL12-TNF- B cells
D(ii)
ρ = 0.61
p < 0.001
Total IFN-γ-production in M.tb-lysate-specific CD3+ T cells
M.tb-specific GB-IFN-γ+IL6-IL12-TNF+ MAIT cells

Figure 4.9 *(previous page)*: Correlation network analysis. The pre- and post-conversion networks are shown in (**A**) and (**B**) respectively and the nodes are colored according to cell type and are square if that cell type is from the innaptive immune response and circular otherwise. Associations between nodes from the different datasets were further looked into and (**C**) shows scatter plots between ((**i**)) *M.tb*-lysate-specific CD4+IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells and total IL-10 production in unstimulated B cells, and ((**ii**)) *M.tb*-lysate-specific CD4+IL-2+CD107-CD154-IFN-γ-TNF+ T cells and total IFN-γ production in *M.tb*-lysate-specific MAIT cells, which both had positive correlations pre-conversion. (**D**) shows a scatter plot between (**i**) unstimulated GB-IL-6-IL-10+IL-12-TNF- B cells and *M.tb*-specific expression of CCR7+CD45RA+CD27+KLRG1+ total Th1 cells who had a negative correlation, and ((**ii**)) GB-IFN-γ+IL-6-IL-12-TNF+ MAIT cells when stimulated with *M.tb* and total IFN-γ production in *M.tb*-specific CD3+ T cells who had a positive correlation post-conversion. Spearman's correlation coefficients and their associated p-values are superimposed onto the plots.

4.3.3 Statistical validation of the kmlShape algorithm

As a statistical validation of the kmlShape algorithm, a network was built to the same 193 variable trajectories from the integrated dataset (Figure 4.10A). The input was not restricted to *M.tb*-lysate stimulated cell subsets or variables that changed upon conversion. This network revealed many associations between different cell types measured in the different datasets.

Without specifying the number of expected communities in the network, the optimal community detection algorithm identified three communities. The nodes belonging to each community are colour coded accordingly in Figure 4.10B. We plotted the variable trajectories of the variables that fell into each community, and superimposed the corresponding mean trajectory over time (Figure 4.10C-E). The first community (light blue) was made up of 70 variables consisting of CD4+ T cells, including the three *M.tb*-specific polyfunctional CD4+ T cells, CD8+ and CD3+ T cells, total Th1 cells, MAIT cells, B cells and NK cells. Wilcoxon's non-parametric test was again applied to the variables to compare pre- and post-conversion time points. Of the 70 variables belonging to the first community, only 37 had significant differences between pre- and post-conversion time points, but they all displayed an increasing trend over time. The second community structure (orange) consisted of 49 variables that appeared to have no clear trend over time. 43 variables then made up the final community structure (green) and seemed to group the variables with a decreasing trend over time. This community consisted of primarily total Th1 cell differentiation phenotypes and cell subsets from the innaptive immune response. Eight variables in this community had significant p-values and they all had a decreasing trend over time. The network and community detection algorithm identified a total of 113 variables that changed, which is double the number of increasing or decreasing variables identified by the kmlShape algorithm. Twelve variables were identified by kmlShape to have changed which were not identified by the correlation network. This included GB-IL6-IL10+IL12-TNF- unstimulated B cells and HLA-DR expression on *M.tb*-specific total Th1 cells, which both have a clear increasing trend.

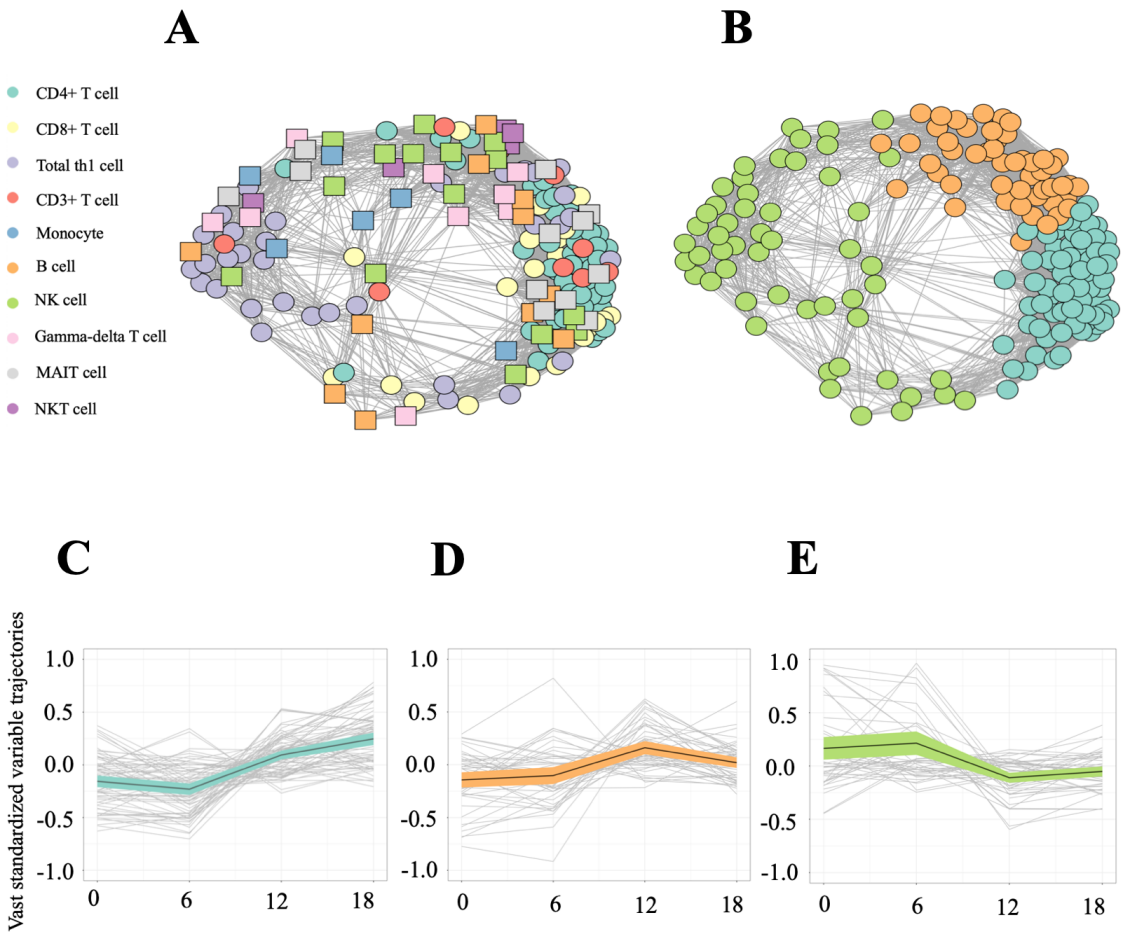

Figure 4.10: Statistical validation of kmlShape using correlation networks. (**A**) shows the correlation network built to all the variable trajectories in the integrated dataset, where the nodes are colored according to cell type and are square if that cell type is from the innaptive immune response and circular otherwise. (**B**) shows the same network after optimal community clustering algorithm has been applied and nodes are coloured according to which community they belonged to. (**C-E**) shows the vast scaled variable trajectories of the variables that fell into each community, with the mean trajectory superimposed onto the plot. The variables in (**C**) seem to have an increasing trend and correspond to the blue cluster, (**D**) have no specific trend corresponding to the orange cluster, and (**E**) appear to have a decreasing trend corresponding to the green cluster.

4.4 Discussion

The aim of this chapter was to identify and define a set of immune features that changed upon infection with *M.tb*. A common approach to identify changes in repeated measures data is to use simple parametric or non-parametric tests to compare the time points. However, suppose t hypothesis tests are simultaneously tested, the probability of observing a significant result, simply due to chance, would be $1-(1-0.05)^t$. This probability tends towards 1 as the number of tests performed, t, increases. It is almost guaranteed that at least one false positive significant result (type I error) would be observed for a large number of tests performed. To account for this, p-values are adjusted for multiple testing. Although employing paired tests and using p-values to draw conclusions from

the data is still common practice, issues surrounding p-values and adjusting for multiple testing are becoming a topic of great debate [86][137][131]. The criticism of p-values is not a recent phenomenon [32], and p-values have been branded as misleading and are strongly depend on the sample size of the data. For multiple testing, it is unclear whether the p-values should be corrected for all tests performed, or just for the tests that are included in the final research [86], and some studies even suggest that multiple comparison adjustments are not necessary [107]. There is no hard and fast rule for when adjustments for multiple tests are necessary [13], however, a general guideline is that when the results of the hypothesis tests are used to make a final decision, as in clinical trials, p-value adjustment in mandatory [113]. Otherwise, in observational studies, it is not necessarily required, but it is recommended that results are interpreted accordingly [42]. For example, an interpretation of a small p-value in an observational study could be that the small p-value corresponds to a greater likelihood the null hypothesis being rejected.

Given all the uncertainties surrounding p-values, and how to correctly adjust for them, this project explored two alternative methods that could identify variables that changed using clustering algorithms. The algorithms included kmlShape, which extends the popular k-means algorithm to cluster longitudinal trajectories according to their shapes over time, and using community detection algorithms on correlation networks. Both algorithms were applied to each variable's median trajectory over time, which was found after the data had been standardized using vast scaling. Therefore, to ensure that the data pre-processing steps did not affect the results, and to simply compare how the clustering algorithms differed to non-parametric tests to identify changes over time, we subsequently employed Wilcoxon's signed rank test. Wilcoxon's test was applied to compare raw and unstandardized pre- and post-conversion time points in the variables that were identified to have changed as a consequence of *M.tb*-infection by the clustering algorithms.

By setting $k = 3$, the kmlShape algorithm clustered the variable trajectories according to three longitudinal trends: an increasing trend, a decreasing trend, and those that remained constant over time. In total, kmlShape identified 55 variables that changed over time, which were assumed to be a consequence of infection with *M.tb*. Of those 55, however, only 49 had p-values less than 0.05 after applying Wilcoxon's signed rank test to the pre- and post-conversion time points. When we explored those variables that were not significant, we found that the kmlShape algorithm did correctly cluster the variables, as the median variable trajectories did either have an increasing or decreasing trend. Therefore, kmlShape is either more sensitive to changes over time than Wilcoxon's test, especially given the small sample sizes for the innaptive dataset, or kmlShape captured the time granularity that was masked in the paired tests as a result of aggregating the time points. Another advantage of the kmlShape algorithm is that no cut-off values was required to conclude that a variable had changed over time, which was the case for Wilcoxon's test. Resulting p-values from the statistical tests needed to be less than a pre-specified threshold, typically chosen to be 0.05, in order to conclude whether a change is significant or not. This threshold can be arbitrary and is another criticism of p-values [32]. However, Wilcoxon's test does take into account variability, where inter-person variability is lost by using the median trajectories in klmShape. In any case, the majority of the variables that increased were *M.tb*-specific CD4+ T cells, while those that decreased were *M.tb*-specific total Th1 cells expressing phenotypic markers mostly consistent with an early differentiated phenotype. It is expected that, upon infection, antigen-specific T cells transition from an early to a late differentiated phenotype.

Correlation networks, built to pre- and post-conversion time points separately, were then used to explore how associations between cells may change as a result of infection with *M.tb*. Only those variables identified to have changed as a result of TBI by the kmlShape algorithm were included in the network. We further restricted the analysis to variables that were either stimulated with *M.tb*-lysate, the common stimulation across the two datasets, or variables that were left unstimu-

lated and had significant pre- and post-conversion differences. Associations that had a Spearman correlation greater than 0.5 with a significant BH adjusted p-value were plotted in the network. In pre- and post-conversion networks, features from the innaptive dataset formed few and weak correlations with features from the adaptive dataset, the most notable being total IL10 production in unstimulated B cells with IL2 and TNF expressing CD4+ T cells pre-conversion, and the association between TNF and IFN-γ producing MAIT and CD3+ T cells post-conversion. There appeared to be no biologically meaningful associations that formed across features from the different datasets, and associations did not differ significantly as a result of *M.tb* infection.

An alternative clustering approach to identify variables that changed as a result of TBI involved building a correlation network to all variable trajectories in the integrated dataset. The assumption was that variables with an increasing trend would correlate with each other, and form negative correlations with those variables that have a decreasing trend. A community detection algorithm was then applied to identify dense sub-graphs, or communities, within the network. Without prespecifying the number of communities, the algorithm identified three sub-networks in the graph. When the trends of the variables belonging to each community were plotted, it seemed that the algorithm had grouped variables according to their longitudinal trends. This algorithm identified 113 variables to have changed over time, which is double the number identified by kmlShape. Upon further exploration of the variables in each community, we found that this method did not correctly classify each longitudinal trajectory as successfully as the kmlShape algorithm nor was it a suitable method of validation. Any two variables that had a positive correlation, for example, regardless of their trend over time, would have clustered with the variables in the first community. This is because, when calculating the correlations between the variable trajectories, the multiple time points were not accounted for, which could have lead to false positive or negative correlations. The kmlShape algorithm was the more robust method.

Regardless of the statistical methods applied in this chapter, we observed that *M.tb*-specific polyfunctional CD4+ T cells increased post-conversion. This provided sufficient evidence to fail to reject the hypothesis that levels of IFN-γ+TNF+IL-2+ *M.tb*-specific CD4+ T cells would be higher post-conversion compared to pre-conversion.

4.5 Conclusion

We employed a sophisticated clustering algorithm, kmlShape, which was more sensitive to changes over time than Wilcoxon's paired tests, to group variables based on their longitudinal trends. To our knowledge, we have provided the most extensive list of immunological features induced by recent acquisition of *M.tb* infection, in both the innaptive and adaptive immune system, to date. Identifying features in the immune system that are modulated by *M.tb* infection, could be useful to understand the biology of early events after *M.tb* infection, beyond the narrow focus of IFN-γ producing T cells measured by current immunodiagnostic tests.

Chapter 5

Identifying biomarkers of recent *Mycobacterium tuberculosis* infection

AIM 2: To investigate which features of the immune response differ between recent and established *M.tb* infection.
We hypothesize that proportions of TNF-only producing CD4+ T cells with a T_{EFF} phenotype and HLA-DR expression on M.tb-specific CD4+ T cells are higher in individuals with recent M.tb infection compared to those with established infection.

5.1 Introduction

Currently, no reliable test exists that can successfully detect the presence or absence of *M.tb* in asymptomatic individuals. The IGRA and TST are classically used to support diagnosis of *M.tb* infection. Neither test can differentiate between active and latent TB nor indicate the time since *M.tb* infection. The risk of TB disease is significantly higher in individuals with recent TBI compared to those with established TBI. Individuals with recent TBI are at the highest risk of progressing to TB disease during the first 2 years post-infection [12]. The work described in this chapter aimed to identify blood-based immune signatures associated with early stages. Being able to distinguish between recent and established *M.tb* infection through specific immune markers, of which no current diagnostics (besides serial testing) can distinguish these two phases of infection, will help to target those individuals requiring preventative treatment. TB treatment involves chemotherapy for a minimum period of 6 months, which may have potential side-effects and is costly in high burden settings. It would be a great benefit to reduce the number of individuals receiving unnecessary LTBI treatment.

The datasets available for this analysis consisted of many highly correlated variables, which, if not handled correctly, could lead to optimistic measures of performance. Therefore, the predictive models were built using regularized regression with sparsity, a common approach for feature selection, particularly in datasets with highly correlated variables. Common methods include ridge regression [64], the LASSO model [128], and the elastic net (EN) model [146]. In their recent paper, Liu and colleagues [81] observed that the standard EN approach does not take into account differing effect sizes in predictors from different datasets in an integrated model. The authors hypothesized that their method, the multiple tuning parameter elastic net (MTP-EN), could account for theses differences and result in a model with higher predictive performance. As a sub-aim, the predictive performance of the MTP-EN model was assessed and compared to the standard EN model to deter-

mine whether the MTP-EN model did in fact improve the predictive performance of the integrated dataset. Further, it was hypothesized that the integrated model, whether it be the standard EN or MTP-EN, built to multiple immune variables would outperform models built on the individual data types in stratifying individuals with recent or established TBI.

For statistical validation, two tree-based machine learning algorithms were additionally built to the data; a simple classification decision tree [23] and a random forest (RF) learning algorithm [21]. A decision tree is a supervised model that is used to predict a response by learning decision rules from features in a given dataset. Decision trees are advantageous as they are easy to interpret, the feature importance is clear and relations between predictors can be viewed easily. A downfall of decision trees is that they often result in an overfitted model and they suffer from high sampling variability [70]. Therefore, a RF model was additionally tuned and built to the data. RF models extend classification trees by building multiple trees and merging them together to make decisions. This helps get more accurate and stable predictions.

5.2 Methods

5.2.1 Study design

Infection with *M.tb* likely occurred between months 6 and 12 in the QFT converters, when the QFT test converted from a negative to a positive result. Recent TBI was therefore characterized by taking the median value of the two QFT positive time points for each individual in the recent QFT converters ($n = 29$) across all variables. The established QFT converters, on the other hand, were infected with *M.tb* prior to the study, and so the median value of the time points available for each individual in the persistent positive cohort across all variables was used to define established TBI ($n = 30$) (Figure 5.1). Differences between the four sampling occasions in this cohort were tested for all the variables using Friedman's test. No significant statistical differences were found and hence we could justify taking their median values (Chapter 3).

All models fitted in this chapter were built to the filtered data, the data that included only biologically meaningful cell subsets identified by the various filtering methods applied (Section 3.4), which was set up to contain the recent and established TBI observations. The variables were then standardized using vast scaling (Section 3.5) and the missing values were accounted for using MFA imputation (Section 3.6). The MFA imputation method was repeated for each CV run during parameter tuning or model validation to ensure that any results found were not a consequence of the imputation method used.

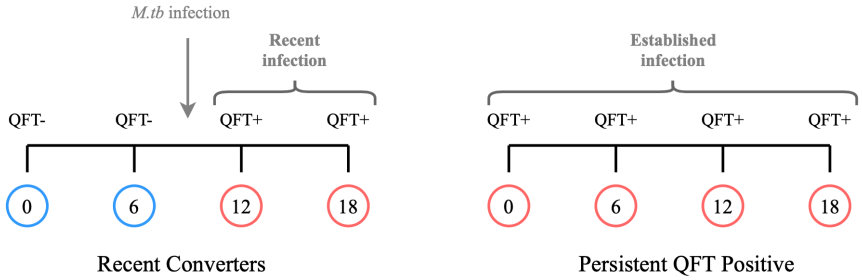

Figure 5.1: Definition of recent and established QFT conversion.

5.2.2 Cross validation

CV is a common resampling procedure that is used to tune parameters and validate models with limited sample sizes. The method is typically referred to as k-fold CV as it splits the data into a user defined k number of groups [89].

A simple graphical display of the method is shown in Figure 5.2. The observations in the dataset are randomly shuffled and then split or "cut" into k groups or "folds". Each fold contains $\frac{n}{k}$ observations, where n is the total number of observations. For the first iteration of the algorithm, the observations in the first fold are held out and form the testing set, while the remaining observations in the $k-1$ folds make up the training set. A model is then built to the training set of observations and its performance is evaluated using the testing set. The performance of the model is then stored and the method is repeated for $2, \ldots, k$ iterations. The result is k measures of performance which are averaged to give the overall CV performance of that model.

The choice for the value of k is dataset dependent and has an associated bias-variance trade-off. A sufficiently large training set is needed to build a model such that it is as close as possible to the model that would be built to the entire dataset. At the same time, the testing set should be large enough to get an accurate testing performance. As k gets larger ($k \to n$), fewer observations make up the testing set, which increases the size of the training set. This will lower the bias but increase the variance, which results in an overfitted model that does not generalize well to unseen data. However, as k gets smaller, the size of the testing set increases at the expense of the training set size. This will decrease the variance but increase bias, which leads to an oversimplified model. k is generally chosen to be $k = 10$, $k = 5$ or $k = n$, which is known as leave-one-out CV (LOOCV).

CV suffers from high sampling variability, which means that different splits of the data will result in different CV performances each time. To alleviate this, and improve the estimated performance of the model, k-fold CV is repeated N times in a process known as Nxk-fold CV. Results from repeated CV procedures are reported as an average across the N CV repeats.

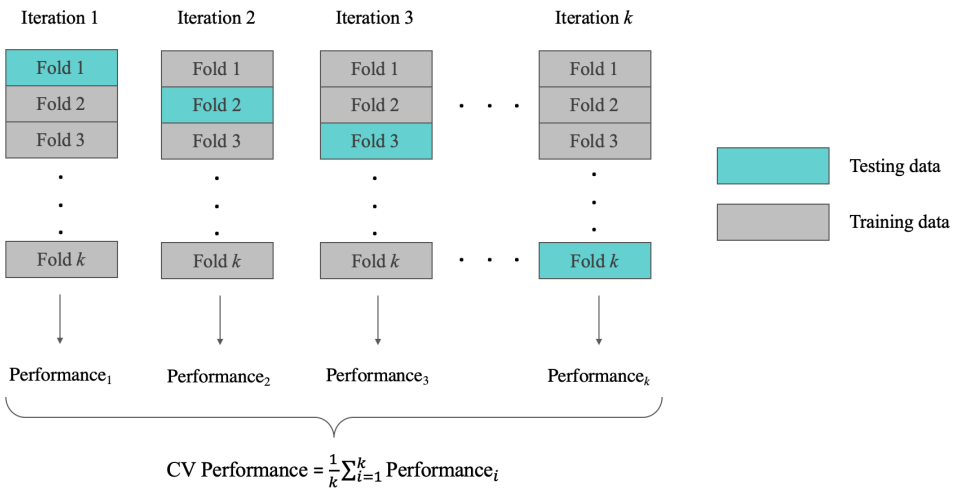

Figure 5.2: Cross validation (CV) protocol. CV is a common resampling procedure that is used to tune parameters and validate models and splits the dataset into k pre-defined groups. The figure shows one CV iteration. Each block, which represents one fold, contains $\frac{n}{k}$ of the total n observations in the dataset. The performance of the model is taken as the average performance across the k iterations.

The method described above is employing CV to internally validate a model's performance. As mentioned earlier, CV can also be used to tune parameters in a model. Suppose a model has a parameter value θ that needs tuning, then k-fold CV is repeated for each candidate value of $\theta = (\theta_1, \ldots, \theta_m)$. The value for θ that either minimizes or maximizes model performance, depending on the performance metric used, is then chosen as the "optimal" parameter value for the given dataset.

CV was used as both an internal validation procedure and to tune the parameters in the regularized regression models and the RF model. CV is the preferred method to data splitting when the sample size is small [61].

5.2.3 Receiver operating characteristic curves

As seen in the previous section, a good performance metric is required to perform CV, as it drives the decision for which parameter value will be optimal for the given dataset. Performance metrics will differ depending on the model used, but a common metric for classification problems in the area under the receiver operating curve. This metric quantifies how capable a model is of distinguishing between classes, where higher values are indicative of high predictive performances [60]. The receiver operating characteristic (ROC) curve is a two-dimensional graph that shows the performance of a classification model for various thresholds. Two measurements make up the ROC, namely the true positive rate and the false positive rate. Consider the following 2x2 contingency table,

	True outcome = 1	True outcome = 0	Total
Predicted outcome = 1	True positive (a)	False positive (b)	$a+b$
Predicted outcome = 0	False negative (c)	True negative (d)	$c+d$
Total	$a+c$	$b+d$	

where a, b, c and d are the number of individuals who fall into each outcome class in a binary classification problem, depending on their true outcome and predicted outcome values. Sensitivity, or the true positive rate, is defined as the proportion of individuals with true response = 1 who have a positive result:

$$\text{Sens} = \frac{a}{a+c}$$

Specificity, or the false positive rate, is then the proportion of individuals with true response = 0 who have a negative result:

$$\text{Spec} = \frac{d}{b+d}$$

The ROC is found by plotting the true positive rate (Sensitivity) $\in [0,1]$ on the y-axis and the false positive rate (1-Specificity) $\in [0,1]$ on the x-axis. The area under the ROC curve (AUC) is then the average performance of the model for all possible classification thresholds.

When tuning the parameters in the regularized regression models and the RF model, we used the AUC as a performance metric. It was also used to quantify and compare the predictive performances of the various models built.

5.2.4 Logistic regression and regularized regression

Logistic regression model setup and notation

A key goal of regression analyses in general is to analyze the relationship between a set of predictor variables (independent variables) and a response (dependent variable). Logistic regression (LR),

which is named for the function used at the core of the method, the S-shaped logistic function [15], is one such method and is used when the response is binary. The LR model assumes that all observations are independent of each other and there is little or no multicollinearity in the data (see below).

For $i = 1, \ldots, n$ independent observations, let $\mathbf{y} = (y_1, \ldots, y_n)$, $y_i \in \{0, 1\}$, be a vector of responses and let $\mathbf{x_1}, \ldots, \mathbf{x_q}$ be q corresponding predictor variables of length n. Further, define $P(y_i = 1|\mathbf{x_i}) = p_i$, where $y_i = 1$ is the default class. Following standard notation, a LR model, which aims to model the probability of the default class, built to a dataset with n observations and q predictors can be written in the following form

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq} = \beta_0 + \mathbf{x_i}^T\boldsymbol{\beta} \tag{5.1}$$

where β_0 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$ are the regression coefficients and $\frac{p_i}{1-p_i}$ is known as the odds of the default class. The left hand side of the equation is hence called the log-odds. Unlike linear regression, which assumes the relationship between the response and the set of predictors is linear, LR makes no assumptions about the linearity between the dependent and independent variables, but rather assumes that there is a linear relationship between the independent variables and the log odds.

The regression coefficients, which quantify the relationship between the response and each predictor variable, are found by solving the following optimization equation

$$\min_{\beta \in \mathbb{R}^q} - l(\beta_0, \boldsymbol{\beta})$$

where $l(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \mathbf{x_i}^T\boldsymbol{\beta}) - \log(1 + \exp(\beta_0 + \mathbf{x_i}^T\boldsymbol{\beta}))\right]$ is the log-likelihood for a LR model. Regression coefficients in logistic regression models are most easily interpreted in terms of the odds ratio, which are found by exponentiating the β coefficients. If the odds ratio is greater than one, it implies that, for every one unit increase in a given predictor variable, the odds of belonging to the default class increases. Otherwise if it is less than one, the odds of belonging to the default class decreases.

Regularized regression

As mentioned previously, a strong assumption in most standard regression models is that there is little or no multicollinearity in the data. Multicollinearity is described as a situation where two or more independent variables are linearly related or correlated. If predictor variables are correlated, then, intuitively, changes in one variable will induce changes in another. Therefore, the precision of the true effect of an independent variable on the dependent variable reduces and the statistical power of the model weakens. There are a number of ways to deal with multicollinearity in the data, such as removing highly correlated variables, or linearly combining the multicollinear variables. Alternatively, regularized regression models are used, which, unlike standard regression models, can effectively handle multicollinearity in the data.

Regularization is becoming an increasingly important statistical concept that is employed to avoid overfitting complex datasets. It is implemented, in a LR framework, by adding penalty terms to the log-likelihood optimization equation that best restricts the influence of some predictor variables by compressing their coefficients. In other words the penalty adds a bias towards certain values. Regularized regression models are also advantageous, and perform well in cases where the number of predictors are significantly larger than the number of observations ($p >> n$).

Ridge regression

Ridge regression imposes an L_2 penalty on the regression coefficients, which is the squared magnitude of the coefficient [64].

$$\hat{\beta}^{ridge} = \min_{\beta \in \mathbb{R}^q} - l(\beta_0, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{5.2}$$

where $l(\beta_0, \boldsymbol{\beta})$ is the log-likelihood for a LR model defined previously, and $\lambda \geq 0$ is known as the shrinkage parameter. When $\lambda = 0$, equation (5.2) will reduce to the standard LR optimization equation. As the value for λ increases, the variance decreases, *i.e.* the uncertainty in the estimates decreases, but the bias increases, where bias is the difference between the true population parameter and the expected estimator. Therefore, the value for λ needs to be tuned such that an optimal value is used for a given dataset to find a balance between the bias and variance in the model.

A parameter value of $\lambda > 0$ results in a model with the same number of features, but a reduction in the magnitude of their regression coefficient. Ridge regression therefore reduces the complexity of a model but does not decrease the number of variables. The method just shrinks their effect. This method hence cannot produce a parsimonious model.

LASSO

Least absolute shrinkage and selection operator (LASSO) is advantageous as it performs variable selection and parameter estimation simultaneously [128]. A LASSO model is achieved by imposing an L_1 penalty on the regression coefficients, which is the absolute magnitude of the coefficient:

$$\hat{\beta}^{lasso} = \min_{\beta \in \mathbb{R}^q} - l(\beta_0, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5.3}$$

where a value of $\lambda = 0$, equation (5.3) will again reduce to the standard LR optimization equation. In contrast to ridge regression, which reduces the magnitude of some coefficients in the model, the LASSO model is able to shrink variable coefficients to zero. The larger the value of λ, the more coefficients of features are shrunk to zero. Therefore, LASSO models can eliminate some features entirely and provide a subset of predictor variables that help to mitigate multicollinearity and decrease model complexity. Non-zero predictors signify that they are important and this is how the method allows for feature selection.

A limitation of the LASSO model as a feature selection method is that, if there is a group of highly correlated variables in the dataset, the LASSO model will select only one of these variables and will select it at random and ignore the rest. For this project, we were interested in which specific binary combinations of cytokines produced by, say CD4+ T cells, could distinguish between recent and established TBI. CD4+ T cells producing various different combinations of cytokines would be highly correlated with each other. Because we were not interested in a random selection of a CD4+ T cell subset, but more on the selection of which specific CD4+ T cell subset, or subsets, that could distinguish between the two phases, LASSO was an inappropriate variable selection technique.

Elastic net

The EN model [146] is a combination of both the ridge regression and the LASSO models, and, similar to the LASSO model, simultaneously performs feature selection and parameter estimation

$$\hat{\beta}^{EN} = \min_{\beta \in \mathbb{R}^q} - l(\beta_0, \boldsymbol{\beta}) + \lambda \left[(1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right]. \tag{5.4}$$

The EN model introduces an additional parameter, α, where $0 \leq \alpha \leq 1$ is the weight given to the L_1 penalty, and $1 - \alpha$ the weight to L_2 penalty, and λ is the overall shrinkage parameter. The

EN model was designed to overcome the issue of variable selection with groups of highly correlated variables and cases when $p >> n$, and therefore is the more suited method for this project.

Multiple tuning parameter elastic net

The authors of the multiple tuning parameter EN (MTP-EN) method observed that EN models tends to shrinks the effect of features simultaneously and does not take into account differing effect sizes of different datasets in an integrated model [81]. The MTP-EN model applies different tuning parameters to the different data types in an integrated dataset to account for these differences. By doing so, the MTP-EN approach could result in a model with higher predictive performance.

For the MTP-EN model set up, assume there are k different datasets, each with $i = 1, \ldots, n$ samples and with the same binary outcome $y_i \in \{0, 1\}$. The design matrix $X = [X^{(1)}|X^{(2)}|, \ldots, |X^{(k)}]$ will then be a partitioned matrix made up k design matrices from the k different datasets. Further define $N(\beta) = (1-\alpha)\sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|$ to be the EN penalty. The MTP-EN therefore extends the standard EN by applying differing penalties, $\lambda_1, \lambda_2, \ldots \lambda_k$ to coefficients from different datasets

$$\min_{\beta \in \mathbb{R}^q} - l(\beta_0, \boldsymbol{\beta}) + \lambda_1 N(\beta^{(1)}) + \lambda_2 N(\beta^{(2)}) + \ldots + \lambda_k N(\beta^{(k)}). \tag{5.5}$$

For this project there were only $k = 2$ data types that made up the integrated dataset. The MTP-EN model was built to the variables in the integrated dataset using the glmnet R package [49] via the "*penalty.factor*" argument. This allows a weighted EN penalty of the form

$$N_\omega(\beta) = \alpha \sum_{j=1}^{p} \omega_j |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \omega_j \beta_j^2$$

where ω is p-dimensional weight vector with 1 in the first $p_1 = 106$ entries corresponding to the variables in the first dataset, and $\kappa = \frac{\lambda_2}{\lambda_1}$ for the $p_2 = 70$ entries in the second dataset. The MTP-EN penalty in Equation 5.5 can be written as

$$\lambda_1 N(\beta^{(1)}) + \lambda_2 N(\beta^{(2)}) = \lambda_1 N_\omega(\beta)$$

Consequently, $\lambda = \lambda_1$ controls the overall degree of shrinkage for both data types and κ controls the shrinkage of one data type relative to the other.

Regularized regression model building and parameter tuning

An overall workflow for the modeling portion of this aim is shown in Figure 5.3. 10-fold CV repeated 500 times was used to find optimal parameter values for λ and α, with the AUC as the performance metric in the MTP-EN model. This was repeated for each candidate weight parameter $\kappa \in [0.2, 1.8]$. The highest AUC values after 10-fold CV were stored after each repeat and the performance of the MTP-EN for each κ was reported as an average of the 500 AUC values. A parameter value of $\kappa = 1$ is equivalent to a standard EN model, hence the performance of the standard EN model could be directly compared to MTP-EN models with varying degrees of penalties applied to each dataset. The result of this experiment was used to determine whether an MTP-EN or standard EN model would be the most suitable for the integrated dataset.

Two standard EN models were further built to the individual datasets separately using the glmnet package. Once again, optimal parameter values for λ and α were tuned using 500x10-fold CV, and the average of the selected parameters across the 500 searches were thereafter defined as the "optimal" parameter values. Relevant candidate biomarkers for classifying TBI were identified as features with non-zero coefficients in the final model, and the predictive performances in terms of AUC values of the models were then compared.

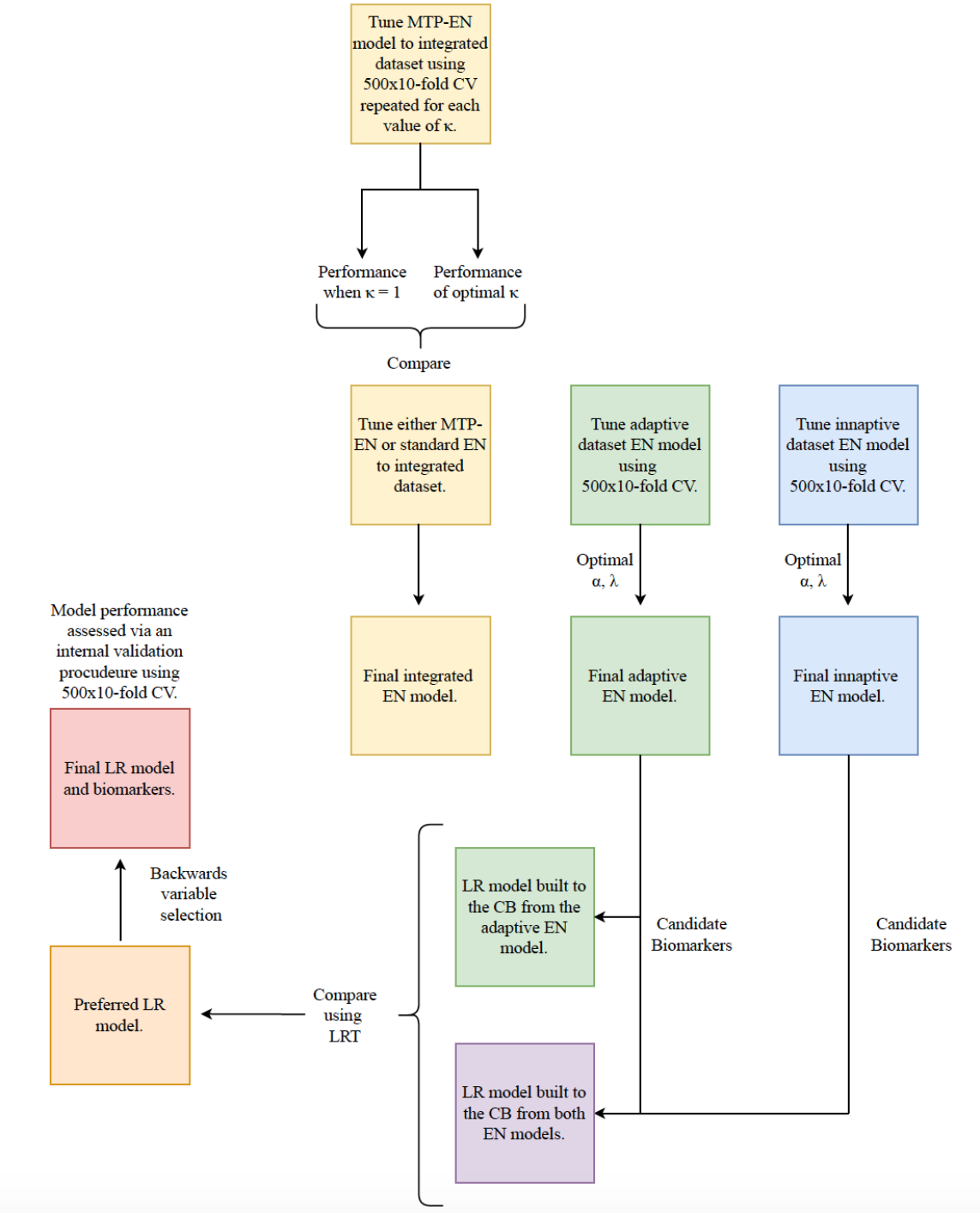

Figure 5.3: A flow chart that outlines the methods for the modeling portion of this aim. AUC stands for area under the curve, LR stands for logistic regression, CB stands for candidate biomarkers, CV stands for cross validation and LRT stands for likelihood ration test.

Building the logistic regression model

The candidate biomarkers identified from the innaptive and adaptive EN models were used to build two LR models with established TBI as the default class: one model was built using the biomarkers identified in the adaptive model, and another using the biomarkers identified from both models. The likelihood ratio test (LRT) was used to assess whether adding the innaptive biomarkers to the LR model resulted in a statistically-significant improvement in the fit of the model. The LRT is a goodness of fit test and identifies the model that better maximizes the likelihood function [95].

For two possible models, with different parameter values θ, the likelihood ratio is defined as follows:

$$\text{LRT} = -2ln\left(\frac{L(\theta_1; y)}{L(\theta_2; y)}\right) \sim \chi^2(k)$$

where L is the likelihood and θ_1 has k fewer parameters than θ_2. The null hypothesis of the LRT states that the simpler model is the preferred model, *i.e.* the model on the numerator with fewer parameter values is the model that provides a better fit for the data. The LRT has an approximate Chi-Squared (χ^2) distribution with k degrees of freedom (dof), and hence the null hypothesis is rejected if the value for the LRT is larger than a χ^2 percentile with k dof. The χ^2 percentile, $100(1-\alpha)$, will be a pre-chosen level of confidence, which is typically chosen to be $\alpha = 0.05$.

The LRT was used to identify the preferred model, whether it be the adaptive features only or a combination of adaptive an innaptive features. Backwards variable selection was performed on the preferred LR model to identify the best subset of predictors. This type of variable selection involves including all predictors in the model and removing insignificant variables in a step-wise fashion [77].

Internal validation

The predictive performance of the final model, and subsequently the set biomarkers, was assessed using an internal validation procedure. 10-fold CV repeated 500 times was performed on the model with AUC and the Brier score [24] as performance metrics. The Brier score is a measure of the accuracy of a probability forecast

$$BS = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2$$

where $0 \leq f_i \leq 1$ is the probability forecast and $y_i \in \{0, 1\}$ is the true outcome. Essentially, the Brier score is the mean squared error (MSE) equivalent for LR models and values closer to zero indicate better forecasting ability. The Brier score is the preferred performance metric over the misclassification error for small sample sizes. Results reported are an average of the performance metrics across the 500 CV repeats.

For all instances in this study when CV was performed, the missing values in the dataset was imputed separately for the training and testing sets using MFA imputation. Therefore, the dataset was imputed several times to ensure that any results found were not just a consequence of the imputation method.

5.2.5 Classification trees

The decision tree algorithm

The overall aim of the decision tree algorithm is to divide a predictor space, $\mathbf{x}_1, \ldots, \mathbf{x}_p$ into non-overlapping regions or nodes, using a set of splitting rules [23]. These regions are found using a greedy, top-down algorithm known as recursive binary splitting (RBS). Top-down means that,

initially, all observations belong to a root node (Figure 5.4), which are then split into separate nodes recursively. The algorithm is greedy because it only considers the best split at each step.

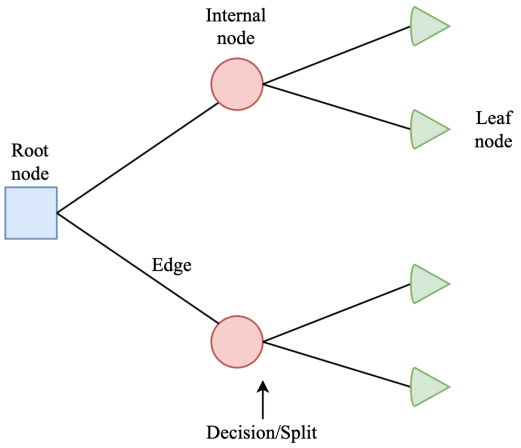

Figure 5.4: A simple illustration of a decision tree. A root node (blue) has no incoming edges and only outgoing edges, an internal node (red) will have one incoming edge and two or more outgoing edges, and a leaf node (green), also known as a terminal node, will have one incoming edge and no outgoing edges.

In a decision tree, each leaf node, also known as the terminal node (Figure 5.4), is assigned a class label. In this setting the class label would be either recent or established infection. The non-terminal nodes, which includes the root node and all internal nodes, will contain a decision rule to split the observations. A decision rule will be made up of a split value for a given predictor variable. Hence, the RBS algorithm runs through each predictor variable $\mathbf{x}_1, \ldots, \mathbf{x}_p$ sequentially and aims to identify a split value s for that predictor variable $\mathbf{x}_j$, $j = 1, \ldots, p$. All continuous variables are discretized. Let x_{ij} be the i^{th} observation in predictor variable $\mathbf{x}_j$. For each row $i = 1, \ldots, n$, x_{ij} is compared to s: if $x_{ij} < s$, row i will be assigned to the left node, otherwise it is assigned to the right node. The dataset has now been split into two groups.

A cost function is then used to quantify the "cost" of this data split at each step. Common cost functions for classification problems include the Gini Index or the cross entropy function. For this project, we used the Gini Index, which is a measure of node impurity and is calculated as follows

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

where $\hat{p}_{mk}$ is the proportion of observations in response category k within leaf node m. Ideally, each leaf node would include observations of only one response category k *i.e.* the leaf node is "pure". Data splitting and evaluation via the Gini Index is repeated for every predictor variable and for every candidate split within that predictor variable. From the set of all predictor and candidate split values, a predictor $\mathbf{x}_j$ and split s that results in the greatest reduction in the Gini Index is chosen as a node in the decision tree. The RBS algorithm is recursive in nature and the internal nodes formed can be further sub-divided using the same strategy. Therefore, the predictor variable and split value that make up the root node will be the best classifying feature in the decision tree.

Intuitively, if the tree is grown continually until each node corresponds to the lowest impurity

(a tree with large depth), then it is likely that the data as been overfitted. Similarly, if the tree growth is stopped too early (a tree with smaller depth), the data may be underfitted. Therefore, to control for this, some pre-defined stopping criteria are applied during the tree building process. These can include stopping the partitioning if the current node does not decrease the Gini Index by some pre-defined threshold, or restricting the depth of the tree to some pre-defined value.

Fitting the decision tree

Typically, a decision tree is built to a training set and the set of rules created are used to predict the testing set. The observations in the testing set are classified as the most commonly occurring class in the region to which it belongs. However, a decision tree was built mainly to visualize the relationship between the variables in the model and assess feature importance in stratifying recent from established TBI. Therefore a simple tree was built to all the observations in the dataset using the R package rpart [126], using the default stopping criteria.

5.2.6 Random forest models

Model setup

A RF model is a nonparametric multivariate technique that is essentially a number of decision trees built on B bootstrapped training samples of the data [21]. Bootstrapping is a common resampling method that involves repeatedly drawing samples of the same size, with replacement, from a single original sample [41]. RF models are a more popular and widely used application of the decision tree machine learning algorithm.

To begin, a bootstrapped dataset is created, which may have the same observation chosen more than once. A decision tree is then built in the same process described above using this bootstrapped dataset. However, RF models differ from decision trees in that only a random subset of $m < p$ variables or features is considered at each step of building the tree. The model therefore aims to identify the best feature among this smaller set of features. These two steps, creating a bootstrapped sample and only considering a subset of predictors at each step, are repeated a large number of times and result in a wide variety of trees, which make up the forest. By doing this, RF models will have an improved predication accuracy over a simple decision tree and also reduce the risk of overfitting the model.

For each variable, the total decrease of the Gini Index across each tree in the forest is accumulated each time the variable is included in the tree. This total is then divided by the number of trees in the forest (B) to give the average total decrease in the Gini Index for a given predictor. Hence, it is simple to establish feature importance in RF models, as a large average decrease in the Gini Index will be indicative of an important predictor. It is important to note here that multicollinearity in the data does not reduce the predictive performance of a RF model, but it does reduce feature importance. Intuitively, this is because once one of the two correlated variables is chosen as a classifier, the other variable is less important to bring any further output variation explanation. Therefore, the importance of some variables may be higher than indicated.

When predicting the testing dataset, every observation is run through each of the B trees and the class outcome, based on the specific decision rules from each tree, is recorded. The final predicted class is then the average prediction across the B trees. Using bootstrapped data and taking the aggregate to make a decision is known as "bagging".

Building the RF model

The RF was built to the data in order to confirm the best classifying features identified by the simple decision tree. The RF model to the data using the randomForest R library [22], which has the following hyperparameters that needed to be tuned

- the number of decision trees to make up the forest (the number of bootstrapped samples),
- the split rule; either "gini" or "extratrees",
- the variable importance measure,
- the number of features, $m < p$, to be considered at each split, and
- the maximum number of edges from root to terminal node to control overfitting (the depth of the tree).

The "extratrees" parameter implements the extremely randomized trees, or extra trees [54], splitting criterion. Essentially, for each feature from the subset of m randomly selected features $\mathbf{x}_i$, $i \in \{1, \ldots, m\}$, a single random cut-point is drawn from a uniformly from the interval (min($\mathbf{x}_i$),max($\mathbf{x}_i$)). Similar to the RF model, the cost of this split is then evaluated using the Gini Index. This method differs from the RF models, which chooses the optimum split, while extra trees chooses the split randomly.

The hyperparameters for the RF model were tuned using 500x10-fold CV. Similar to the EN models, the "optimal" hyperparameters were taken as the average across the 500 repeats and used to build the final RF model. For the splitting rule, the optimal parameter was chosen as the most commonly occurring class out of "gini" or "extratrees". The final RF model was used to identify the 10 most important features corresponding to the largest mean decrease in the Gini Index. The performance of the model was then assessed via the same internal validation procedure as with the final LR model.

5.3 Results

5.3.1 The MTP-EN model

The MTP-EN model was built to all 176 variables from both datasets and the effect on the testing AUC for different values of κ in the MTP-EN is shown in Figure 5.5. The model performance reached a plateau for $\kappa > 0.8$, where values of $\kappa < 1$ result in a smaller penalty applied to the innaptive dataset. Differential penalization therefore resulted in lower comparative AUC values when the effect sizes of the features in the adaptive dataset in particular were decreased relative to the innaptive dataset ($\lambda_1 > \lambda_2$). In terms of the predictive performance as determine by the AUC, the standard EN ($\kappa = 1$) and the highest performing MTP-EN model with $\kappa = 1.7$ were identical with an average AUC of 0.93. In addition, the computation time for the MTP-EN model was significantly longer (24.4 hours) compared to the EN model (2.52 hours). Accordingly, for this specific dataset, there was no added benefit to fitting the MTP-EN model over the standard EN model. The EN model was therefore used to build the integrated model.

5.3.2 Biomarker discovery and internal validation

Three EN models were subsequently built: one on the integrated dataset and one on each of the adaptive and innaptive data types separately. The final EN models built on the integrated and adaptive variables had identical parameter values for α and λ (0.25 and 0.84 respectively) and consequently identified the same candidate biomarkers corresponding to an average AUC value of 0.91.

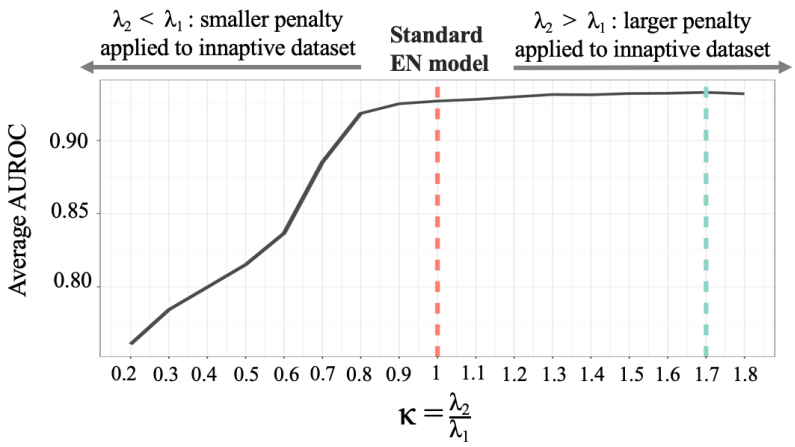

Figure 5.5: The MTP-EN model. Figure shows the average of 500 AUC values as a function of κ. A red dashed line is plotted at $\kappa = 1$, which is equivalent to a standard EN model, and a blue line at the "optimal" $\kappa = 1.7$, which corresponds to the highest mean AUC.

The biomarkers identified were total Th1 cells expressing the phenotypic marker HLA-DR, identified by stimulation with either E6C10 or *M.tb*-lysate, and the frequency of IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells stimulated with Esp. Comparing the two stages of *M.tb* infection using a Wilcoxon non-parametric test, recent TBI values for these variables were found to be significantly higher ($p < 0.001$) than established TBI (Figure 5.6). Hence, these variables could be validated as potential biomarkers for recent infection.

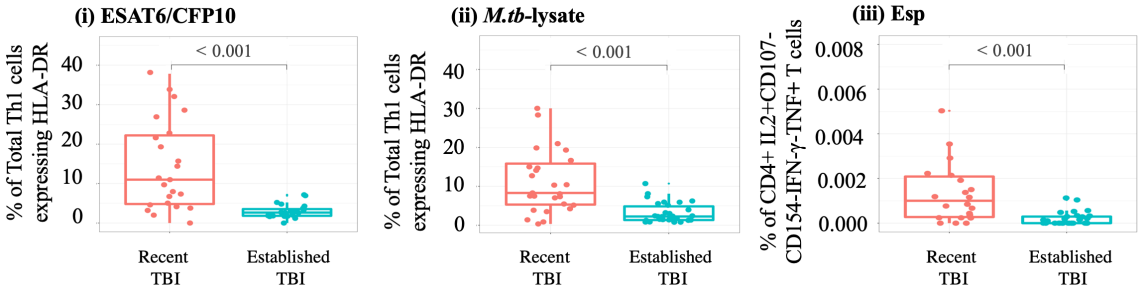

Figure 5.6: The identified biomarkers. Figure shows boxplots comparing raw values of recent (red) and established (blue) TBI for expression of HLA-DR on total Th1 cells when stimulated with **(i)** E6C10 or **(ii)** *M.tb*-lysate, and **(iii)** IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells stimulated with Esp. Wilcoxon tests were used to compare the two groups and the resulting p-values are superimposed onto the plots.

The EN model built on the innaptive dataset yielded optimal values for α and λ as 0.15 and 0.37 respectively. This model identified 10 candidate biomarkers, which corresponded to a poor average AUC of 0.62.

LR models were then built to the candidate biomarkers from the single dataset EN models. The results from the LRT indicated that a combination of the non-zero coefficients from both EN models did not significantly improve model fit ($\chi^2 = 6.09$, $p = 0.808$), and the LR model built with the

adaptive biomarkers only was the preferred model:

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{E6C10\_HLADR}_i + \beta_2 \text{Mtb\_HLADR}_i + \beta_3 \text{Esp\_CD4IL-2TNF}_i.$$

Backwards variable selection on this model further identified *M.tb*-lysate-specific total Th1 cells expressing HLA-DR as a statistically insignificant biomarker. *M.tb*-lysate contains a mix of all *M.tb* antigens, including E6C10, and therefore HLA-DR expression under the two stimulations is highly correlated (Spearman's Rho = 0.82, $p < 0.01$; Figure 5.7) and including it in the model would invalidate model assumptions. The final LR model that was built to the data was

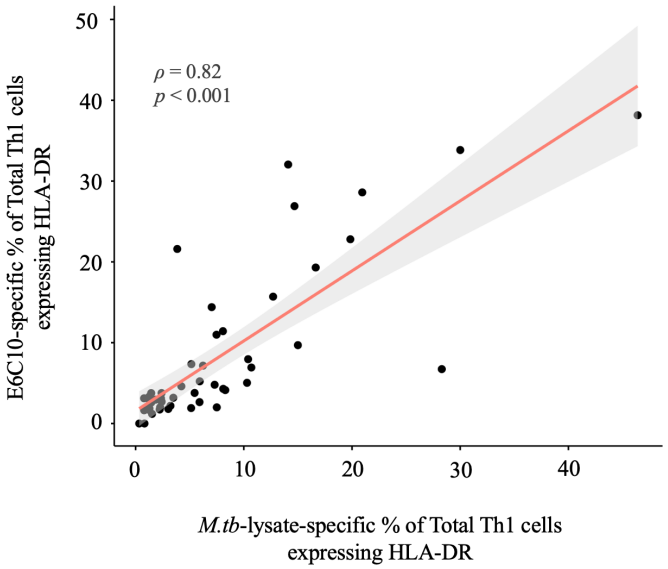

Figure 5.7: The correlation between *M.tb*-lysate (x-axis) and E6C10 (y-axis) stimulation on total Th1 cells expressing HLA-DR. Spearman's non-parametric correlation coefficient was used to quantify the degree of correlation and the R value (correlation coefficient) and its associated p-value are superimposed onto the plot.

$$\begin{aligned}\text{logit}(p_i) &= \beta_0 + \beta_1 \text{E6C10\_HLADR}_i + \beta_2 \text{Esp\_CD4IL-2TNF}_i \\ &= -1.55 - 4.34 \text{ x E6C10\_HLADR}_i - 2.79 \text{ x Esp\_CD4IL-2TNF}_i.\end{aligned}$$

LR models are most easily interpreted in terms of the odds ratio (OR), which are found by exponentiating the regression coefficients. Holding all other variables fixed, the odds of being an established QFT converter (the default class) decreases (OR = $e^{\beta} < 1$) by 99% (100[1-$e^{-4.34}$]) = 100[1-0.01]) for every one standardized unit increase total th1 expression of HLA-DR in response to E6C10. Similarly, for every one standardized unit increase of Esp-specific CD4+IL-2+CD107-CD154-IFN-γ-TNF+ T cells, the odds of being an established QFT converter also decreases, now by 94%. Both these estimated ORs were significantly different from one ($p < 0.05$). Accordingly, as the value of either one of these biomarkers increases, the chances that an individual is a recent QFT converter, or has recently been infected with *M.tb*, increases, and this relationship is statistically significant. The performance of this model was then assessed via an internal validation procedure and performed satisfactorily (average AUC and Brier scores are 0.89 and 0.008 respectively). Table 5.1, model (i), summarizes the regression coefficients, odds ratios and performance metrics for this model.

Including the frequency Esp-specific IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells as a coefficient in the LR model statistically improved model fit (LRT: $\chi^2 = 12.76$, $p < 0.001$). We noted that the ability for this subset to successfully distinguish between recent and established stages of infection was dependent on the cell subset being CD107-CD154-IFN-γ. This was tested by comparing the predictive performance of a LR model built to Esp-specific IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells and one built to any IL-2+TNF+ CD4+ T cell when stimulated with Esp. The average AUC value for the latter LR model was lower (average AUC = 0.72) compared to the former (average AUC = 0.86). E6C10-specific HLA-DR was then included as an additional variable to the two LR models, and the model with Esp-specific IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells had a lower comparative Akaike information criterion (AIC) [3] value (42 compared to 57). Given a collection of models for the data, the AIC estimates the quality of each model. It takes into account goodness of fit through the likelihood function and imposes an additional penalty for increasing number of parameters. Lower AIC scores are indicative of a good model fit to the data and hence the AIC provides a means for model selection [77].

The ability of E6C10-specific HLA-DR to distinguish between the different stages of infection as a single biomarker was then assessed. The performance of this model and the model including Esp-specific IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells were similar and equally high (average AUC and Brier scores are 0.87 and 0.007 respectively; Table 5.1, model (ii)).

$$\begin{aligned}\text{logit}(p_i) &= \beta_0 + \beta_1 \text{E6C10\_HLADR}_i \\ &= -0.91 - 4.06 \text{ x E6C10\_HLADR}_i.\end{aligned}$$

Table 5.1: Outcome from LR models. The table summarizes the model estimates and the average performance metrics after internal validation of (i) the final LR model and (ii) LR model built to E6C10-specific total Th1 cells expressing HLA-DR only.

	Coefficients	β	e^{β}	p-value	Avg. AUC	Avg. Brier
i	(Intercept)	-1.55	0.21	0.032	0.89	0.008
	E6C10_HLA-DR	-4.34	0.01	0.003		
	Esp_CD4IL-2TNF	-2.79	0.06	0.028		
ii	(Intercept)	-0.91	0.40	0.075	0.87	0.007
	E6C10 HLA-DR	-4.06	0.02	< 0.001		

5.3.3 Statistical validation

Using the same vast standardized and MFA-imputed dataset, a simple classification tree was built to all the 176 features in the dataset. The plot of this decision tree is shown in Figure 5.8A. The tree identified two features from the set of all variables in the integrated dataset to best discriminate between recent and established TBI; the expression of HLA-DR on total Th1 cells when stimulated with E6C10 and the frequency of Esp-specific IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells.

Initially, all $n = 59$ observations from the dataset, of which 30 are established converters and 29 are recent converters, are stored in the root node (node 1). E6C10-specific expression of HLA-DR on total Th1 cells was identified as the best classifying feature in the dataset. Rows with standardized values of HLA-DR on total Th1 cells greater than or equal to -0.098 (Figure 5.8B) were assigned to leaf node 2 on the left. This node contains 17 observations, of which all of them are recent TBI individuals. Otherwise, the remaining observations were assigned to leaf node 3. The frequency of Esp-specific CD4+IL-2+CD107-CD154-IFN-γ-TNF+ T cells was identified as the

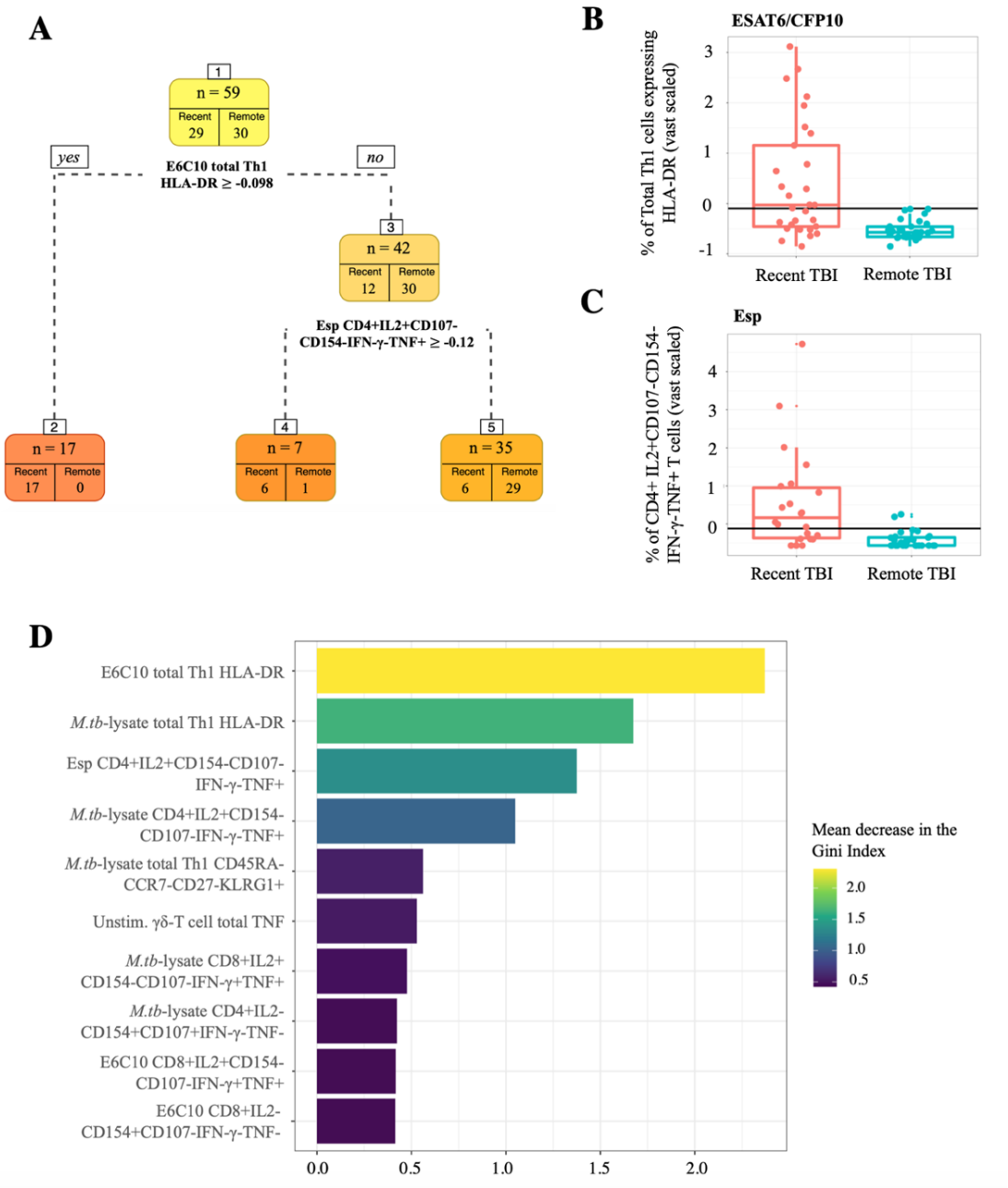

Figure 5.8: Statistical validation of the regression models. (**A**) a simple classification tree built to all the observations in the dataset. To illustrate the cut-offs identified by the decision tree, boxplots were used to compare the vast scaled values of recent (red) and established (blue) TBI. Black lines are superimposed onto the plots at **(B)** -0.098 for the expression of HLA-DR on total Th1 cells when stimulated with E6C10 and at **(C)** -0.12 for the frequency IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells stimulated with Esp. **(D)** is the variable importance plot from the final RF model.

next best classifying feature in the dataset and was used to further discriminate recent from established stages of TBI. Those observations with standardized values greater than or equal to -0.12 (Figure 5.8C) for this feature were assigned to leaf node 4. This node contains 7 observations, of

which 6 were recent TBI individuals and 1 misclassified established TBI individual. The remaining observations were then allocated to the final leaf node (node 5). The majority of the observations in this node were correctly classified as established converters ($n = 29$), and the 6 were misclassified as recent converters. The decision rules identified by the tree resulted in 12% (7 out of 59) of the observations being misclassified.

For the RF model, the following parameter values were identified by CV, and therefore the model was built such that:

- 500 decision trees built to 500 random bootstrapped samples of the data made up the forest,
- a subset of 25 out of the 176 features were considered at random at each split,
- each tree built was allowed no more than 10 nodes from root to terminal node, and
- "extratrees" was used as a splitting rule.

The Gini Index was then used to measure variable importance, and the 10 variables that resulted in the largest mean decrease in the Gini Index in the final RF model are shown Figure 5.8D. Among the 10 top performing variables were 9 from the adaptive and one single variable of the innaptive dataset. The only innaptive variable identified was TNF production of $\gamma\delta$ T cells when left unstimulated.

After an internal model validation procedure, the average AUC for the final RF model was 0.84 and the average misclassification error was 0.12, precisely the misclassification error of the simple classification tree. The misclassification error is a natural measure of performance for the RF model, however, it can be misleading in this instance given the small sample size of the dataset.

5.4 Discussion

Several regularized regression modelling approaches and machine learning algorithms were applied to a flow cytometry dataset to identify immunological biomarkers that could distinguish recent and established TBI. Because our focus in this aim was less on estimating model coefficients, but more on identifying predictive markers, instead of using Rubin's rule to take into account imputation variability, we rather repeated the MFA imputation for each CV run. Therefore, we were confident that the results found here were not a consequence of the imputation method used.

A MTP-EN model, which applies differential penalties to the different datasets in the integrated model, was built to the full dataset. The results indicated that applying a greater penalty to the adaptive dataset ($\kappa < 1$) yielded comparatively worse average AUC values. However, when a smaller penalty was applied to this dataset ($\kappa > 1$), the predictive performances were identical to the model with no additional penalty (the standard EN model). In terms of computing power and predictive performance, there was insufficient evidence to justify building the MTP-EN over the standard EN model for the integrated dataset. The MTP-EN results did, however, demonstrate the importance of the features of the adaptive immune response in stratifying the two stages of *M.tb* infection.

The EN model built to the integrated dataset identified only non-zero coefficients from the adaptive dataset as important features. These candidate biomarkers from this model were the exact same two biomarkers identified by the adaptive EN model, namely total Th1 cells expressing HLA-DR and Esp-specific IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells. HLA-DR is a cell-surface receptor that has been found to be an early immune marker reflecting T cell activation in response to bacteria during TB disease [2][59][91][105][139]. The robustness of HLA-DR as a biomarker was confirmed with *M.tb*-lysate, however, E6C10 is a more appropriate biomarker for use in diagnostic

tools as it is a single antigen whereas *M.tb* lysate contains a mix of different antigens and is therefore less standardized. TNF and IL-2 produced by CD4+ T cells, on the other hand, are early response cytokines that both play an important role in the context of TB [47].

The true effect of these two identified biomarkers on the probability of an individual being remotely *M.tb* infected was estimated through a LR model. Higher standardized frequencies of both these biomarkers were associated with a larger probability, or odds, of an individual being recently infected. This reflects the relationship that is seen in the raw data plots, where individuals recently infected with *M.tb* had significantly higher values in these features compared to remotely infected individuals and confirms the hypothesis of this aim. The performance of LR model built to these two biomarkers was assessed via an internal validation procedure and, given the small sample size, were sufficiently high.

Diagnostic tests made up of a single biomarker are generally simpler and more cost effective. Therefore, the ability of HLA-DR to successfully distinguish between the two stages of TBI as a biomarker on its own was tested. Moving forward with HLA-DR as a single diagnostic measure was justified by substantial literature showing excellent performance of this biomarker to distinguish different stages of the TB spectrum [59], and the small number of markers necessary to measure this biomarker (as few as four [91]). In addition, using this same ACS study data as this project, Mpande et al. (2020) [90] found that the difference in expression of HLA-DR, measured by median fluorescent intensity, between *M.tb*-specific (IFN-γ+TNF+CD3+ upon stimulation with E6C10 or *M.tb*-lysate) and total T cells could differentiate individuals with recent TB infection (TBI) from those remotely infected with high sensitivity and specificity. The frequencies of Esp-specific IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells were also extremely low (values range between 0 and 0.006), which could be challenging to measure in a robust way. Lastly, because the discriminatory ability of Esp-specific IL-2+CD107-CD154-IFN-γ-TNF+ CD4+ T cells was in fact dependent on the subset being negative for CD107, CD154 and IFN-γ, the panel for a diagnostic test including this biomarker would be complex. The final LR model included Esp-specific IL-2+CD107-CD154-IFN-γ-TNF+ frequencies of CD4+ T cells, however, it is clear based on predictive performances that E6C10-specific HLA-DR expression alone is an equally strong single biomarker of recent TBI.

The validity of HLA-DR as a biomarker of recent infection was then further confirmed in both the simple decision tree and the results of the RF model. HLA-DR expression in response to E6C10 was found to be the best classifying feature among the set of all individuals in stratifying recent from established TBI in the classification tree, and, in the RF model, corresponded to the largest average decrease in the Gini Index. At the beginning of the study, we set out to test two hypotheses. From the results of the analyses above, there was sufficient evidence against rejecting the hypothesis that HLA-DR expression on *M.tb*-specific CD4+ T cells are higher in individuals with recent *M.tb* compared to those with established infection. The other hypothesis was that proportions of TNF-only producing CD4+ T cells with a T_{EFF} phenotype would also be higher in recently infected individuals. This was a biomarker identified by Halliday and colleagues [59]. Similar variables were included in this dataset, but were not selected in the models. Therefore, one of the pre-defined hypotheses was true, while the other was not.

A final observation from this aim is that it was evident that integrating variables from the innaptive dataset did not improve the fit of the regression models. Only features from the adaptive dataset were identified as important predictors in the integrated EN model, and results from a LRT demonstrated that the non-zero coefficients from the innaptive dataset were unable to outperform the strongest candidate biomarkers from the adaptive dataset. One variable from the innaptive dataset, namely total TNF production in $\gamma\delta$ T cells when left unstimulated, was found to be the sixth most important variable in the final RF model. However, there was a significant difference

between the average decrease in the Gini Index for the top four most important variables compared to the rest, therefore the likely effect of this innaptive variable is negligible in comparison.

5.5 Conclusion

As far as it is known, this study includes the most comprehensive integrated evaluation of adaptive and innaptive immune responses induced by recent TBI in humans published to date. The results show that the innaptive immune responses were poor predictors of recent TBI, and did not improve the performance of the integrated model. Based on the results reported here, the expression of HLA-DR on E6C10-specific T cells was selected as the strongest candidate biomarker for recent TBI and was validated using a different machine learning algorithm and in a separate test cohort [90]. HLA-DR as a biomarker holds the potential to identify individuals who would benefit from preventive TB treatment, however, due to the small sample size in this study, further validation is recommended in an independent cohort for conclusive results.

Chapter 6

QFT Reverters

AIM 3: To identify which features of the immune response are associated with QFT reversion.
Under the assumption that QFT reversion is associated with clearance of M.tb infection, we hypothesize that immune features induced by M.tb infection will have opposite trends upon QFT reversion.

6.1 Introduction

The clinical implications of QFT reversion are still unclear. It is possible that reversion is associated with the clearance of *M.tb* or could be a consequence of technical assay variability. Alternatively, the QFT reverters could be a biologically distinct group of individuals who have a different immune response upon infection with *M.tb*. In this chapter, we set out to analyze the QFT reverters in more detail to explore which of the above categories best explains this cohort.

Assuming that QFT reversion is associated with the clearance of *M.tb*, we hypothesized that variable trajectories for immune responses in the reverters would have opposite trends compared to the those from the recent QFT converters over time. The trajectories of the variables that were identified in Chapter 4 to have changed as a result of TBI were investigated in QFT reverters (Figure 6.1A).

The reverters were also then compared to the two control groups, namely the persistent QFT negatives and persistent QFT positives. The relationship between the reverters and the two control cohorts was initially assessed using principal component analysis (PCA) [101] and then projection to latent space, or partial least squares discriminant analysis (PLS-DA) [142]. PCA is a popular dimension reduction method that decomposes an input matrix into latent structures of a lower dimension, such that as much of the information as possible from the original input matrix is preserved. By taking the eigenvalue decomposition of the covariance matrix of a set of predictors, the principal components (PC) are the eigenvectors corresponding to the largest eigenvalues. Although primarily a dimension reduction technique, PCA can be extended to a regression problem, where the PCs are used to predict an output. This is known as principal component regression (PCR). Given the unsupervised nature of PCA, however, there is no guarantee that the linear combinations that best represent the input will also be the best to linear combinations to predict the output. Therefore, PLS is the preferred method as it chooses features that best explain both the output and the input.

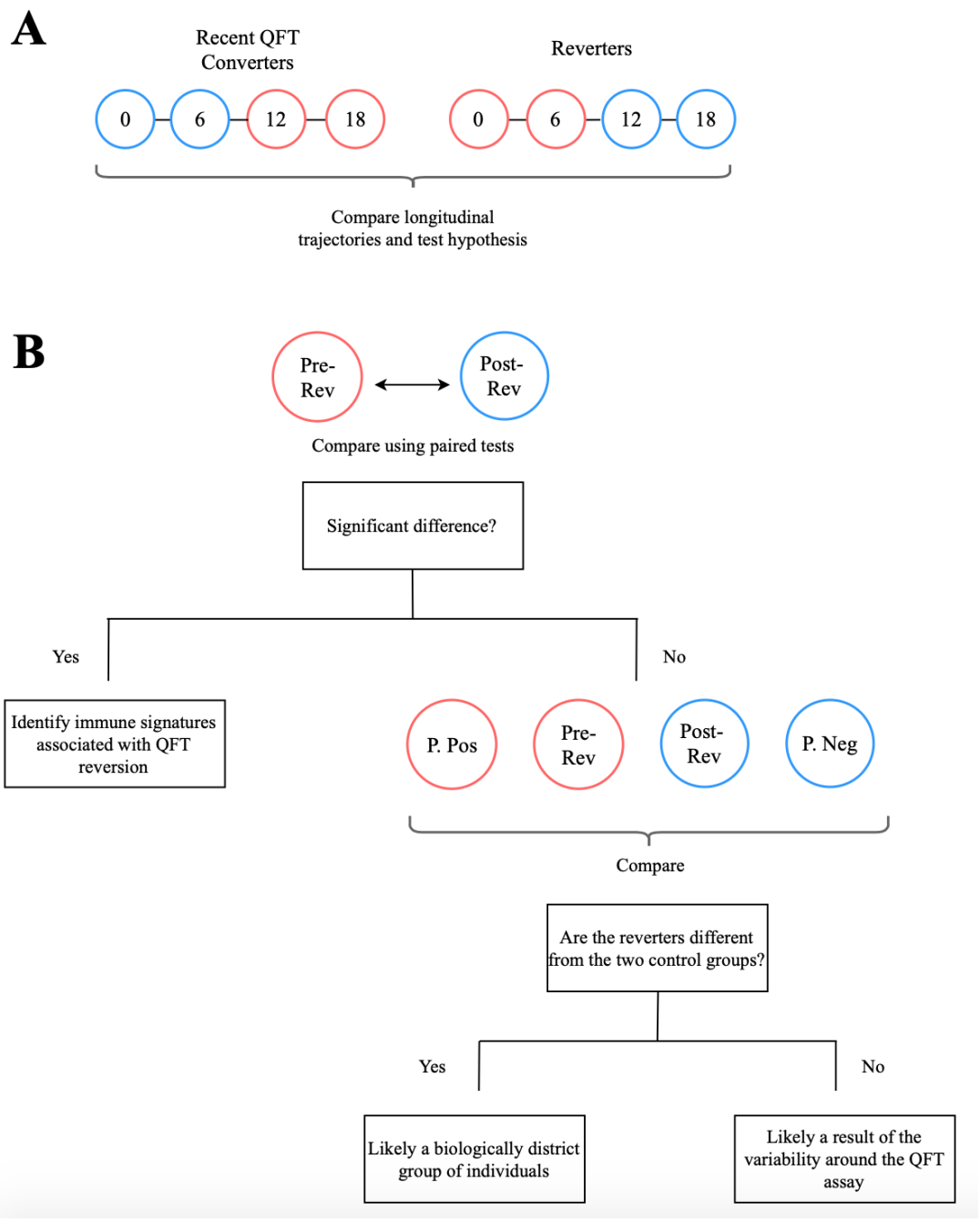

Figure 6.1: Workflow. (**A**) the longitudinal trajectories of the immune features in the reverters were compared to the immune features in the recent QFT converters to test whether QFT reversion reversion is likely associated with loss of *M.tb* infection and to test the set-out hypothesis. (**B**) pre- and post-reversion time points were then compared using simple, non-parametric tests to assess which immune features are associated with QFT reversion. If no features changed over time in the QFT reverters, they were compared to the two control groups, the persistent QFT positive and negative cohorts.

PLS regression was introduced by Wold in 1996 [143] as an alternative to the standard OLS re-

gression approach, as it can handle many, multicollinear variables. PLS regression generalizes PCA and multiple linear regression, and is used to predict a set of dependent response variables from a set of independent predictor variables. PLS differs from OLS, however, in that the prediction is achieved by extracting a set of latent variables from the explanatory variables which have the best predictive power [1]. Using these latent variables is also how the method handles multicollinearity in the data.

Although PLS was principally designed for regression purposes, it can be extended to classification and discrimination problems by simply re-coding the response vector with dummy variables to indicate class membership of each sample. This method is known as PLS-DA and is preferred to linear discriminant analysis (LDA), which has numerical limitations for large, highly correlated datasets [79]. PLS-DA is advantageous if this goal of the analysis is to classify samples into known groups and identify variables that drive the discrimination between groups.

6.2 Methods

6.2.1 Study design

To infer which features of the immune response may be associated with QFT reversion, paired tests were performed between pre- (QFT+) and post-reversion (QFT-) time points for all variables in the integrated dataset. The resulting p-values were adjusted for multiple comparisons and a significant difference was defined by an adjusted p-value less than 0.05 (Figure 6.1B). If no significant differences were found in the immune features, then the reverters were likely a biologically distinct group of individuals or a result of a variability around the QFT assay. Pre- ($n = 30$) and post-reversion ($n = 30$) time points were compared to the two control groups, namely the persistent QFT negatives ($n = 30$) and persistent QFT positives ($n = 30$) to explore this further (Figure 6.1B).

All analyses were run on the vast scaled (Section 3.5) and MFA imputed data (Section 3.6), which included all biologically relevant cell subsets identified by the various filtering methods applied (Section 3.4), except E6C10-specific phenotypic markers on total Th1 cells. These cells were not measured in the QFT reverters.

6.2.2 Variable trajectories in recent converters and reverters

In Chapter 4, the kmlShape algorithm [52] was used to identify which variables changed as a consequence of infection with *M.tb* in the recent converters. kmlShape is a clustering algorithm which aims to group longitudinal trajectories based on their shapes over time (Figure 4.6). The algorithm identified 67 variables in total that either increased or decreased post-infection. The trends of these 67 variables were then assessed in the QFT reverters.

For the 48 variables that increased post-conversion, median variable trajectories were calculated and plotted for the QFT reverters, and compared to the recent converters. The 8 variables that decreased post-conversion, however, were all the expressions of *M.tb*-lysate-specific phenotypic markers on total Th1 cells. All four time points were not available for these cells, but rather an average of the pre- and post-reversion time points were measured (Figure 3.4). The longitudinal trends of these variables could not be assessed, and, therefore, Wilcoxon's signed rank test was used to compare pre- and post-reversion time points to infer their trend over time.

6.2.3 Assessing longitudinal trends in the reverters

Wilcoxon's signed rank test (Section 4.2.1) was used to compare the paired pre- and post-reversion time points for the variables in the integrated dataset. The BH p-value adjustment protocol (Sec-

tion 4.2.1) was applied to correct for multiple comparisons, and significant differences between the paired samples were defined by an adjusted p-value less than 0.05. Any variable with a significant difference would then be tested as an immune signature that was associated with QFT reversion.

In addition, the kmlShape algorithm was applied to the median variable trajectories from the reverters. As discussed in Chapter 4, this method was found to be more sensitive to changes over time than Wilcoxon's paired test. Similar to the algorithm run on the recent QFT converters, the choice of k was taken to be three.

6.2.4 Comparing the reverters to the persistent QFT positives and negatives

Exploratory data analysis

The raw values of the QFT reverters, stratified according to QFT status, were compared to the two control cohorts. Median values were taken across all four longitudinal time points in the persistent QFT positive and negative cohorts, and at the pre- and post-reversion time points. Boxplots were used to visualize the reverters in relation to the controls and the groups were compared using Wilcoxon's rank sum unpaired test.

Principal component analysis

Consider an input vector $\mathbf{x} \in \mathbb{R}^p$. $\mathbf{x}$ can be written in terms of

$$\begin{aligned}\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} &= x_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \ldots + x_p \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \\ &= x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \ldots + x_p\mathbf{e}_p\end{aligned}$$

where $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ are orthonormal (orthogonal, or perpendicular along a line, and of unit length) vectors. PCA seeks an orthonormal basis $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_q$ such that $q < p$ and

$$\mathbf{x} \approx \hat{\mathbf{x}}(q) = b_1\boldsymbol{\phi}_1 + b_2\boldsymbol{\phi}_2 \ldots + b_q\boldsymbol{\phi}_q$$

and $\hat{\mathbf{x}}(q)$ has maximum variance. Then, $b_1, b_2, \ldots, b_q$ can be stored instead of $x_1, x_2, \ldots, x_p$ for each input of $\mathbf{x}$.

Let $\boldsymbol{\Sigma}$ (p x p) be the covariance matrix of an input matrix $\mathbf{X}^T$ (p x n) = $(\mathbf{x_1}, \ldots, \mathbf{x_n})$ that has been centred and scaled. It happens that the approximation error $E[|\mathbf{x} - \hat{\mathbf{x}}(q)|^2]$ is minimized when $\boldsymbol{\Sigma}\boldsymbol{\phi_i} = \lambda_i\boldsymbol{\phi_i}$ [40]. That is, the optimal orthonormal bases are the q eigenvectors of $\boldsymbol{\Sigma}$ corresponding to the largest eigenvalues, after the eigenvalue decomposition of the covariance matrix. Hence

$$\mathbf{x} \approx \hat{\mathbf{x}}(q) = \sum_{i=1}^{q} b_i\boldsymbol{\phi}_i$$

where $b_i = \boldsymbol{\phi}_i^T\mathbf{x}$ and $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_q$ are known as the PCs.

Alternatively, the PCs can be found by the singular value decomposition (SVD) [58], which generalizes the eigenvalue decomposition. The SVD decomposes a matrix $\mathbf{X}$ (n x p) into:

$$\mathbf{X} = \mathbf{USV}^T$$

where $\mathbf{U}$ (n x n) and $\mathbf{V}$ (p x p) are orthonormal matrices and are the left and right singular vectors respectively, and S (n x p) is a diagonal matrix that stores the singular values along the main diagonal. By taking the SVD of an input $\mathbf{X}$, the PCA model, which is given as $\mathbf{X} = \mathbf{TP}^T$, is found by setting $\mathbf{T} = \mathbf{US}$ and $\mathbf{P} = \mathbf{V}$. The matrix $\mathbf{T}$ (n x p) is known as the loading matrix and $\mathbf{P}$ (p x p) is the score matrix. The principal components $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_q$ then correspond to the q rows in $\mathbf{P}^T$ (the singular vectors) that have the q highest corresponding singular values, and the values along the main diagonal of S hold the variances of each principal component. These variances can be plotted in a scree plot to determine the optimal number of PCs that will explain as much of the variance of the original space as possible, while in a lower dimension.

We applied PCA to the vast-scaled (Section 3.5) and MFA-imputed (Section 3.6) integrated dataset to assess whether any unsupervised groupings of observations formed.

Projection to latent space-discriminant analysis

Let $\mathbf{Y}$ (n x k) be a multivariate matrix of continuous response variables, where n is the number of observations in the matrix and k is the number of dependent variables. Further, define $\mathbf{X}$ (n x p) to be a matrix of p predictor variables for each of the n observations. If each of the columns of $\mathbf{X}$ are linearly independent, in other words $\mathbf{X}$ is a matrix of full ranks, then performing multiple linear regression is sufficient. However, when $p >> n$, which was the case in this particular project, $\mathbf{X}$ is a singular matrix and multiple linear regression is no longer feasible. PLS regression is the preferred method,

PLS begins by iteratively decomposing $\mathbf{X}$ and $\mathbf{Y}$ into their latent structures. Let $\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F}$ and $\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$ be the basic decomposition for the matrices, where $\mathbf{T}$ (n x n) and $\mathbf{U}$ (n x n) are score matrices, $\mathbf{P}$ (p x n) and $\mathbf{Q}$ (k x n) are the matrices of loadings, and $\mathbf{E}$ (n x p) and $\mathbf{F}$ (n x k) are matrices of random errors. PLS regression aims to find a set latent vectors that performs a the simultaneous decomposition of both $\mathbf{X}$ and $\mathbf{Y}$, such that these components explain as much of the covariances between $\mathbf{X}$ and $\mathbf{Y}$ as possible. Therefore, the latent structure with the most variation in $\mathbf{Y}$, $\mathbf{u}_1$, is extracted and explained by the latent structure of $\mathbf{X}$, $\mathbf{t}_1$, that best explains $\mathbf{u}_1$. Note that $\mathbf{t}_1$ does not necessarily explain the most variation in $\mathbf{X}$, as with PCA. Consequently, $\mathbf{X}$ and $\mathbf{Y}$ are modelled using the same latent variables and hence will be a product of a common set of orthogonal factors, $\mathbf{T}$, and a set of specific loadings.

The score matrix $\mathbf{T}$, who's columns are the latent variables, can be expressed as the following decomposition $\mathbf{T} = \mathbf{XW}(\mathbf{P}^T\mathbf{W})^{-1}$. The matrix $\mathbf{W}$ (p x n) contains the regression coefficients from the regression of $\mathbf{X}$ on the latent variable $\mathbf{t}_h$. We can now express $\mathbf{Y}$ as

$$\begin{aligned}\mathbf{Y} &= \mathbf{UQ}^T + \mathbf{F} \\ &= \mathbf{TQ}^T + \mathbf{F} \\ &= \mathbf{XW}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q}^T + \mathbf{F}.\end{aligned}$$

The PLS regression coefficient, $\mathbf{B}$, is given as $\mathbf{XW}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q}^T$ and is a vector of length h, where h is the dimension of the PLS. A new sample, $\mathbf{X}_{\text{new}}$ is predicted as

$$\mathbf{Y}_{\text{new}} = \mathbf{X}_{\text{new}}\mathbf{B}.$$

PLS regression can be extended to a classification problem in a process known as PLS-DA. To get into a PLS-DA framework, the response vector of length n with c classification levels is simply recoded into a dummy matrix of size n x c to indicate class membership. To illustrate this, assume there is vector of responses with three classification levels, $c \in \{1, 2, 3\}$, the dummy response matrix

will be an n x 3 matrix as follows

$$\begin{bmatrix}1\\2\\3\end{bmatrix} \rightarrow \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}.$$

The PLS protocol is then performed as before and a new sample is assigned as the column index of the element with the largest predicted value in the row.

Prior to fitting the PLS-DA model to the four groups, which were the persistent negatives, the QFT- reverters, the QFT+ reverters and the persistent positives, LASSO feature selection was performed to identify the most stratifying features in the dataset. The LASSO method was introduced in Chapter 4 and the shrinkage parameter, λ, was defined as the average λ value across 500 repeats of 10-fold CV. The LASSO model was built to the integrated dataset with the control cohorts only as responses. This model would then select features that could stratify the two control groups and was blinded to the differences between reverters and the other two groups.

PLS-DA models were built using the R package ropls [127] once again to the vast standardized and MFA imputed dataset containing the features selected by the LASSO model. The group status, namely persistent QFT negative, QFT- reverter, QFT+ reverter and persistent QFT positive was then defined as the response.

The performances of the PLS-DA models built were then assessed via an internal validation procedure. MFA imputation was repeated for 1000 random bootstrapped samples of the data, where 70% of observations were set aside to make up training set, and the remaining 30% the testing set. The data was split into the training and testing sets such that there were equal proportions of the four groups in the testing set (10 observations per group). The PLS-DA model was then built to the training set of observations and the test data was used assess how well the model could predict the new outcome based on its training. The misclassification error was a natural choice for a performance metric, and the overall performance of the PLS-DA models were reported as the average misclassification error for each bootstrapped sample.

Similar to what was done in Chapter 5, LASSO, and hence PLS-DA models were built to the integrated dataset and then to the two single datasets separately. Based on each models' average misclassification errors, the results could be used to to conclude whether an integrated model could outperform or add to models built to the single datasets.

6.3 Results

6.3.1 Longitudinal trends in the QFT reverters

The kmlShape algorithm built to the variable trajectories in the recent converter cohort in Chapter 4 identified 51 variables that increased, and 16 that decreased post infection with *M.tb*. These variables are summarized in Tables 4.1 and 4.2.

For those variables that increased in the recent converters (Figure 6.2A(i)), median variable trajectories were plotted for the QFT reverters (Figure 6.2A(ii)). The trajectories and the mean trajectory superimposed onto the plot indicated that there was no specific trend, let alone a decreasing trend, over time. For the variables that decreased in the recent converters, no significant differences were found after adjusting for multiple comparisons in the reverters, and none of the variables had an increasing trend (Figure 6.2B(i) and (ii)).

Not restricting the analysis to just the variables that changed in the recent converters, paired tests

were performed on between pre- and post-reversion time points for all the variables in the integrated dataset. After adjusting for multiple comparisons, no significant differences were found. Further, when the kmlShape algorithm was run on the median variables trajectories from the reverters, the algorithm was unable to identify any specific trends over time (Figure 6.3).

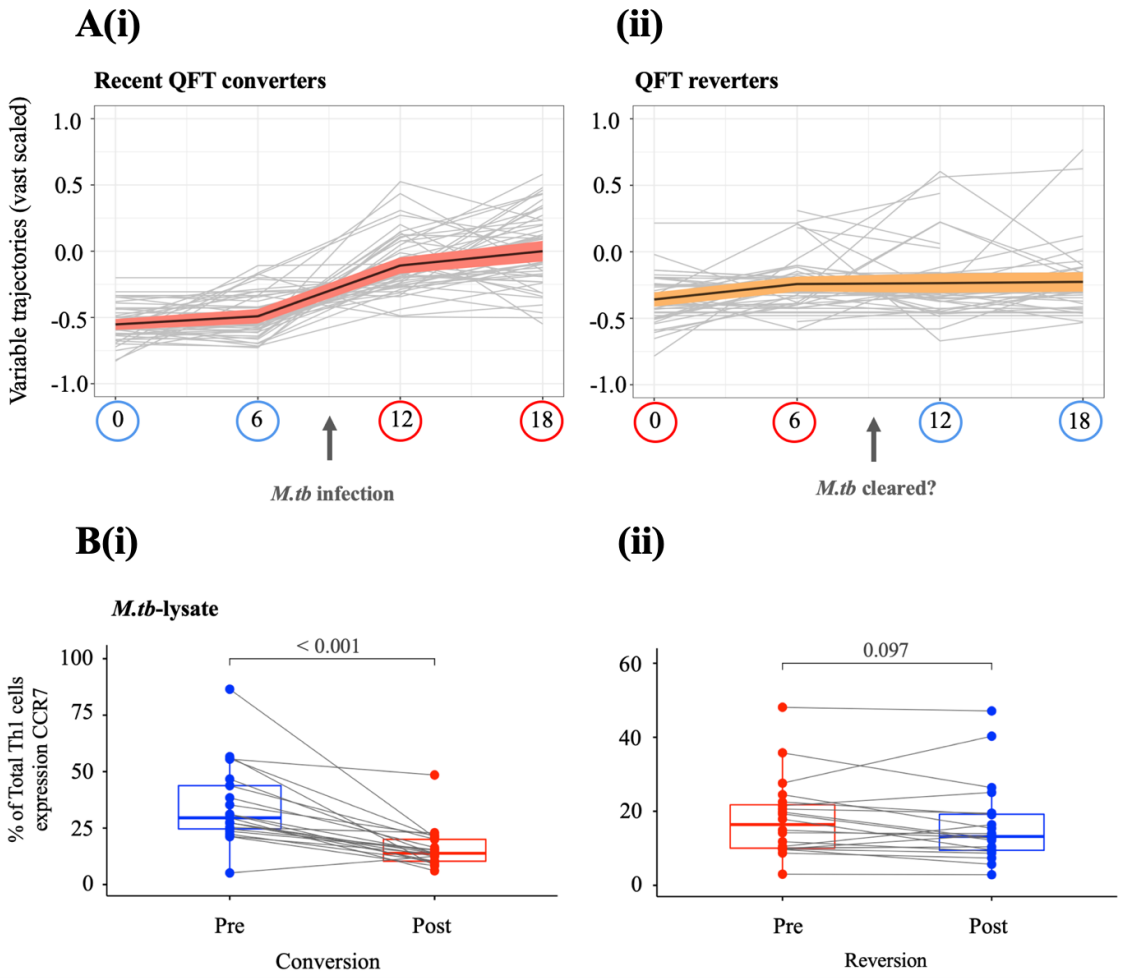

Figure 6.2: Longitudinal trends of the QFT reverters compared to the recent converters. (**A**) Variable trajectories of those cell subsets that increased post infection with *M.tb* identified by kmlShape plotted for (**i**) the recent QFT converters and (**ii**) the QFT reverters, and mean trajectories are superimposed onto the plots (red and orange respectively). (**B**) Wilcoxon's signed rank test was used to compare pre- and post- (**i**) conversion and (**ii**) reversion time points of *M.tb*-lysate-specific expression of CCR7 on total Th1 cells, one of the variables identified by kmlShape that decreased upon TBI. The resulting p-values from the test are superimposed onto the plots.

6.3.2 Relationship between QFT reverters and the control cohorts

Wilcoxon's ranked sum test was used to compare the raw values of the persistent negatives, QFT- reverters, QFT+ reverters and persistent positives. The different relationship patterns that formed between reverters and the control cohorts in particular was observed.

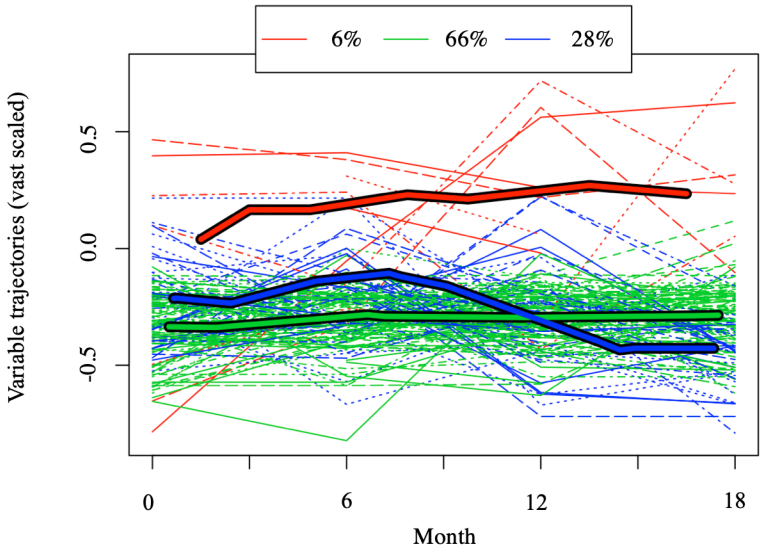

Figure 6.3: The three clusters identified by the kmlShape algorithm built to the reverters. The algorithm was unable to identify and cluster any variable trajectories according to their longitudinal trends over time.

In the majority of cell subsets, 71% of the variables, no significant differences were detected between groups. For 16% of the variables, the reverters were different from the persistent negatives but not from the persistent positives (Figure 6.4(i)). This pattern was observed most frequently in *M.tb*-specific CD4+ T cells, and all E6C10-specific cells were IFN-γ-. The three groups were significantly different from each other for 7% of the variables (Figure 6.4(ii)). In the remaining 6% of variables, the reverters were different from the persistent positives but not from the persistent negatives (Figure 6.4(iii)). Interestingly, the 6% included three IFN-γ+ E6C10-specific T cells subsets. In general, however, the reverters did have detectable E6C10- specific T cell responses, which indicates that they have been exposed to *M.tb*, but contrarily to what is measured by QFT these responses did not decrease upon reversion.

Since several variables were different in the reverters compared to control groups, we applied a PCA model to further explore the data in a more holistic way. The model was built to the integrated dataset and the two principal components captured 26% of the total variance. When the first two principal components were plotted against each other, there was no significant groupings of observations, and the groups appeared to overlap considerably (Figure 6.5A). A univariate plot of each group's loading scores on the first principal component indicated that this component was able to separate the persistent negatives from the other three groups (Figure 6.5B), but the total variance of this component was too low to draw any conclusions. Principal component two, on the other hand, was unable to separate any of the groups from each other.

The unsupervised nature of PCA was unable to significantly separate the groups from each other. Therefore, a PLS-DA modelling approach was more suitable to this dataset. Feature selection was performed prior to fitting the PLS-DA models using a LASSO model. The binomial LASSO model was blinded to the QFT reverters and identified seven variables from the adaptive dataset that could stratify the control cohorts from each other: *M.tb*-lysate-specific IL-2-CD107-CD154-IFN-γ+ TNF-, IL-2-CD107-CD154+IFN-γ-TNF+, IL-2+CD107-CD154+IFN-γ-TNF+ CD4+ T cells; E6C10-specific IL-2+CD107-CD154-IFN-γ+TNF+, IL-2+CD107-CD154+IFN-γ+TNF+, IL-2-CD107+CD154+IFN-γ+TNF+ CD4+ T cells; and the expression of CCR7 on *M.tb*-lysate-

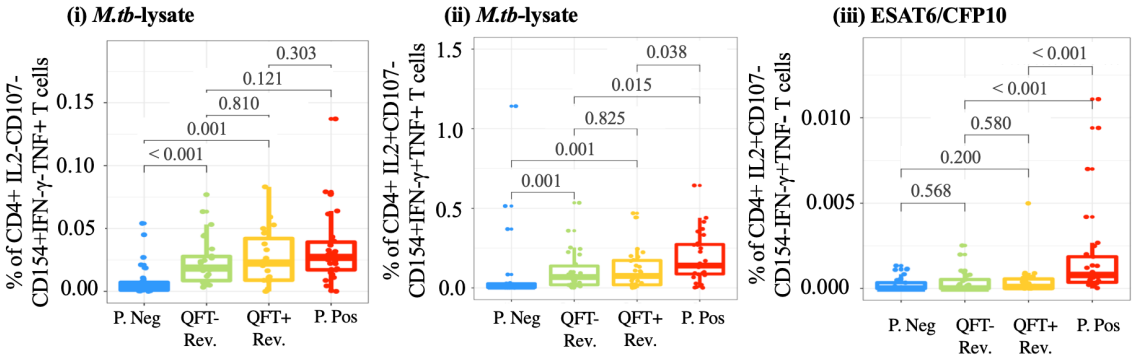

Figure 6.4: Comparisons between QFT- reverters (green), QFT+ reverters (yellow), persistent QFT negative (blue) and positive (red) groups. Examples of variables that were significantly different across (**i**) the reverters and the QFT negatives, but not the QFT positives (*M.tb*-lysate-specific CD4+IL2-CD107-CD154+IFN-γ-TNF+ T cells); or (**ii**) all 3 groups (*M.tb*-lysate-specific CD4+IL2+CD107-CD154+IFN-γ+TNF+ T cells); or (**iii**) the reverters and QFT positives, but not the QFT negatives (E6C10-specific CD4+IL2+CD107-CD154- IFN-γ+TNF- T cells).

specific total Th1 cells. The coefficients of these variables indicated that the *M.tb*-specific CD4+ T cell subsets were more associated with an individual being persistent QFT positive, while the expression of CCR7 on *M.tb*-lysate-specific total Th1 cells was associated with an individual being persistent QFT negative.

Using these seven feature-selected variables, a PLS-DA model was built with the four groups (persistent negative, QFT- reverter, QFT+ reverter, persistent positive) as a response. The results from the PLS-DA model are summarized in Figure 6.6. In three dimensions, the feature-selected PLS-DA model captured 80% of the total variance, with the first component capturing 50% (Figure 6.6A). Plotting the first two latent components against each other, the model was able to successfully separate the two control cohorts from each other. The QFT- and QFT+ reverter groups of observations overlapped significantly with each other and were positioned in between the two control groups (Figure 6.6B). The reverters, regardless of QFT status, also had intermediate loading scores on latent component one from the two controls and were significantly different from them (Figure 6.6C(i)). Figure 6.6C(ii) shows how the original variables contributed to creating latent component 1. The persistent positives loaded negatively on the first latent component and we can see that the key variables that drove this were were two IFN-γ+ E6C10-specific CD4+ T cells subsets. The persistent negatives, however, loaded positively on this first component, driven by the expression of CCR7 on *M.tb*-lysate-specific total Th1 cells, which also had a positive loading on the first component.

The average misclassification error after an internal validation procedure of this PLS-DA model was 0.48. On average, about half of the observations were misclassified: the persistent positives were misclassified 9% of the time, the persistent negatives 15% of the time, and the QFT- and QFT+ reverters misclassified 79% and 87% of the time, respectively. The reverters, therefore, regardless of QFT status, were more often misclassified than correctly classified, which explains the high overall misclassification rate. The average confusion matrix across the 1000 bootstrapped samples (Table 6.1) showed that roughly half of the observations from both reverter groups were classified as persistent negatives, which is twice as many as those that were classified as persistent positives, or correctly classified as reverters.

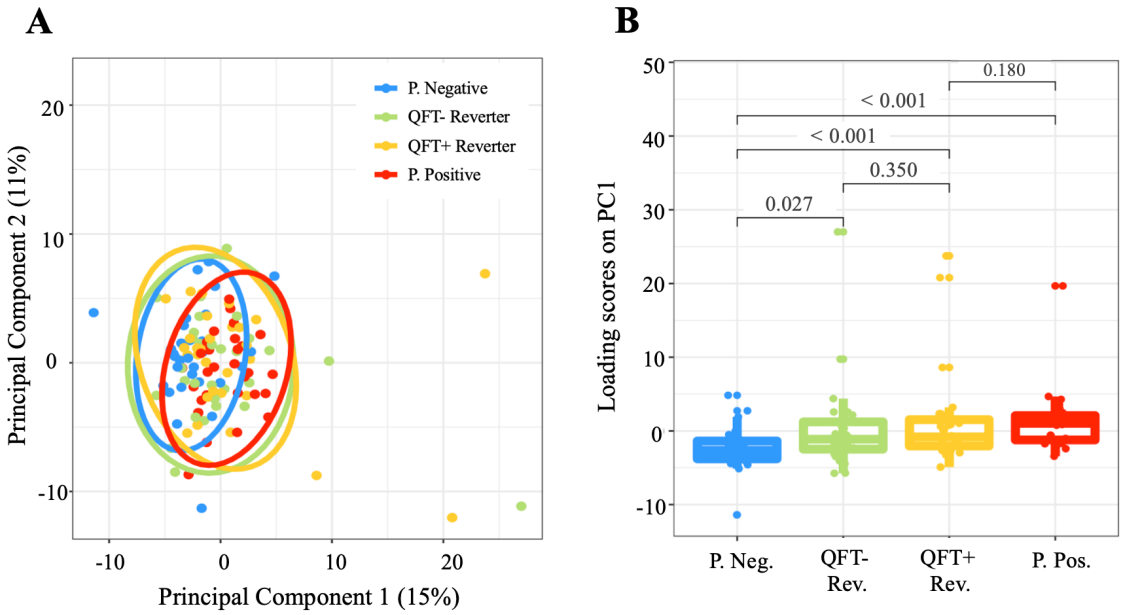

Figure 6.5: PCA model built to the integrated dataset. The two axes in (**A**) are the first and second principal components (PCs) from the model, and the total variance each component accounted for is shown as a percentage in brackets. Each dot is an observation and is colour coded according to which cohort they belong to: the persistent positives (blue), the QFT- reverters (green), the QFT+ reverters (yellow) and the persistent QFT positives (red). 95% confidence ellipses are shown for each group. (**B**) shows the loading scores of each group on the first principal component. Wilcoxon's sum rank test was then used to compare the groups and the resulting p-values are superimposed onto the plots.

As most reverters were getting misclassified as persistent negatives, we tested whether the classification performance of the model improved when the observations from the persistent negatives and reverters were combined into a "non-positive" group. The internal validation procedure of the PLS-DA model was re-run with persistent QFT positive and non-positive as new responses and, although still high, the misclassification error did improve to an average of 0.24.

Table 6.1: Internal validation of the PLS-DA model performance. Table shows an element-wise average of 1000 confusion matrices produced for each of the bootstrapped samples of the data during internal validation. The rows are the true outcomes in the testing set and the columns are the outcomes predicted by the model built to the training set.

		Predicted				
		P. Negative	QFT- Reverter	QFT+ Reverter	P. Positive	Total
True	P. Negative	8.5	1.2	0.3	0.0	10
	QFT- Reverter	4.9	2.8	1.2	1.1	10
	QFT+ Reverter	4.7	2.5	0.8	2.0	10
	P. Positive	0.0	0.6	0.4	9.0	10

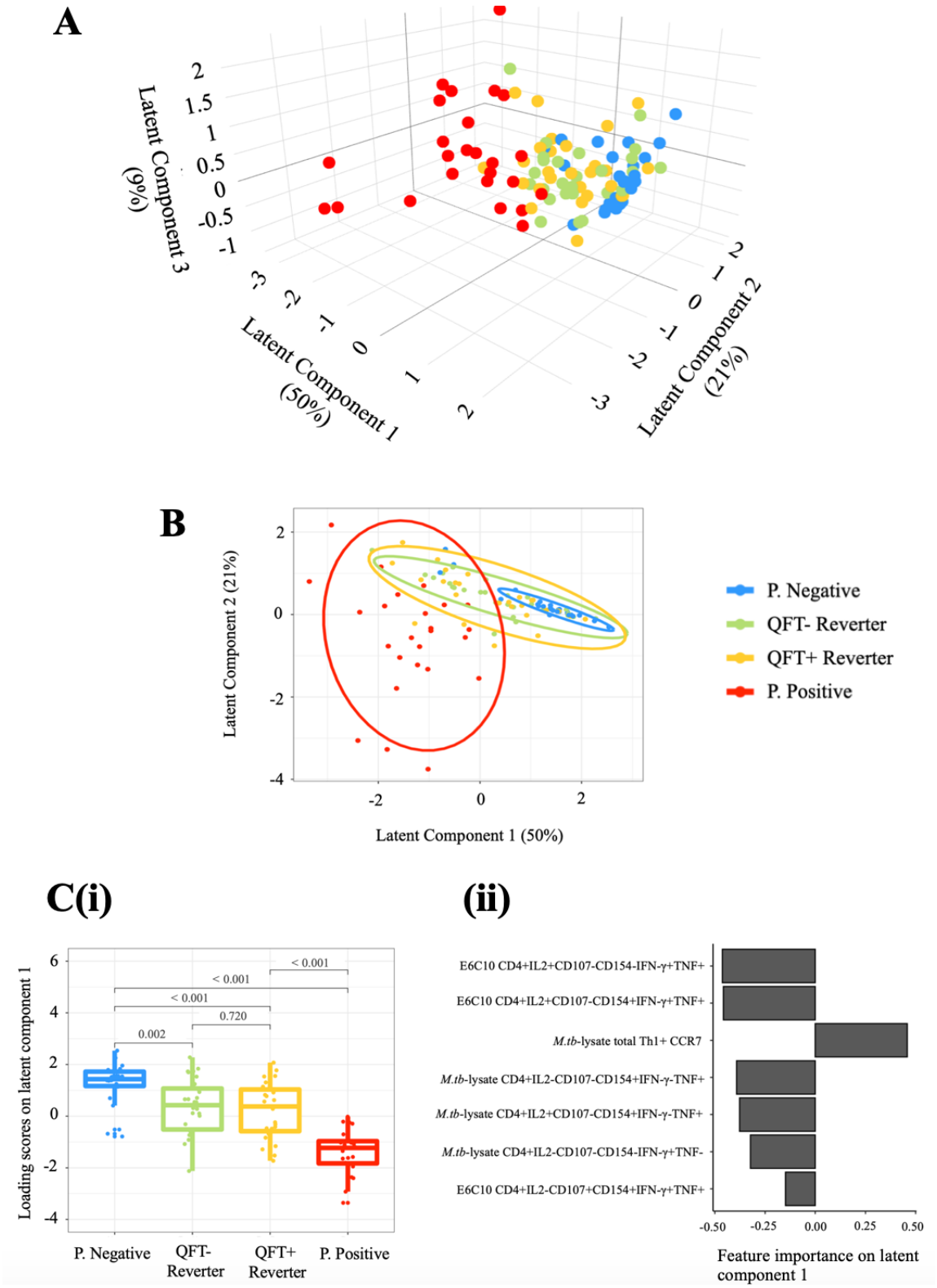
A
Latent Component 3 (9%)
Latent Component 1 (50%)
Latent Component 2 (21%)
B
Latent Component 2 (21%)
Latent Component 1 (50%)
P. Negative
QFT- Reverter
QFT+ Reverter
P. Positive
C(i)
Loading scores on latent component 1
< 0.001
< 0.001
< 0.001
0.720
0.002
P. Negative
QFT- Reverter
QFT+ Reverter
P. Positive
(ii)
E6C10 CD4+IL2+CD107-CD154-IFN-γ+TNF+
E6C10 CD4+IL2+CD107-CD154+IFN-γ+TNF+
M.tb-lysate total Th1+ CCR7
M.tb-lysate CD4+IL2-CD107-CD154+IFN-γ-TNF+
M.tb-lysate CD4+IL2+CD107-CD154+IFN-γ-TNF+
M.tb-lysate CD4+IL2-CD107-CD154-IFN-γ+TNF-
E6C10 CD4+IL2-CD107+CD154+IFN-γ+TNF+
Feature importance on latent component 1

Figure 6.6 *(previous page)*: PLS-DA model built to the LASSO feature selected variables. The three axes in (**A**) are the first, second and third latent components from the model, and the total variance each component accounted for is shown as a percentage in brackets. Each dot is an observation and is colour coded according to which cohort they belong to, the persistent negatives (blue), the QFT- reverters (green), the QFT+ reverters (yellow) and the persistent QFT positives (red). (**B**) Observations from latent component one plotted against latent component two with a corresponding 95% confidence ellipse for are shown for each group. (**Ci**) Loading scores of each group on component one. Wilcoxon's sum rank test was used to compare the groups and the resulting p-values are superimposed onto the plots. (**Cii**) Feature importance on latent component one.

The results outlined above are from the PLS-DA built to LASSO selected features from the integrated dataset. When the binomial LASSO model was fitted to the adaptive dataset only, the model identified the same seven variables that were found in the integrated LASSO model. Therefore, the performances of the integrated and adaptive PLS-DA models were identical. The PLS-DA model built to the LASSO selected features from the innaptive dataset, however, selected 23 variables and had a much larger average misclassifcation error of 0.64.

In addition to the PLS-DA model built to the seven LASSO selected features with persistent negative, QFT- reverter, QFT+ reverter and persistent positive group status as a response, three other PLS-DA models were built to further explore the observed relationships between the groups. The first model was built to the same selected features, but included the recent QFT converters, which were separated according to QFT status, as a response. The univariate plot of each group's loading scores on the first latent component from this PLS-DA model indicated that this component could successfully separate the groups based on their QFT status, except for the reverters. The reverters, once again, regardless of QFT status, showed intermediate scores between QFT- and QFT+ groups (Figure 6.7A).

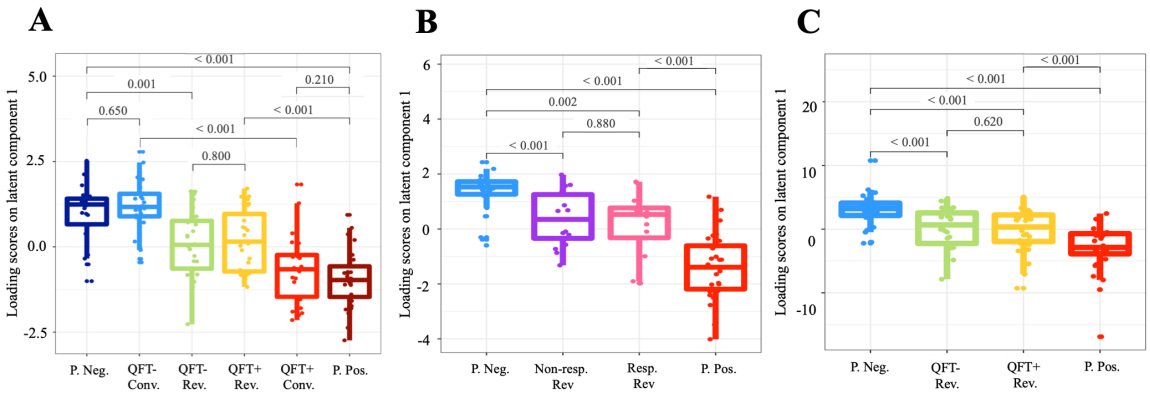

Figure 6.7: Loading scores of additional PLS-DA models built. Univariate boxplots of the loading scores on latent component 1 from the PLS-DA models built: (**A**) including the recent QFT converters, (**B**) with the reverters stratified according to responder status identified by MIMOSA, and (**C**) where no feature selection was performed prior to building the PLS-DA model. Wilcoxon's rank sum test was used to perform pairwise comparisons of the groups and the resulting p-values are superimposed onto the plots.

Since reverters were not separated according to QFT status, the second PLS-DA model was built

to assess whether the reverters could be stratified according to E6C10 responder status defined by MIMOSA (Section 3.4.1). Responding and non-responding reverters were still not significantly different from each other based on their loadings on the first latent component and were significantly different from the control cohorts (Figure 6.7B).

Lastly, to ensure that the relationships that have been observed between the reverters and the control groups in the various PLS-DA models were not a result of the selected features, a PLS-DA model was built to all the variables in the integrated dataset. The loading scores on component one are shown in Figure 6.7C. The relationship was the same, however the total variance captured by the first component was much lower (12%) and the average misclassification error was 0.77.

6.4 Discussion

Several modelling techniques and non-parametric tests were employed to better understand the immunological features associated with QFT reversion. QFT reversion was defined by two positive QFT test results, followed by two negative QFT test results. Given that the QFT test is the best indication of *M.tb* infection to date, a reasonable hypothesis to explain these findings is that QFT reversion could be associated with the clearance or control of *M.tb*. Since measuring actual *M.tb* presence (and therefore its clearance) in tissues of healthy individuals is not possible, we further hypothesized that if immune variables that significantly changed upon recent QFT conversion showed opposite trends during reversion, they may offer indirect evidence supporting *M.tb* clearance. If this hypothesis were true, immune signatures associated with QFT reversion could be tested as correlates of protection from *M.tb*. However, when we compared the longitudinal trends of the variables that changed as a result of TBI in the recent converters to the trends in reverters, the reverters did not change over time, nor was a trend seen that was opposite the QFT converters. Therefore, there was sufficient evidence to reject the hypothesis that reversion is associated with the clearance of *M.tb*.

Not restricting the analysis to just the variables that changed in the QFT converters, paired non-parametric tests were performed between pre- and post-reversion time points for all the variables in the integrated dataset. After adjusting for multiple comparisons, no significant differences between these two time points were found. Further, the kmlShape algorithm was unable to identify, nor cluster any of the variable trajectories according to any specific longitudinal trends. The algorithm was run with $k = 3$. This was the value used when the same algorithm was run on the recent converters, where we expected three groups, or clusters, to form based on results from exploratory data analysis. From exploratory data analysis on the reverters, however, we did not expect any specific clusters to form and $k = 3$ was therefore arbitrary. Different values for k ($k \in \{2, 3, 4\}$) were tested, but still no specific longitudinal trends were identified in the reverters. Hence, it is unlikely that immune features in the QFT reverters change over time, and no specific cell subset seemed to be associated with reversion.

When compared to the QFT- and QFT+ control groups, raw values of several variables were distinct in the reverters, suggesting that QFT reversion may identify a distinct group of individuals. To explore these relationships further, a PCA model was built followed by a PLS-DA model with LASSO selected features. The PCA model built to the integrated dataset captured little of the total variability and the unsupervised nature of the model was unable to separate out the groups of observations. PLS-DA, a supervised method, was therefore more suitable for the data. Feature selection was performed before building the PLS-DA model using LASSO regularized regression. This was done to improve the computational efficiency of the model and to identify the most stratifying features in the dataset. A LASSO model built to the integrated dataset, with the control groups as a response, identified seven features from the adaptive dataset. Quite fittingly, the E6C10-specific

CD4+ T cells that were chosen by the LASSO models were all IFN-γ+, which is precisely how the persistent positives were defined. We blinded the model to the differences between the reverters and the controls to validate the true relationship between the reverters and the persistent QFT positives and negatives.

In three components, the PLS-DA model captured 80% of the total variability. The first latent component, which accounted for 50% of the total variability, could successfully separate the observations from the two control groups from each other. The reverters, regardless of QFT status, showed an intermediate profile between the control groups. The LASSO model was blinded to the reverters, and the features were chosen such that they could discriminate the control groups from each other. It was interesting, therefore, that the model did not identify the reverters as either one of the controls, and could not distinguish between the QFT+ and QFT- reverter groups.

To confirm that the model could successfully stratify QFT- from QFT+ individuals, a PLS-DA model was built using the same seven selected features and included the recent QFT converters, stratified according to QFT status, as a response. Despite the LASSO model also being blinded to the converter group, QFT- and QFT+ time points in the converters were indistinguishable from the persistent negatives and positives, respectively. The first latent component, and hence the selected features, could successfully separate the groups according to QFT status, with the reverters as an exception. These results suggest that reverters may be a biologically distinct group of individuals, who have different immune responses to true QFT+ and QFT- individuals.

Since reverters were not separated according to their QFT status (pre- and post-reversion) another PLS-DA model was built to test whether other immune features could account for the intermediate phenotype of this group. According to the MIMOSA's responder status, about half of the reverters did not respond to E6C10 stimulation (similarly to QFT- individuals), and half did (similarly to QFT+ individuals). When looking at the loading scores on component one, there were no differences between responders and non-responders, and the reverters still showed intermediate scores between the control groups.

Lastly, to ensure any relationships seen between the reverters and the controls was not a consequence of the feature selection, a PLS-DA model was built to all the variables in the integrated dataset, and the relationships were unchanged.

The various PLS-DA models built provide strong evidence to suggest that the reverters are a biologically distinct group of individuals with intermediate features between QFT+ and QFT- control groups, rather than unique features specific for the reverter group. However, when the feature-selected PLS-DA model built to the the persistent QFT positive, persistent QFT negative, QFT- reverters and QFT+ reverters groups as responses was validated via an internal validation procedure, more than half of the reverters were misclassified. Further, it appeared that the majority of the observations in both reverter cohorts (QFT- and QFT+) were classified as persistent negatives, more than they were correctly classified as reverters or misclassified as persistent QFT+ individuals. These findings may suggest that immune responses measured in reverters are indeed more similar to persistent QFT- rather than QFT+ individuals, indirectly supporting the hypothesis that they have cleared or at least controlled *M.tb* infection. However, given the small sample size of the dataset, using the misclassification rate is a misleading measure of performance and is difficult to draw definitive conclusions from.

The LASSO model built to the integrated dataset and adaptive dataset selected features from the adaptive dataset. No variable from the innaptive dataset was chosen in the integrated LASSO model. This could be because adaptive features, specifically IFN-γ production from T cells in

response to E6C10, were used to define the cohorts. However, the PLS-DA model built to the innaptive LASSO selected features only had a much higher average misclassification error compared to the model built to the adaptive features. The model was also unable to separate the control cohorts from each other. The immune features from the adaptive dataset alone were therefore once again sufficient for the analysis, suggesting that innaptive features are not permanently affected by *M.tb* infection and therefore don't contribute to the group stratification.

6.5 Conclusion

We report a comprehensive and thorough analysis of immunological features in QFT reverters, a group of individuals little is known about. We concluded that immune responses in reverters do not change over time, and that no specific immune feature was associated with reversion. Individuals from the QFT reverter group did have detectable T cell responses, which indicates that they have been exposed to *M.tb*, but these responses did not significantly change upon reversion. Given that, overall, the reverters appeared more similar to persistent QFT negative rather than QFT positive individuals, the hypothesis that QFT reversion may be associated with clearance of *M.tb* infection merits further and more definitive investigations.

Chapter 7

Overall discussion and conclusions

All the studies reported in this document focus on a unique longitudinal cohort of adolescents (Adolescent Cohort Study, ACS, described under preliminary studies) established at the South African TB Vaccine Initiative (SATVI), the world largest dedicated TB vaccine clinical site, located in a high TB burden setting and equipped with state of the art immunology facilities. This project explored a number of multivariate modeling and clustering techniques to analyze the immune response to *M.tb* infection, and was based on the integration of multiple datasets. By integrating the datasets we hoped to better understand the interplay between features from different arms of the immune response, which, to our knowledge, has not been done before in the context of TB.

Prior to any data analysis, several pre-processing steps had to be addressed. In Chapter 3, we stressed the importance of trying a number of different methods to find the best pre-processing method for a given dataset, such that unbiased and valid results are yielded. A large amount of time was spent on the data pre-processing steps for this project, due to the complexities and differences of the datasets, to ensure that they would not affect or bias any results. The first pre-processing step involved filtering the datasets to identify and remove biologically irrelevant cell subsets, and to decrease the number of variables in the integrated dataset. COMPASS, a Bayesian hierarchical model, was applied to filter the CD4+ and CD8+ T cell responses. T cell subsets with less than ten observations at any of the sampling occasions with posterior probabilities greater than 0.1 were removed. Due to the high background (unstimulated) values in cell types from the innaptive dataset, COMPASS could not be used to filter this dataset, and we had to design a novel filtering approach, which we believe can be used in future studies. After the various filtering methods were applied, we were satisfied that we had preserved all biologically meaningful cell subsets in the final integrated dataset.

The next step included standardizing the integrated dataset, such that the values from the different data types were comparable. After trying and contrasting a number of standardization and normalization techniques on the raw data, we found that a method known as vast scaling performed the best. Vast scaling, which is achieved by multiplying the Z-score by the mean divided by the standard deviation, retained the original densities of the raw data values. It did this without inflating any of the responses, and brought the values between, and within, the two datasets to a similar and more comparable range.

As a final pre-processing step, missing values were accounted for using an MFA-based imputation method. Similar to the standardization step, we implemented a variety of imputation methods and compared their performances based on how well they could preserve the densities of the raw data values. Because we assumed the data was MCAR, we had no reason to have any other aim but to replicate the data. We were pleased with the performance of MFA imputation. A potential lim-

itation, however, was not fitting models using a multiple imputation framework and using Rubin's rules to account for between and within-imputation variability. However, we argue that because we were less interested in estimating model coefficients than identifying predictive markers, that simply repeating the imputation for every CV run was sufficient. This way we could ensure that any results found were not a consequence of the imputation method used, and we were confident that it did not affect or bias the results found in the dissertation.

Chapter 4 focused on identifying and defining immunological features that changed after recent acquisition of *M.tb* infection. We explored a number of clustering algorithms, as an alternative to employing simple, non-parametric tests, on the recent converter cohort. The recent converter cohort was defined by two negative QFT results followed by two positive QFT results. kmlShape, an extension of the k-means algorithm, which aims to cluster longitudinal trajectories based on their shapes, identified 55 variables that changed as a result of *M.tb* infection. When comparing the results from kmlShape to applying simple non-parametric paired tests, we observed that kmlShape was more sensitive to changes over time. This is likely because the kmlShape algorithm could successfully capture time granularity that was masked by aggregating the time points for the paired tests.

Due to the high dimensionality of the data, and the fact that there were only four repeated measures per individual, we were unable to successfully apply longitudinal modeling techniques to the recent converter cohort, and there is certainly scope for this in future studies. However, to our knowledge, the results from Chapter 4 provided the most extensive list of immunological features to date that could be useful to understand early biological events after *M.tb* infection.

Based on reports in Chapter 5, which compared immune responses in individuals with established *M.tb* infection to individuals that were recently infected, we identified HLA-DR on E6C10-specific T cells as a strong biomarker for recent *M.tb* infection. This biomarker was found by applying a number of regularized regression models, including a multiple tuning parameter elastic net (MTP-EN) model and standard EN models to the data, and was statistically validated using machine learning algorithms. We also concluded that features from the innaptive immune response were poor predictors of recent infection, and did not improve the performance of the integrated model. A limitation in this chapter was, however, the small sample size of the dataset. We recommended that further validation in an independent cohort with a larger sample size is required for conclusive results.

In Chapter 6, we explored the phenomenon of QFT reversion. The QFT reverters were identified by two positive QFT results followed by two negative QFT results, and very little is actually known about this group of individuals. From performing simple non-parametric tests, we concluded that immune features in reverters do not change over time, and that no specific immune feature is associated with QFT reversion. Two mulitvariate methods, namely PCA and PLS-DA, were then used to explore the relationship between the reverters and the persistent QFT positive and negative cohorts. PCA was unable to successfully separate out the groups and captured little of the total variability in the data. Non-linear PCA methods were also tested, specifically kernel PCA, however this did not improve the ability of the model to successfully separate the groups. We found that a LASSO feature selected PLS-DA model worked best on the data. The model could successfully separate the control cohorts from each other, and the reverters had intermediate profiles. However, after an internal validation of the PLS-DA model, it appeared that the reverters, regardless of QFT status, were more similar to persistent QFT negatives than QFT positive individuals.

From the results reported in Chapter 6, we believe there is sufficient evidence to conclude that the reverters do not change over time, and are either a biologically distinct group of individuals, or

have cleared *M.tb* infection. There is immense potential for future studies on QFT reversion, and I believe the best way to fully understand reversion is to compare them to true QFT positive and QFT negative control cohorts in a larger study.

The main overall limitation of this project, however, was the fact that the two data types, namely the adaptive and innaptive datasets, were not generated consistently. As a result, median values had to be taken for each individual, stratified according to cohort and QFT status. Non-parametric tests indicated that there were no significant differences where the median values were taken, however, it meant that we could not explore the longitudinal nature of the data or use all the data points available. As discussed in Chapter 4, aggregating time points can mask important data trends. Further, not generating the datasets consistently also resulted in a larger degree of missingness in the data, as not all of the same participants and time points were measured in each data type. For future data integration projects of this nature, it is therefore strongly recommended that the datasets be generated in a consistent manner for successful and holistic analyses.

In conclusion, this dissertation represents the most comprehensive analysis of the immune response in QFT converters and reverters to date. Despite data challenges, I do believe each of the study aims were addressed adequately and in as much detail as possible. Findings from this dissertation will make a significant contribution to the TB field.

Bibliography

[1] Abdi, H. Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley* (2010).

[2] Adekambi, T., Ibegbu, C., Cagle, S., Kalokhe, A., Wang, Y., et al. Biomarkers on patient T cells diagnose active tuberculosis and monitor treatment response. *Journal of Clinical Investigation* (2015).

[3] Akaike, H. Information theory and an extension of the maximum likelihood principle. *Akademiai Kiado* (1973), 267–281.

[4] Allen, M., Bailey, C., Cahatol, I., Dodge, L., Yim, J., et al. Mechanisms of control of *Mycobacterium tuberculosis* by NK cells: role of glutathione. *Frontiers Immunology 6* (2015), 508.

[5] Andersen, P., and Scriba, T. Moving tuberculosis vaccines from theory to practice. *Nature Reviews Immunology* (2019).

[6] Andrews, J., Hatherill, M., Mahomed, H., Hanekom, W., Campo, M., et al. The dynamics of QuantiFERON-TB gold in-tube conversion and reversion in a cohort of South African adolescents. *American Journal of Respiratory and Critical Care Medicine 191*, 5 (2015), 584–91.

[7] Andrews, J., Nemes, E., Tameris, M., Landry, B., Mahomed, H., et al. Serial QuantiFERON testing and tuberculosis disease risk among young children: an observational cohort study. *The Lancet: Respiratory Medicine 5*, 4 (2017), 282–90.

[8] Ault, R., Headley, C., Hare, A., Carruthers, B., Mejias, A., and Turner, J. Blood RNA signatures predict recent tuberculosis exposure in mice, macaques and humans. *Scientific Reports 10*, 16873 (2020).

[9] Axelsson-Robertson, R., Loxton, A., Walzl, G., Ehlers, M., Zumla, A., et al. A broad profile of co-dominant epitopes shapes the peripheral *Mycobacterium tuberculosis*-specific CD8+ T-cell immune response in South African patients with active tuberculosis. *PLoS 8* (2013), 3.

[10] Batushansky, A., Toubiana, D., and Fait, A. Correlation-based network generation, visualization, and analysis as a powerful tool in biological studies: A case study in cancer cell metabolism. *Big Data and Network Biology* (2016).

[11] Bean, A., Roach, D., Briscoe, H., France, M., Korner, H., et al. Structural deficiencies in granuloma formation in TNF gene-targeted mice underlie the heightened susceptibility to aerosol *Mycobacterium tuberculosis* infection, which is not compensated for by lymphotoxin. *Journal of Immunology 162* (1999), 3504–11.

[12] Behr, M., Edelstein, P., and Ramakrishnan, L. Revisiting the timetable of tuberculosis. *Analysis* (2018).

[13] Bender, R., and Lange, S. Adjusting for multiple testing-when and how? *J. Clin. Epidemiol. 54* (2001), 343–349.

[14] Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JSTOR 57*, 1 (1995), 289–300.

[15] Berkson, J. Application of the logistic function to bio-assay. *Journal of the American Statistical Association 39*, 227 (1944), 357–365.

[16] Bermudez, L., Wu, M., and Young, L. Interleukin-12-stimulated natural killer cells can activate human macrophages to inhibit growth of *Mycobacterium avium*. *Infectious Immunology 63* (1995), 4099–104.

[17] Betts, M., Brenchley, J., Price, D., De Rosa, S., Douek, D., et al. Sensitive and viable identification of antigen-specific CD8+ T cells by a flow cytometric assay for degranulation. *J. Immunol. Methods 281*, 1-2 (2003), 65–78.

[18] Blanchard, D., Stewart, W., Klein, T., Friedman, H., and Djeu, J. Cytolytic activity of human peripheral blood leukocytes against legionella pneumophila-infected monocytes: characterization of the effector cell and augmentation by interleukin 2. *Journal Immunology 139* (1987), 551–6.

[19] Borgström, E., Fröberg, G., Correia-Neves, M., Atterfelt, F., Bellbrant, J., et al. CD4+ T cell proliferative responses to PPD and CFP-10 associate with recent *M. tuberculosis* infection. *Tuberculosis 123*, 101959 (2020).

[20] Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefe, M., et al. On modularity clustering. *IEE Transactions on Knowledge and Data Engineering 20*, 2 (2009), 172–188.

[21] Breiman, L. Random forests. *Machine Learning 45*, 5 (2001), 5–32.

[22] Breiman, L., Cutler, A., Liaw, A., and Wiener, M. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*, 2018. R package version 4.6-14.

[23] Breiman, L., Friedman, J., Stone, C., and R.A., O. *Classification and regression trees*. Wadsworth Books, 1984.

[24] Brier, G. Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78* (1950), 1–3.

[25] Brigitte, E., and Jerome, P. *Analyses factorielles simples et multiples; objectifs, méthodes et interprétation*. Dunod, 2008.

[26] Caccamo, N., Guggino, G., Meraviglia, S., Gelsomino, D., and Paolo, T. Analysis of *Mycobacterium tuberculosis*-specific CD8 T-cells in patients with active tuberculosis and in individuals with latent infection. *PLoS 4* (2009), 5.

[27] Caruso, A., Serbina, N., Klein, E., Triebold, K., Bloom, B., and Flynn, J. Mice deficient in CD4 T cells have only transiently diminished levels of IFN-γ, yet succumb to tuberculosis. *Journal Immunology 162*, 9 (2005), 5407–16.

[28] Chakravarty, S., Zhu, G., Tsai, M., Mohan, V., Marino, S., et al. Tumor necrosis factor blockade in chronic murine tuberculosis enhances granulomatous inflammation and disorganizes granulomas in the lungs. *Infectious Immunology 76* (2008), 919–26.

[29] Chattopadhyay, P. K., Yu, J., and Roederer, M. Live-cell assay to detect antigen-specific CD4+ T-cell responses by CD154 expression. *Nature Protocols 1*, 1 (2006).

[30] Cheng, C., Wang, B., Gao, L., Liu, J., Chen, X., Huang, H., et al. Next generation sequencing reveals changes of the gammadelta T cell receptor repertoires in patients with pulmonary tuberculosis. *Sci Rep 8* (2018), 2956.

[31] Cho, S., Mehra, V., Thoma-Uszynski, S., Stenger, S., Serbina, N., et al. Antimicrobial activity of MHC class i-restricted CD8+ T cells in human tuberculosis. *Proc Natl Acad Sci USA 97*, 22 (2000), 12210–5.

[32] Cohen, J. The earth is round (p<.05). *American Psychologist 49* (1994), 997–1003.

[33] Csárdi, G. e. a. *igraph: Network Analysis and Visualization*, 2020. R package version 1.2.5.

[34] Cyktor, J., Carruthers, B., Stromberg, P., Pircher, H., and Turner, J. Killer cell lectin-like receptor g1 deficiency significantly enhances survival after *Mycobacterium tuberculosis* infection. *Infection and Immunity 81*, 4 (2013), 1090–1099.

[35] Day, C., Abrahams, D., Lerumo, L., Janse van Rensburg, E., Stone, L., et al. Functional capacity of *Mycobacterium tuberculosis*-specific T cell responses in humans is associated with mycobacterial load. *J. Immunol. 187*, 5 (2011), 2222–2232.

[36] De Jong, R., Brouwer, M., Hooibrink, B., Van der Pouw-Kraan, T., Miedema, F., and Van Lier, R. The CD27- subset of peripheral blood memory CD4+ lymphocytes contains functionally differentiated T lymphocytes that develop by persistent antigenic stimulation in vivo. *Eur. J. Immunol. 22* (1992), 993–999.

[37] Deusch, K., Luling, F., Reich, K., Classen, M., Wagner, H., and Pfeffer, K. A major fraction of human intraepithelial lymphocytes simultaneously expresses the gamma/delta T cell receptor, the CD8 accessory molecule and preferentially uses the V delta 1 gene segment. *European Journal of Immunology 21* (1991), 1053–9.

[38] Dharmadhikari, A., Basaraba, R., Van Der Walt, M., Weyer, K., Mphahlele, M., et al. Natural infection of guinea pigs exposed to patients with highly drug-resistant tuberculosis. *Tuberculosis 91*, 4 (2011), 329–38.

[39] Ding, Y., Ma, F., Wang, Z., and Li, B. Characteristics of the Vdelta2 CDR3 sequence of peripheral gammadelta T cells in patients with pulmonary tuberculosis and identification of a new tuberculosis-related antigen peptide. *Clin Vacc Immunol 22* (2015), 761–768.

[40] Eckart, C., and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika 1* (1936), 211–218.

[41] Efron, B., and Tibshirani, R. *An introduction to the Bootstrap*. Chapman and Hall, 1993.

[42] Elm, E., Altman, D., Egger, M., Pocock, S., Gotzsche, P., et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med. 4* (2007), 1–5.

[43] Esin, S., and Batoni, G. Natural killer cells: a coherent model for their functional role in *Mycobacterium tuberculosis* infection. *Journal of Innate Immunity 7* (2015), 11–24.

[44] Fabri, M., Stenger, S., Shin, D., Yuk, J., Liu, P., et al. Vitamin D is required for IFN-γ mediated antimicrobial activity of human macrophages. *Sci. Transl. Med.* (2011).

[45] Finak, G., McDavid, A., Chattopadhyay, P., Dominguez, M., De Rosa, S., et al. Mixture models for single-cell assays with applications to vaccine studies. *Biostatistics 15*, 1 (2014), 87–101.

[46] Flynn, J. Immunology of tuberculosis and implications in vaccine development. *Tuberculosis 84*, 1 (2004), 93–101.

[47] Flynn, J., and Chan, J. Immunology of tuberculosis. *Ann Rev Immunol. 19* (2001), 93–129.

[48] Fréchet, M. Sur quelques points du calcul fonctionne. *Rendiconti del Circolo Matematico di Palermo 22*, 1 (1906), 1–72.

[49] Friedman, J., et al. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, 2019. R package version 3.0-2.

[50] Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association 32*, 200 (1937), 675–701.

[51] Geldmacher, C., Ngwenyama, N., Schutez, A., Petrovas, C., Reither, K., et al. Preferential infection and depletion of *Mycobacterium tuberculosis*-specific CD4 T cells after HIV-1 infection. *Journal of Experimental Medicine 13* (2010), 2869–81.

[52] Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S., and Subtil, F. kmlShape: An efficient method to cluster longitudinal data (time-series) according to their shapes. *PLOS ONE 11*, 6 (2016), 1–24.

[53] Genolini, C., and Guichard, E. *kmlShape: K-Means for Longitudinal Data using Shape-Respecting Distance*, 2016. R package version 0.9.5.

[54] Geurts, P., Ernst, D., and Wehenkel, L. Extremely randomized trees. *Machine Learning 63* (2006), 3–42.

[55] Gold, M., Cerri, S., Smyk-Pearson, S., Cansler, M., Vogt, T., Delepine, J., et al. Human mucosal associated invariant T cells detect bacterially infected cells. *PLoS Biology* (2010).

[56] Gold, M., Eid, T., Smyk-Pearson, S., Eberling, Y., Swarbrick, G., Langley, S., et al. Human thymic MR1-restricted MAIT cells are innate pathogen-reactive effectors that adapt following thymic egress. *Mucosal Immunology 6* (2013), 35–44.

[57] Gold, M., McLaren, J., Reistetter, J., Smyk-Pearson, S., Ladell, K., et al. MR1-restricted MAIT cells display ligand discrimination and pathogen selectivity through distinct T cell receptor usage. *Journal of Experimental Medicine 211* (2014), 1601–10.

[58] Golub, G., and Reinsch, C. Singular value decomposition and least squares solutions. *Numerische Mathematik* (1970).

[59] Halliday, A., Whitworth, H., Kottoor, S., Niazi, U., Menzies, S., et al. Stratification of latent *Mycobacterium tuberculosis* infection by cellular immune profiling. *Journal of Infectious Disease 215* (2017), 1480–7.

[60] Hanley, J., and McNeil, B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143* (1982), 29–36.

[61] Harrell, F. *Regression Modeling Strategies*. Springer Series in Statistics, 2001.

[62] Hawn, T., Day, T., Scriba, T., Hatherill, M., Hanekom, W., et al. Tuberculosis and prevention of infection. *Microbiology and Molecular Biology Reviews 78*, 4 (2014), 650–71.

[63] Henao-Tamayo, M., Irwin, S., Shang, S., Ordway, D., and Orme, I. T lymphocyte surface expression of exhaustion markers as biomarkers of the efficacy of chemotherapy for tuberculosis. *Tuberculosis 91* (2011), 308–313.

[64] Hoerl, A., and Kennard, R. Ridge regression. *Encyclopaedia of Statistical Sciences 8* (1988), 129– 136.

[65] Honaker, J., King, G., and Blackwell, M. Amelia II: A program for missing data. *Journal of Statistical Software 45*, 7 (2011), 1–47.

[66] Houben, R., and Dodd, P. The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. *PloS Medicine* (2016).

[67] Hu, Z., Zhao, H., Li, C., Liu, X., Barkan, D., et al. The role of KLRG1 in human CD4+ T-cell immunity against tuberculosis. *J Infect Dis. 217*, 9 (2018), 1491–1503.

[68] Husson, F., and Josse, J. Handling missing values in multiple factor analysis. *Quality and Preferences 30*, 2 (2013), 77–85.

[69] Husson, F., and Josse, J. *missMDA: Handling Missing Values with Multivariate Data Analysis*, 2017. R package version 1.16.

[70] James, G., Witten, D., Hastie, T., and Tibshirani, R. *An introduction to statistical learning*. Springer, 2013.

[71] Jenum, S., Grewal, H., Hokey, D., Kenneth, J., Vaz, M., et al. The frequencies of IFNγ+IL2+TNFα+ PPD-specific CD4+CD45R0+ T-cells correlate with the magnitude of the QuantiFERON Gold in-tube response in a prospective study of healthy Indian adolescents. *PLoS* (2014).

[72] Jiang, J., Wang, X., Wang, X., et al. Reduced CD27 expression on antigen-specific CD4+ T cells correlates with persistent active tuberculosis. *J. Clin. Immunol. 30* (2010), 566–573.

[73] Joosten, S., Ottenhoff, T., Lewinsohn, D., Hoft, D., Moody, D., and Seshadri, C. Harnessing donor unrestricted T-cells for new vaccines against tuberculosis. *Vaccines 37* (2019), 3022–30.

[74] Kaufmann, S. Recent findings in immunology give tuberculosis vaccines a new boost. *Trends in Immunology 26*, 12 (2005), 660–7.

[75] Keun, H., Ebbels, T., Antti, H., Bollard, M., Beckonert, O., et al. Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Anal Chim Acta. 490* (2003), 265–276.

[76] Koay, H., Gherardin, N., Enders, A., Loh, L., Mackay, L., et al. A threestage intrathymic development pathway for the mucosal-associated invariant T cell lineage. *Nature Immunology 17* (2016), 1300–11.

[77] Kuo, L., and Mallick, B. Variable selection for regression models. *The Indian Journal of Statistics 60*, 1 (1998), 65–81.

[78] Latorre, I., Fernandez-Sanmartin, M., Muriel-Moreno, B., Villar-Hernandez, R., Vila, S., et al. Study of CD27 and CCR4 markers on specific CD4+ T-cells as immune tools for active and latent tuberculosis management. *Frontiers in Immunology* (2019).

[79] Lê Cao, K., Boitard, S., and Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics 12*, 1 (2011), 253.

[80] Lin, L., Finak, G., Ushey, K., Seshadri, S., Hawn, T., et al. COMPASS identifies T-cell subsets correlated with clinical outcomes. *Nat Biotechnol. 33*, 6 (2015), 610–616.

[81] Liu, J., et al. Data integration by multi-tuning parameter elastic net regression. *BMC Bioinformatics 19*, 1 (2018), 369.

[82] Machingaidze, S., Verver, S., Mulenga, H., Abrahams, D., Hatherill, M., et al. Predictive value of recent QuantiFERON conversion for tuberculosis disease in adolescents. *American Journal of Respiratory and Critical Care Medicine 186*, 10 (2012), 1051–6.

[83] MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proc. 5th Berkeley Symp. Math. Statistics and Probability. 1* (1967), 281–297.

[84] Maglione, P., Xu, J., and Chan, J. B cells moderate inflammatory progression and enhance bacterial containment upon pulmonary challenge with *Mycobacterium tuberculosis*. *Journal Immunology 178* (2007), 7222–34.

[85] Mahomed, H., Hawkridge, T., Verver, S., Abrahams, D., Geiter, L., et al. The tuberculin skin test versus QuantiFERON TB Gold in predicting tuberculosis disease in an adolescent cohort study in South Africa. *PLoS One 6*, 3 (2011), e17984.

[86] Martinez-Camblor, P., Pérez-Fernández, S., and Díaz-Coto, S. The role of the p-value in the multitesting problem. *Journal of Applied Statistics 47*, 9 (2020), 1529–1542.

[87] Metcalfe, J., Everett, C., Steingart, K., Cattamanchi, A., Huang, L., et al. Interferon-gamma release assays for active pulmonary tuberculosis diagnosis in adults in low- and middle-income countries: systematic review and meta-analysis. *Journal of Infectious Disease 204* (2011), 1120–9.

[88] Mills, C., O'Grady, F., and Riley, R. Tuberculin conversion in the "naturally infected" guinea pig. *Bill Johns Hopkins Hospital 106* (1960), 36–45.

[89] Mosteller, F., and Tukey, J. Data analysis, including statistics. *Handbook of Social Psychology* (1968).

[90] Mpande, C., et al. *Mycobacterium tuberculosis*-specific T cell activation identifies individuals at high risk of tuberculosis disease. *medRxiv* (2020).

[91] Musvosvi, M., Duffy, D., Filander, E., et al. T-cell biomarkers for diagnosis of tuberculosis: candidate evaluation by a simple whole blood assay for clinical translation. *European Respiratory Journal 51* (2018).

[92] Nemes, E., Geldenhuys, H., Rozot, V., Rutkowski, K., Ratangee, F., et al. Prevention of *M. tuberculosis* infection with H4:IC31 vaccine or BCG revaccination. *The New England Journal of Medicine 379*, 2 (2018), 138–49.

[93] Nemes, E., Rozot, V., Geldenhuys, H., Bilek, N., Mabwe, S., et al. Optimization and interpretation of serial QuantiFERON testing to measure acquisition of *Mycobacterium tuberculosis* infection. *American Journal of Respiratory and Critical Care Medicine 196*, 5 (2017), 638–48.

[94] Nemeth, J., Rumetshofer, R., Winkler, H., Burghuber, O., Müller, C., and Winkler, S. Active tuberculosis is characterized by an antigen specific and strictly localized expansion of effector T cells at the site of infection. *Eur. J. Immunol.* (2012).

[95] Neyman, J., and Pearson, E. *Joint Statistical Papers of J. Neyman and E. S. Pearson.* Cambridge University Press, 1967.

[96] Nikitina, I., Kondratuk, N., Kosmiadi, G., et al. *Mtb*-specific CD27low CD4 T cells as markers of lung tissue destruction during pulmonary tuberculosis in humans. *PLoS One 7* (2012), e43733.

[97] O'Garra, A., Redford, P., McNab, F., Bloom, C., Wilkinson, R., et al. The immune response in tuberculosis. *Annual Review Immunology 31* (2013), 475–527.

[98] Organization, W. H. Global tuberculosis report.

[99] Pai, M., Denkinger, C., Kik, S., Rangaka, M., Zwerling, A., et al. Gamma interferon release assays for detection of *Mycobacterium tuberculosis* infection. *Clinical Microbial Review 27*, 1 (2014), 3–20.

[100] Pasparakis, M., Alexopoulou, L., Douni, E., and Kollias, G. Tumour necrosis factors in immunoregulation: everything that's interesting is ... new! *Cytokine Growth Factor Review 7* (1996), 223–29.

[101] Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine 2*, 11 (1901), 559–572.

[102] Peters, W., and Ernst, J. Mechanisms of cell recruitment in the immune response to *Mycobacterium tuberculosis*. *Microbes Infection 5* (2003), 151–58.

[103] Phuah, J., et al. Effects of B cell depletion on early *Mycobacterium tuberculosis* infection in cynomolgus macaques. *Infectious Immunology 84* (2016), 1301–11.

[104] Riley, R., Mills, C., Nyka, W., Weinstock, N., Storey, P., et al. Aerial dissemination of pulmonary tuberculosis. a two-year study of contagion in a tuberculosis ward. *American Journal of Hygiene 70* (1995), 2.

[105] Riou, C., Berkowitz, N., Goliath, R., Burgers, W., and Wilkonson, R. Analysis of the phenotype of *Mycobacterium tuberculosis*-specific CD4+ T cells to discriminate latent from active tuberculosis in HIV-uninfected and HIV-infected individuals. *Frontiers Immunology 8*, 968 (2017), 1–11.

[106] Riou, C., Du Bruyn, E., Ruzive, S., Goliath, R., Lindestam Arlehamn, C., et al. Disease extent and anti-tubercular treatment response correlates with *Mycobacterium tuberculosis*-specific CD4 T-cell phenotype regardless of HIV-1 status. *Clinical & Translational Immunology 9*, 9 (2020), e1176.

[107] Rothman, K. J. No adjustments are needed for multiple comparisons. *Epidemiology 1*, 1 (1990), 43–46.

[108] Rozot, V., Vigano, S., Mazza-Stalder, J., Idrizi, E., Day, C., et al. *Mycobacterium tuberculosis*-specific CD8+ T cells are functionally and phenotypically different between latent infection and active disease. *Eur. J. Immunol. 43*, 6 (2013), 1568–1577.

[109] Rubin, D. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons, 2004.

[110] Ruhwald, M., deThurah, L., Kuchaka, D., Zaher, M., Salman, A., et al. Introducing the ESAT-6 free IGRA, a companion diagnostic for TB vaccines based on ESAT-6. *Scientific Reports* (2017).

[111] Russell, D., Cardona, P., Kim, M., Allain, S., and Altare, F. Foamy macrophages and the progression of the human tuberculosis granuloma. *Nature Immunology 10* (2009), 943–8.

[112] Sakai, S., Kauffman, K., Oh, S., Nelson, C., Barry, C. E., et al. MAIT cell-directed therapy of *Mycobacterium tuberculosis* infection. *Mucosal Immunology* (2020).

[113] Sankoh, A., Huque, M., and Dubey, S. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Stat. Med. 16* (1997), 2529–2542.

[114] Schuetz, A., Haule, A., Reither, K., et al. Monitoring CD27 expression to evaluate *Mycobacterium tuberculosis* activity in HIV-1 infected individuals *in vivo*. *PLoS One 6* (2011), e27284.

[115] Seder, R., Darrah, P., and Roederer, M. T-cell quality in memory and protection: implications for vaccine design. *Nature Reviews Immunology 8* (2008), 247–58.

[116] Sester, M., Sotfiu, G., Lange, C., Giehl, C., Girardi, E., et al. Interferon-g release assays for the diagnosis of active TB: a systematic review and meta-analysis. *European Respiratory Journal 27* (2011), 100–11.

[117] Shen, L., Frencher, J., Huang, D., Wang, W., Yang, E., et al. Immunization of Vγ2Vδ2 T cells programs sustained effector memory responses that control tuberculosis in nonhuman primates. *PNAS* (2019).

[118] Silveria-Mattos, P., Barreto-Duarte, B., Vasconcelos, B., Fukutani, K., Vinhaes, C., et al. Differential expression of activation markers by *Mycobacterium tuberculosis*-specific CD4+ T cell distinguishes extrapulmonary from pulmonary tuberculosis and latent infection. *Clinical Infectious Diseases* (2019).

[119] Sologuren, I., et al. Partial recessive IFN-γ R1 deficiency: genetic, immunological and clinical features of 14 patients from 11 kindreds. *Human Molecular Genetics 20*, 8 (2011), 1509–23.

[120] Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol. 15* (1904), 72–101.

[121] Spencer, C., Abate, G., Blazevic, A., and Hoft, D. Only a subset of phosphoantigen-responsive 9 2T cells mediate protective tuberculosis immunity. *Journal Immunology 181* (2008), 4471–84.

[122] Stekhoven, D. *missForest: Nonparametric Missing Value Imputation using Random Forest*, 2013. R package version 1.4.

[123] Stekhoven, D., and Buhlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics 28*, 1 (2012), 112–118.

[124] Stenger, S., and Modlin, R. An antimicrobial activity of cytolytic T cells mediated by granulysin. *Science 282* (1998), 5386.

[125] Streitz, M. Tesfa, L., Yildirim, V., et al. Loss of receptor on tuberculin-reactive T-cells marks active pulmonary tuberculosis. *PLoS One 2* (2007), e735.

[126] Therneau, T., Atkinson, B., and Ripley, B. *rpart: Recursive Partitioning and Regression Trees*, 2019. R package version 4.1-15.

[127] Thevenot, E. *ropls: PCA, PLS(-DA) and OPLS(-DA) for multivariate analysis and feature selection of omics data*, 2020. R package version 1.22.0.

[128] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society 58*, 1 (1996), 267–88.

[129] Torgo, L. *DMwR: Functions and data for "Data Mining with R"*, 2015. R package version 0.4.1.

[130] Toulon, A., Breton, L., Taylor, K., Tenenhaus, M., Bhavsar, D., et al. A role for human skin-resident T cells in wound healing. *Journal of Experimental Medicine 206* (2009), 743–50.

[131] Trafimow, D., and Marks, M. Editorial. *Basic Appl. Soc. Psych. 37* (2015), 1–5.

[132] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., et al. Missing value estimation methods for DNA microarrays. *Bioinformatics 17*, 6 (2001), 520–525.

[133] van den Berg, R., Hoefsloot, H., Westerhuis, J., Smilde, A., and van der Werf, M. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics 7* (2006), 142.

[134] Vickers, M., Darboe, F., Muefong, C., Mbayo, G., Barry, A., et al. Monitoring anti-tuberculosis treatment response using analysis of whole blood *Mycobacterium tuberculosis* specific T cell activation and functional markers. *Front. Immunol.* (2020).

[135] Voillet, V., Besse, P., Liaubet, L., San Cristobal, M., and González, I. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics 17*, 1 (2016), 402.

[136] Walzl, G., Ronacher, K., Hanekom, W., Scriba, T., and Zumla, A. Immunological biomarkers of tuberculosis. *Nature Reviews 11* (2011), 343–54.

[137] Wasserstein, R., and Lazor, N. ASA's statement on p-values: context, process and purpose. *Am. Stat. 70* (2016), 129–133.

[138] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 6 (1945), 80–83.

[139] Wilkinson, K., et al. Activation profile of *Mycobacterium tuberculosis*-specific CD41 T cells reflects disease activity irrespective of HIV status. *American Journal of Respiratory and Critical Care Medicine 192*, 11 (2016), 1307–10.

[140] Williams, M., Tyznik, A., and Bevan, M. Interleukin-2 signals during priming are required for secondary expansion of CD8+ memory T cells. *Nature 441*, 7095 (2006), 890–3.

[141] Winje, B., White, R., Syre, H., Skutlaberg, D., Oftung, F., et al. Stratification by interferon-γ release assay level predicts risk of incident TB. *Thorax 0* (2018), 1–10.

[142] Wold, H. Partial least squares. *Encyclopedia of the Statistical Sciences 6* (1982), 1–53.

[143] Wold, H. Estimation of principal components and related models by iterative least squares. *Multivariate Analysis* (1996), 391–420.

[144] Xi, X., Han, X., Li, L., and Zhao, Z. Identification of a new tuberculosis antigen recognized by gammadelta T cell receptor. *Clin Vacc Immunol 20* (2013), 530–539.

[145] Yoneda, T., and Ellner, J. CD4(+) T cell and natural killer cell-dependent killing of *Mycobacterium tuberculosis* by human monocytes. *American Journal of Respiratory and Critical Care Medicine 158* (1998), 395–403.

[146] Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. 67*, 2 (2005), 301–320.