**UNIVERSITY OF SOUTHAMPTON**

FACULTY OF MEDICINE

Clinical and Experimental Sciences

# Proteomic Discovery and Validation of Diagnostic Plasma Biomarkers for Pulmonary Tuberculosis

https://doi.org/10.5258/SOTON/T0009

by

**Diana Jazmín Garay Baquero**

0000-0002-9450-8504

Thesis for the degree of Doctor of Philosophy

November 2018

# ABSTRACT

*Despite more than a century fighting against tuberculosis, the World Health Organisation has estimated that around 1.7 million people died of tuberculosis in 2016 and over a quarter of the world's population is infected (1). One of the critical hurdles for stopping tuberculosis transmission is early and effective diagnosis of patients with the active pulmonary disease. Although important innovations in molecular diagnosis have been recently developed (e.g. Xpert MTB/RIF, Cepheid Inc., USA), there are no suitable tests for population screening at point-of-care (2, 3). The current tuberculosis diagnosis pipeline presents a highly variable performance and requires access to reference laboratory facilities (3). A non-sputum based rapid test with high specificity and sensitivity could save ~400,000 lives per year (4). Therefore, new biomarkers for diagnosis are urgently required for identifying patients with early symptoms and to expedite treatment. Variable sensitivity and specificity can be overcome using a combination of multiple biomarkers (5). Proteins, as ultimate biological effectors, are ideal candidates for diagnostic biomarkers; consequently, proteomic studies are a crucial platform for biomarker discovery in tuberculosis. This work aims to develop a multi-marker panel for tuberculosis diagnosis with high performance capable of differentiating tuberculosis patients from relevant controls. Quantitative Multidimensional Protein Identification Technology (qMudPIT) is applied for biomarker discovery identifying candidates for early diagnosis of tuberculosis. The multidimensional method optimised in this work led to the identification of 5022 plasma proteins and 3577 quantified proteins using iTRAQ labelling. Known and completely novel markers for active tuberculosis in plasma were identified including a peptide derived from Mycobacterium tuberculosis. Complementary statistical and bioinformatic analysis were applied to prioritise candidates for validation in one or two independent cohorts. The plasma proteomic profile here described represents a power strategy for biomarker discovery and the panel proposed has the potential to be translated to a rapid test and which might contribute to tuberculosis control.*

## TABLE OF CONTENTS

**TABLE OF CONTENTS** .......... 7

**LIST OF FIGURES** .......... 11

**ACADEMIC THESIS: DECLARATION OF AUTHORSHIP** .......... 17

**ACKNOWLEDGMENT OF CONTRIBUTION** .......... 19

**CHAPTER 1** .......... 21

**Overview of PhD** .......... 21

**1.1 Tuberculosis** .......... 24

1.1.1 Current Diagnostic Pipeline .......... 27

1.1.2 Urgency for Novel Diagnostic Tools .......... 31

**1.2 Biomarkers and Tuberculosis Diagnosis** .......... 32

1.2.1 Biomarker Definitions and Classification .......... 32

1.2.2 Current Screening Biomarker Pipeline .......... 34

1.2.3 Biomarker Validation and Qualification .......... 36

1.2.4 Biomarkers in Tuberculosis Diagnosis .......... 39

**1.3 Plasma Proteomics** .......... 42

1.3.1 The Depletion Dilemma .......... 45

1.3.2 MudPIT strategy .......... 47

1.3.2.1 Size Exclusion Chromatography (SEC) .......... 47

1.3.2.2 Reverse Phase Chromatography RP-HPLC .......... 48

1.3.2.3 Orthogonality in Multidimensional separations .......... 49

1.3.3 Current Status of Plasma/Serum Proteomics in Tuberculosis .......... 50

**1.4 Shotgun Proteomics** .......... 53

1.4.1 Enzymatic Digestion .......... 54

1.4.2 Mass Spectrometry and Tandem Mass Spectrometry ($MS^2$) .......... 58

1.4.2.1 Electrospray Ionisation (ESI) .......... 60

1.4.2.2 Fragmentation Methods .......... 61

1.4.2.3 LTQ-Orbitrap Technology .......... 62

1.4.3 Quantitative Proteomics .......... 63

1.4.3.1 Label-free Quantification .......... 63

1.4.3.2 Stable-isotope Labelling .......... 63

1.4.4 Data Analysis .......... 68

1.4.4.1 Protein Inference .......... 69

1.4.4.2 Statistical Significance of Protein Identification: FDR and PEP .......... 70
**HYPOTHESIS** .......... 73
**CHAPTER 2** .......... 75
**2.1 General Experimental Design** .......... 75
2.1.1 Sample Size .......... 76
2.1.2 Sample Collection .......... 78
**2.2 Patient Recruitment and Ethics Statement** .......... 80
**2.3 Sample Collection and Processing** .......... 80
**2.4 Reagents and Chemicals** .......... 80
**2.5 Plasma proteomics analysis** .......... 81
2.5.1 Multidimensional Protein Identification Technology (MudPIT) Analysis .......... 81
2.5.2 High Performance Size Exclusion Chromatography .......... 81
2.5.3 Dialysis Purification .......... 82
2.5.4 Protein Quantification .......... 82
2.5.5 Trypsin Digestion .......... 82
2.5.6 iTRAQ labelling .......... 82
2.5.7 Peptide Prefractionation with Offline alkaline RP-HPLC .......... 82
2.5.8 LC-FT-Orbitrap MS Analysis .......... 83
**2.6 MS Data Processing** .......... 83
2.6.1 Statistical Pipeline for iTRAQ based quantification .......... 84
**2.7 ELISA and Luminex validation** .......... 85
2.7.1 ELISA .......... 85
2.7.2 Luminex .......... 86
**2.8 Contribution statement** .......... 86
**CHAPTER 3** .......... 89
**3.1 Introduction** .......... 89
**3.2 Methods** .......... 90
3.2.1 Patients Cohort .......... 90
3.2.2 Sample Processing and Labelling .......... 91
3.2.3 Peptide Fractionation .......... 91
3.2.4 LC-FT-Orbitrap MS Analysis .......... 92
3.2.5 Bioinformatics and Statistical Analysis .......... 92
3.2.6. Solid Phase Extraction Cleaning Protocol (SPE) and C4 Chromatography .......... 92
**3.3 Results** .......... 93
3.3.1 Evaluation of Missingness .......... 93
3.3.2 Quality Control and Normalisation at Peptide Level .......... 94
3.3.3 Partial Differential Plasma Proteome in Active Pulmonary Tuberculosis .......... 96

3.3.4 Size Exclusion Chromatography Optimisation ........ 98
3.3.5 Dialysis Purification and Protein Quantification ........ 101
3.3.6 Solid Phase Extraction Cleaning Protocol and C4 Chromatography ........ 101
**3.4 Discussion** ........ 105
**CHAPTER 4** ........ 111
**4.1 Introduction** ........ 111
**4.2 Methods** ........ 112
4.2.1 Patient Cohort ........ 112
4.2.2 Experimental Design ........ 112
4.2.3 Sample Processing ........ 113
4.2.4 Data Analysis ........ 114
**4.3 Results** ........ 115
4.3.1 Prefractionation of Plasma Samples by Size Exclusion Chromatography ........ 115
4.3.2 Peptide Fractionation by C4 nUPLC ........ 119
4.3.3 Partial Profiling of Tuberculosis Plasma Proteome ........ 121
4.3.3.1 Analysis of independent segments: Exploration of the plasma proteome ........ 121
4.3.3.2 Multiconsensus report: an integrated evaluation of the plasma proteome ........ 126
4.3.3.3 Statistical assessment of proteomic data ........ 127
4.3.3.5 Functional analysis of the plasma proteome in active tuberculosis ........ 131
**4.2 Discussion** ........ 133
**CHAPTER 5** ........ 139
**5.1 Introduction** ........ 139
**5.2 Methods** ........ 140
5.2.1 Patient Cohort ........ 140
5.2.2 Experimental Design ........ 140
5.2.3 Sample processing and data analysis ........ 141
5.2.4 Validation of prioritised candidates ........ 142
**5.3 Results** ........ 144
5.3.1 Peptide Fractionation of Segment Four by C4 HPLC ........ 144
5.3.2 In-Depth Analysis of Plasma Subproteome of Pulmonary Tuberculosis Infection ........ 145
5.3.3 Selection of Candidates for Validation ........ 147
5.3.4 Validation of Selected Candidates in Independent Cohorts ........ 151
5.3.5 A potential marker from *Mycobacterium tuberculosis* ........ 163
5.3.6 Bioinformatic mining of proteomic profile from segment four ........ 166
**5.4 Discussion** ........ 175
**CHAPTER 6** ........ 185
**6.1 The most comprehensive plasma proteomic profile of pulmonary tuberculosis to date** . 186

**6.2 Towards a universal biosignature for pulmonary tuberculosis**.......................................... 187

**6.3 A novel multi-marker panel for pulmonary tuberculosis diagnosis** .................................. 189

**6.5 Publications and participation in congresses** ................................................................ 192

**ANNEXES**........................................................................................................................ 195

**Annexe 1. Proteins significantly modulated (p < 0.05) common to preliminary experiments 1 and 2.**.................................................................................................................................. 195

**Annexe 2. Summary of proteins significantly regulated from independent analysis of SEC segments**............................................................................................................................ 198

**Annexe 3. Summary of proteins significantly regulated from multiconsensus report and *PL3* analysis** ............................................................................................................................ 209

**Annexe 4. Proteins significantly regulated from segment 4 fully quantified and common to sets A, B and C by *PL2***........................................................................................................ 214

**Annexe 5. Proteins significantly regulated from segment 4 fully quantified and common to sets A, B and C by *PL3***........................................................................................................ 216

**Annexe 6. Common proteins with reported signatures**............................................................ 225

**REFERENCES** ................................................................................................................ 227

## LIST OF FIGURES

Figure 1. Tuberculosis global incidence ..................... 22
Figure 2. Tuberculosis pathogenesis ..................... 26
Figure 3. Tuberculosis natural history and diagnostic tools ..................... 28
Figure 4. Screening algorithm in cases where chest radiography and Xpert MTB/RIF are not available ..................... 30
Figure 5. Overview of biomarker categorisation ..................... 33
Figure 6. Biomarker development pipeline ..................... 34
Figure 7. Common biomarker failure sources and possible solutions ..................... 37
Figure 8. Proteome abundance distribution. ..................... 43
Figure 9. Protein dynamic range in human plasma ..................... 44
Figure 10. Size Exclusion chromatography principle. ..................... 47
Figure 11.  Graphical representation of orthogonality in an LC x LC system. ..................... 49
Figure 12.  Proteomic strategies. ..................... 54
Figure 13. Trypsin. ..................... 56
Figure 14. TCEP reaction mechanism. ..................... 57
Figure 15. MMTS reaction mechanism ..................... 58
*Figure 16. MS-based proteomics: General workflow and available instrumentation.* ..................... 59
Figure 17. Representation of electrospray ionisation process. ..................... 60
Figure 18. Schematic representation of CID fragmentation ..................... 61
Figure 19. Schematic representation of Orbitrap Elite. ..................... 62
Figure 20. General chemical structure of iTRAQ reagents. ..................... 65
Figure 21. Illustration of iTRAQ quantification using MS/MS. ..................... 66
Figure 22. Effect of isotopic purity on intensity. ..................... 67
Figure 23. Reporter ion quantification workflow ..................... 68
Figure 24. FDR and PEP for statistical assessment of PSMs ..................... 71
Figure 25. General workflow for TB diagnosis biomarkers discovery. ..................... 75
Figure 26. Power analysis simulation for proteomic analysis ..................... 76
Figure 27. Experimental strategy for discovery ..................... 77
Figure 28. Multidimensional protein identification technology (MudPIT) analysis ..................... 81
Figure 29.  Statistical and bioinformatic pipelines ..................... 84
Figure 30. iTRAQ labellig scheme of preliminary study. ..................... 91
Figure 31. Experimental design for SPE cleaning and C4 separation method development ..................... 93
Figure 32. Exploration of missingness ..................... 94

Figure 33. Quality control analysis of raw data intensities ........................................................95
Figure 34. Experiment comparison at protein level: clustering, correlation and statistical assessment ..........................................................................................................................................97
Figure 35. Common proteins differently modulated between experiments 1 and 2....................98
Figure 36. Optimisation of chromatographic parameters for size exclusion fractionation (SEC)100
Figure 37. Solid phase extraction protocol for iTRAQ/TMT labelled peptides........................102
Figure 38. C4 Peptide fractionation.........................................................................................104
Figure 39. Experimental design active pulmonary tuberculosis plasma proteome profile........113
Figure 40. Daily SEC calibration curves..................................................................................116
Figure 41. SEC prefractionation of plasma samples ................................................................118
Figure 42. Quality assessment of collection time-points in SEC separations ...........................119
Figure 43. C4 HPLC chromatographic traces of iTRAQ labelled peptides .............................120
Figure 44. Percentage of co-isolation excluding peptides from quantification.........................121
Figure 45. iTRAQ labelling efficiency segments 1 to 4...........................................................122
Figure 46. Data exploration at peptide level of plasma proteome............................................124
Figure 47. Data exploration at protein level of plasma proteome ............................................125
Figure 48. Multiconsensus report: coverage and dynamic range of profiled proteins...............127
Figure 49. Comparison of four statistical pipelines for assessment of significant protein regulation ........................................................................................................................................128
Figure 50. Variability relative quantification of 20 shared proteins across independent experiments ........................................................................................................................................129
Figure 51. Differential plasma proteome driven by active tuberculosis...................................130
Figure 52. Gene ontology enrichment of the plasma proteome in active tuberculosis.............132
Figure 53. Correlation network analysis of plasma proteome associated to pulmonary tuberculosis ........................................................................................................................................133
Figure 54. Experimental design for depth profiling of plasma segment 4 in pulmonary tuberculosis context ..........................................................................................................................................141
Figure 55. C4 HPLC chromatographic traces of iTRAQ labelled peptides – Segment 4 .........144
Figure 56. General metrics for the plasma profile resulting from segment 4............................145
Figure 57. Data distribution and clustering from sets A to C...................................................146
Figure 58. Differential expressed proteins common among sets A to C from segment 4.........148
Figure 59. General workflow for antibody-based validation of selected candidates................152
Figure 60. Assessment of candidates by ELISA on validation subset: Proteins with no significant differences ......................................................................................................................................153
Figure 61. Proteins validated by ELISA on MIMIC cohort.....................................................155
Figure 62. Proteins evaluated on MIMIC validation cohort by Luminex .................................157
Figure 63. Proteins evaluated on South African validation cohort by Luminex .......................158
Figure 64. Receiver operator characteristic curves for validated candidates ...........................160

Figure 65. Mass spectrometric evidence of an Mtb derived peptide........................................ 165

Figure 66. Batch effect correction........................................................................................ 167

Figure 67. WGCNA analysis of plasma proteomic profile resulting from segment 4.............. 168

Figure 68. Independent WGCNA analysis of Peruvian and South African samples................ 170

Figure 69. Strongly correlated proteins to tuberculosis infection: module turquoise ............... 173

Figure 70. Gene ontology enrichment analysis for strongly correlated proteins to tuberculosis infection ........................................................................................................................... 175

## LIST OF TABLES

Table 1. Clinical meaning of parameters of biomarkers performance. 39
Table 2. Minimum requirements for the ideal point-of-care diagnostic test for tuberculosis 50
Table 3. Serum proteomic studies for novel biomarkers in active pulmonary tuberculosis 51
Table 4. General description of iTRAQ experiments for comprehensive plasma profile 78
Table 5. Criteria for inclusion and exclusion of participating individuals 79
Table 6.  Clinical information of individuals participating in pilot study 91
Table 7. Comparison IR quantification system and Nanodrop performance 101
Table 8. Clinical information of individuals participating in the proteome profiling study 112
Table 9. Clinical information of individuals participating in the proteome profiling study 140
Table 10. Validation cohort from South Africa 143
Table 11. Validation cohort from UK (MIMIC Study) 143
Table 12. Common proteins to PL2, PL3 and Biolayout significantly modulated by tuberculosis infection in segment 4 149
Table 13. Proteins selected for validation from segments 1, 2 and 3. 151
Table 14. Dilution factors for validation of selected proteins 152
Table 15. Coordinates of the ROC curves 161

UNIVERSITY OF Southampton

## ACADEMIC THESIS: DECLARATION OF AUTHORSHIP

I, Diana Jazmín Garay Baquero, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Proteomic Discovery and Validation of Diagnostic Plasma Biomarkers for Pulmonary Tuberculosis

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as: [please list references below]:

Signed: Diana Jazmín Garay Baquero

Date: 19/11/2018

# ACKNOWLEDGMENT OF CONTRIBUTION

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

My heartily gratitude to my supervisor Professor Paul Elkington for all the support and encouragement he gave me. Without his guidance and constant feedback this PhD would not have been achievable and especially for his constant optimistic thinking that drove us further during challenging times. I particularly appreciate his capacity to keep in mind the big picture for tuberculosis control which gave a profound purpose to my work.

I gratefully acknowledge all the support and scientific direction from my supervisor Dr. Spiro Garbis who during our many inspirational discussions continuously encouraged me to embrace the beauty of the mass spectrometry and expanded the perspectives of this work.

I am very grateful to Dr. Christopher Woelk for his supervisory role during the first two years of my PhD. His support and guidance on the bioinformatics and data analysis strengthened my scientific education and definitively encouraged me to immerse myself in the bioinformatic analysis of the proteomic data which opened a whole new perspective of my work.

I thank Dr. Cory White for his insight on the statistical analysis, and for providing the R script used to normalise raw peptides intensities and calculate protein expressions. Dr. Ashley Heinson for the initial training provided in R environment. Dr. Antigoni Manousopoulou for providing initial training on the MudPIT platform, and Hannah Schiff for her experimental assistance during long hours of Luminex and ELISA optimisation and validation experiments. I would like to thank also Dr. Andres Vallejo for his input on the normalisation tools such as ComBat and co-expression based network analysis using Biolayout and WGCNA his contribution allowed me to unveil novel insights in the tuberculosis biology derived from the proteomic data generated in my work.

This work would not have been possible without the contribution of our collaborators who generously provided the clinical samples for the discovery phase and validation. I thank Dr. Naomi Walker who provided the samples from South Africa, Dr. Cesar Ugarte Gil who provided the samples from Peru and Dr. Marc Tebruegge who provided the samples from the MIMIC cohort.

I am deeply grateful for the funding that I received towards my PhD from COLCIENCIAS, the Government of Colombia, and the Faculty of Medicine of the University of Southampton. I specially thank Dr. David Wilson for all his support and mentorship throughout my PhD, his unconditional support encouraged me to move from Colombia and embark on this fulfilling journey. My thanks to all the members of the tuberculosis group and the proteomics group, who all helped me in numerous ways during various stages of my PhD.

I would also like to say a heartfelt thank you to my Dad, Mum, Cindy and Cristina for always believing in me and encouraging me to follow my dreams. And my friends Catalina and Yurany for helping in whatever way they could during this challenging period.

And finally to Andres, who has been by my side throughout this PhD, living every single minute of it, and without whom, I would not have had the courage to embark on this journey in the first place making it possible for me to complete what I started.

# CHAPTER 1

# Introduction

## Overview of PhD

Pulmonary tuberculosis is a highly infectious communicable disease that has accompanied humanity for centuries and still today claims over 1.7 million lives per year (1), more than any other infection. In spite of decades of efforts and scientific progresses in this field, around a quarter of world's population is infected and the strategies of control and elimination are insufficient (6-8). *Mycobacterium tuberculosis* (*Mtb*), an obligate human pathogen, is the etiological cause of tuberculosis. The host-pathogen interaction is highly complex and often results in a wide dynamic spectrum of heterogeneous clinical outcomes. However, from a public health and clinical perspective, tuberculosis patients are classified into one of two groups: latent disease that can persist asymptomatically for lifetime and active tuberculosis, which is the most infectious stage when pulmonary (9, 10). Subsequent to *M. tuberculosis* infection, most individuals will contain the infection as latent disease and only 5-10% will develop reactivation leading to the active tuberculosis (11). *M. tuberculosis* bacilli are transmitted directly between individuals through the airways when an individual inhales an infective aerosol generated by coughing of a patient with active pulmonary tuberculosis (12). Consequently, the *M. tuberculosis* transmission cycle relies completely on patients with active pulmonary tuberculosis and the rapid diagnosis of this population is a potentially effective constraining step to control the disease.

According to the latest World Health Organisation (WHO) report, in 2016, an estimated of 10.4 million people developed active tuberculosis, of whom over  died. The incidence of tuberculosis is heterogeneously distributed; the 30 highest burden countries including India, Indonesia, China, the Philippines and Pakistan accounted for 87% of all worldwide incidence cases (1). Figure 1A presents the global distribution of tuberculosis incidence presented as incidence per 100000 population per year. Estimates of incidence disaggregated by sex and age are shown in Figure 1B. Most cases occur during productive age, which exerts a considerable pressure on economies of high burden countries. The WHO estimated that the global economic burden of tuberculosis is approximately $12 billion yearly due to a 30% decline in average productivity among the population with the active disease (13).

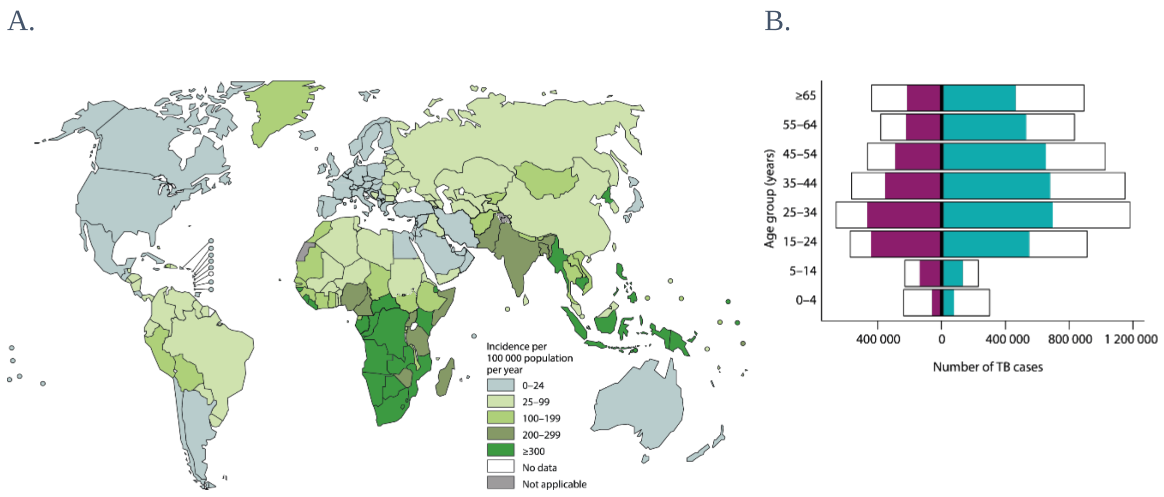


***Figure 1. Tuberculosis global incidence***

***A**. WHO estimated tuberculosis incidence cases of the active disease in 2016 relative to 100000 population per year (incidence rate) **B.** Tuberculosis incidence disaggregated by sex and age. Black line represents global estimate, purple incidence in women and turquoise in men. Adapted from Global Tuberculosis Report 2017, World Health Organisation, figures 3.4 and 3.30 (1)*

The host-pathogen interactions during *Mtb* infection is heterogeneous, not only in terms of the specific immune response, but also in terms of broad variables such as gender and ethnicity that significantly modulate some features of the immunopathogenesis (14, 15). Worldwide, reports across many countries indicate that tuberculosis notifications are approximately double in men compared to women (16). In 2016, 65% of global cases were males (Figure 1B) with a male to female ratio (M:F) of bacteriologically confirmed tuberculosis ranging from 1.3 in Ethiopia to 4.5 (in Viet Nam) (1). Despite sociocultural differences between countries, biological factors regulating the immune response to tuberculosis have been associated to gender. For instance, a genetic association has been proposed between tuberculosis and the locus in the X chromosome for the toll-like receptor 8 (TLR8, on Xp22) (17). CYBB and IKBKB, two of the nine genes known to correlate with mendelian susceptibility to *Mtb* infection are X-linked and therefore these are only observed in males (18). Furthermore, a possible relationship between steroid hormones and tuberculosis (18, 19) has been suggested considering that most of immune cells involved in tuberculosis control express specific receptors for sex hormones. While the mechanism behind this particular association is poorly understood, animal models and clinical observations indicate that the male bias mainly results from a biological effect rather than an epidemiological artifact (18).

Similar to age, ethnicity has been suggested as a host factor associated to differential pathogenesis in tuberculosis. For instance, a significantly higher prevalence of tuberculosis among black populations has been repeatedly reported in the literature for more than 30 years (20-23). The proposed causes of this disparity include differences in comorbidities, socioeconomic status and other environmental and host factors. Although social underpinnings of tuberculosis are known and social/health inequities associated to ethnic differences are well recognised in some social settings, epidemiological studies adjusting for socioeconomic factors suggest that other host factors contribute to the ethnic disparities in tuberculosis. For example, Nahid, P. *et al.*,  (2011) found that black participants were 2-fold more likely to self-report tuberculosis disease (95% confidence interval, 1.5–2.9) after adjusting for clinical and demographic indicators in a well characterised cohort of 5115 black and white participants in the United States (21). Additionally, inflammatory profiles of tuberculosis patients exhibit an ethnic heterogeneity associated to the host variability rather than the *Mtb* genotype (14, 15).

Most deaths caused by tuberculosis could be prevented with early diagnosis and appropriate treatment. Millions of deaths (53 million in total 2000–2016) are adverted every year due to successful detection and treatment of the infection, however there are still considerable gaps in diagnosis and treatment (1). The current tuberculosis diagnosis pipeline is focused on passive-case finding of active pulmonary tuberculosis and depends upon old and inadequate technologies. Therefore, it is widely accepted amongst key stakeholders that early and fast tuberculosis diagnosis remains as an urgent challenge to address in order to control this disease (4, 7, 24-26).

The tuberculosis diagnostic techniques recommended by the WHO have important limitations that prevent their application for population screening. For instance, techniques such as microscopy examination, culture and Xpert MTB/RF require facilities available in intermediate to reference level laboratories (3). A recent survey of several stakeholders identified as a top priority for tuberculosis control the development of a non-sputum-based test suitable for point-of-care (POC) based on biomarkers or biosignatures (27). Mathematical modelling suggests that development of new strategies for vaccination and treatment and the introduction of a POC test are essential to increase the current decline of the incidence global rate from 1.5% to 17%/year for 2035 (28). This reduction in the incidence rate is required to meet the ambitious targets proposed by 2035 in the WHO's End TB Strategy (29)  and the development of novel POCs will require the discovery and validation of new biosignatures.

Biomarkers are measurable characteristics that allow identification of a particular physiological state or process and can be investigated at virtually any level of a biological system (30). Particularly, proteins as ultimate biological effectors are reliable markers for disease states especially relevant when post-translational processes are driven by pathogenic agents. Non-targeted mass spectrometry proteomics is the comprehensive study on a large scale of protein profiles and usually implicates identification and quantification of all proteins constituting diverse biomedical specimens (31). Therefore, proteomics is an ideal approach for biomarker discovery, frequently on basis of relative quantification over a control or healthy state, and the candidates are accordingly defined using fold changes in protein expression (32).  In order to quantify a proteome, diverse label-free (spectral counting), chemical isobaric-stable isotope tagging strategies have been developed, such as iTRAQ (Isobaric tags for relative and absolute quantitation), TMT (Tandem Mass Tag) or metabolic SILAC (Stable isotope labelling by amino acids in cell culture) with subsequent liquid chromatography – mass spectrometry (LC-MS) based analysis (31). Combined quantitative techniques and shotgun proteomics (non-targeted approaches) allow the unbiased and systems-based interrogation of protein profiles resulting from a pathological state in complex biological samples.

Biological significance and quality of the samples are crucial for biomarker discovery using proteomics (5). Although active tuberculosis infection is confined mainly to the lungs, pathological progression may be tracked in the peripheral blood. Therefore, analytical matrixes such as plasma are ideal for biomarker discovery through proteomics. The ease of collection and biological relevance allows the use of plasma as an indicator of the overall physiological state (33). Nevertheless, plasma proteomics involves analysis of a highly complex matrix, since plasma exhibits a significant variability between individuals and also presents a wide dynamic concentration range of proteins spanning twelve orders of magnitude (34). This challenge has been usually addressed using depletion methods; however, depletion can lead to unspecific co-removal of less abundant proteins. Accordingly, new analytical methods that avoid depletion will allow larger coverage of the plasma proteome and therefore increase the opportunities for discovering novel biomarkers, and such approaches include Multidimensional Protein Identification Technique or MudPIT (35).

Although new and more effective diagnostic tools for tuberculosis will require discovering novel biomarkers of the disease, validation of those candidates is crucial for clinical translation. Some studies have interrogated the plasma proteome in active tuberculosis using various methods, and have suggested that proteins such as, APOCII, CD5L, HABP2, RBP4 S100A9, SOD3, and MMP9 are biomarkers (36-38). However, these candidates were only verified or validated in limited cohorts, preventing their possible clinical translation (38). Typically biomarker validation is a challenging process that involves the evaluation of: sensitivity, specificity, variability, precision, reproducibility, accuracy, range of use, limit of detection, and probability of false negatives (39). On the other hand, biomarker qualification is aimed to determine the clinical validity of the candidate and implies diagnostic accuracy studies (40). Combined efforts in both analytical studies and clinical translation will be necessary to deliver a suitable test for tuberculosis population screening critical for transmission control of this disease.

### 1.1 Tuberculosis

Tuberculosis (TB) is one of the oldest recorded diseases but still remains as a leading cause of mortality in the world (41, 42). 9000-years-old archaeological evidence have demonstrated TB lesions in human bones from settlements of the Neolithic period (43). Despite more than a century of research, the growing development of a range of antituberculous drugs, the availability of a vaccine (BCG) and global strategies implemented to fight this health threat (44, 45), the latest report in 2017 estimated that 10.4 million people have active TB and  died as a consequence of this disease (1).   Therefore, tuberculosis control is lagging far behind other major diseases such as HIV and malaria.

The obligate intracellular pathogen *Mycobacterium tuberculosis* (*Mtb*) is the causative agent of tuberculosis and was identified for first time by Robert Koch in 1882. Due to the aerobic nature of this acid-fast bacillus, it grows most successfully in highly oxygenated tissues, such as the lungs (46). Although tuberculosis is mainly a pulmonary disease, the infection may develop as extra-pulmonary disease affecting other organs, and it can even involve multiple organs, especially in the context of human immunodeficiency virus (*HIV*) co-infection (47).

The life cycle of *Mtb* is initiated when an individual inhales aerosolised bacilli from a patient with active pulmonary disease (46, 48). Figure 2 depicts the tuberculosis pathogenesis cycle. After the infectious droplets are inhaled and deposited in the alveoli at the well ventilated base of the lungs, innate responses are triggered mainly through alveolar macrophages and dendritic cells. *Mtb's* ability to infect macrophages seems to promote bacterial dissemination (49). Induction of the adaptive response is triggered later when dissemination of the mycobacteria to draining lymph nodes occurs, and antigen presentation by dendritic cells lead to priming and expansion of effector T cells. Granuloma formation is promoted by the migration of these effector cells to the lungs in combination with other leukocytes (50).  The granuloma is the hallmark structure of tuberculosis; at its most basic is an organised and compact aggregate of epithelioid cells, mainly macrophages that have undergone transformation to develop tightly interdigitated cell membranes that link adjacent cells. In immunocompetent hosts, the immune response elicits the formation of granulomas, where *Mtb* is successfully contained but not eliminated (48). The initial granulomas may heal, resulting in a calcified Ghon focus in the lower portion of the lungs indicating the sites of the primary infection (Figure 2).

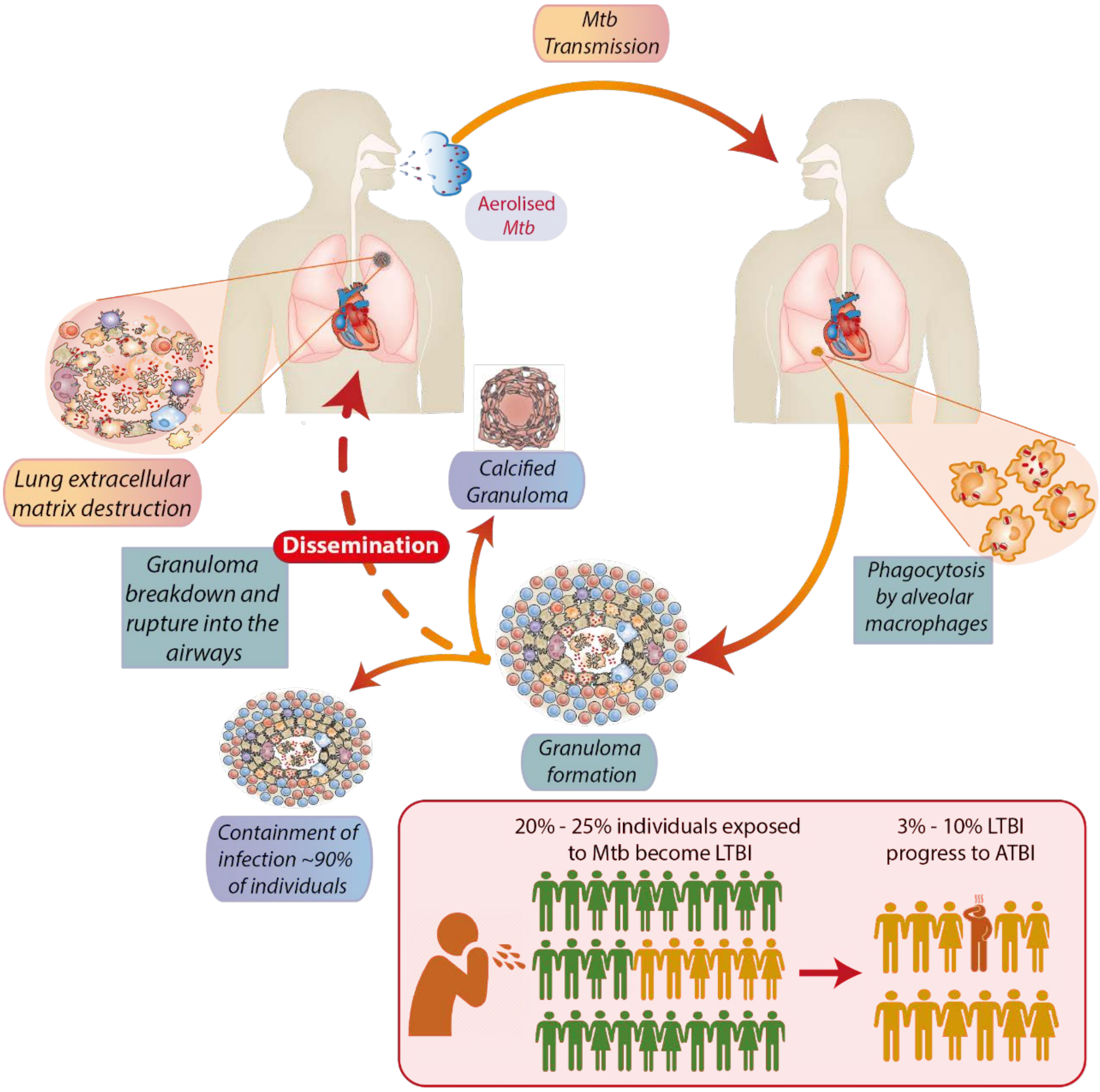

Mtb Transmission
Aerolised Mtb
Lung extracellular matrix destruction
Calcified Granuloma
Dissemination
Granuloma breakdown and rupture into the airways
Phagocytosis by alveolar macrophages
Granuloma formation
Containment of infection ~90% of individuals
20% - 25% individuals exposed to Mtb become LTBI
3% - 10% LTBI progress to ATBI

*Figure 2. Tuberculosis pathogenesis*

*Inhalation of aerosol droplets containing the bacteria Mycobacterium tuberculosis initiates the infection. Macrophages take up the bacteria and transport across the alveolar epithelium to the lungs. Subsequent to dissemination of bacilli to the draining lymph node, dendritic cell presentation of antigens drives T cell priming and expansion of antigen-specific T cells, which migrates to the lung. Recruitment of activated macrophages, T cells, B cells and other leukocytes leads to granuloma establishment. Many different cell populations are part of the granulomas, such as dendritic cells, neutrophils, natural killer (NK) cells, B and T cells, fibroblasts and cells that secrete extracellular matrix components (51). Infection is mainly contained latently (LTBI) but approximately 10% of patients progress to the active state (ATBI) when the bacilli can be coughed up and spread (50, 52, 53).*

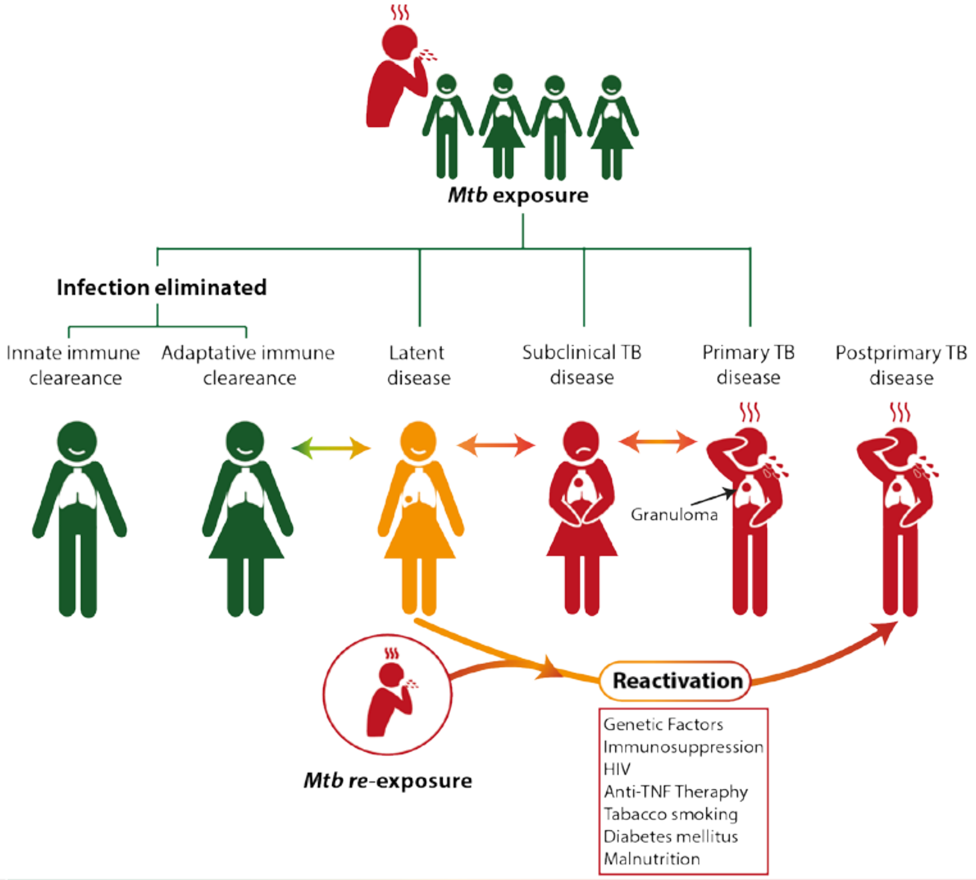
Classically, diverse clinical phenotypes has been recognised following *Mtb* infection: *primary active* disease, which is referred as symptomatic primary infection occurring soon after infection; *latent* disease (LTBI), in most cases, around 90%, the infection is contained asymptomatically for life and this population constitutes an enormous reservoir of potential transmission. Around 5% to 10% of these *latent* patients will progress to *secondary active* disease (ATBI) at any time of their lifetime typically as apical pulmonary tuberculosis (48). More recently, tuberculosis has become recognised as a highly heterogeneous disease encompassing a variety of immune responses resulting in a wide range of clinical manifestations. These may include: infection with clearance without detectable adaptive response, localised immune response not detectable systematically, bacterial persistence with active immune control and subclinical active disease (9).

Considering the range of clinical outcomes resulting from the heterogeneity of the immune responses triggered by *Mtb* infection, diagnosis is highly challenging. Importantly, solely individuals with the active disease can transmit the disease since the immune control is disrupted and the bacilli can disseminate through the airway.  Symptoms of active disease can range from systemic responses such as weight loss, night sweats and fever to cough and haemoptysis in pulmonary disease. Radiological examination can show pulmonary abnormalities such as consolidation, cavities and thoracic lymphadenopathy. In spite of these overt clinical manifestations, confirming the active state is challenging but crucial to break the cycle of transmission. In addition, the clinical presentation of active tuberculosis overlaps with diseases such as pneumonia, bronchitis and lung cancer (9).

### 1.1.1 Current Diagnostic Pipeline

The complex biology of tuberculosis has hampered the development of accurate and rapid point-of-care diagnostic tests, which remains as one of the major hurdles to global control of this disease as presented in Figure 3 (54). Currently, the gold standard tests for tuberculosis are laboratory based, and the diagnosis can take weeks or even months (55). In latent disease, infection or exposure to *Mtb* can be only demonstrated by the reactivity of the host to *Mtb* antigens, and until the beginning of this century the tuberculin skin test (TST) was the only diagnostic tool available for the LTBI diagnosis. This test measures the induration formed after intradermal inoculation of a culture filtrate of *Mtb* known as PPD (Purified Protein Derivate) which contains around 200 antigens of *Mtb* into the volar forearm (9, 11). A delayed-type hypersensitivity reaction is promoted in patients previously exposed to *Mtb* and the size of this reaction is measured 48 to 72 hours after the initial inoculation  (56).

Although TST is widely used as the clinical applications are well established, it has important limitations. TST testing requires two visits to the healthcare centre which results in significant loss of readings, it has a limited sensitivity in immunocompromised patients and exhibits cross-reactivity with Bacillus Calmette–Guérin (BCG) vaccination and non-tuberculous bacteria (11). Although some of these drawbacks have been addressed by antigen-specific interferon-γ (IFN-γ) release assays (IGRA) which are performed using peripheral blood after stimulation with culture filtrate protein-10 (CFP-10) and early secretory antigenic target 6 (ESAT-6), neither TST or IGRA distinguish between active and latent stages or estimate the risk of tuberculosis progression to the active disease (11, 55). Additionally, IGRA tests require extensive laboratory infrastructure and training.



| | | | | | | |
|---|---|---|---|---|---|---|
| TST | Negative | Positive | Positive | Usually Positive | Usually Positive | Usually Positive |
| IGRA | Negative | Positive | Positive | Usually Positive | Usually Positive | Usually Positive |
| Culture | Negative | Negative | Negative | Usually Positive | Usually Positive | Usually Positive |
| Sputum Smear | Negative | Negative | Negative | Usually Positive | Usually Positive | Usually Positive |
| Infectious | No | No | No | Sporadically | Yes | Yes |
| Symptoms | None | None | None | Mild or none | Mild or severe | Mild or severe |

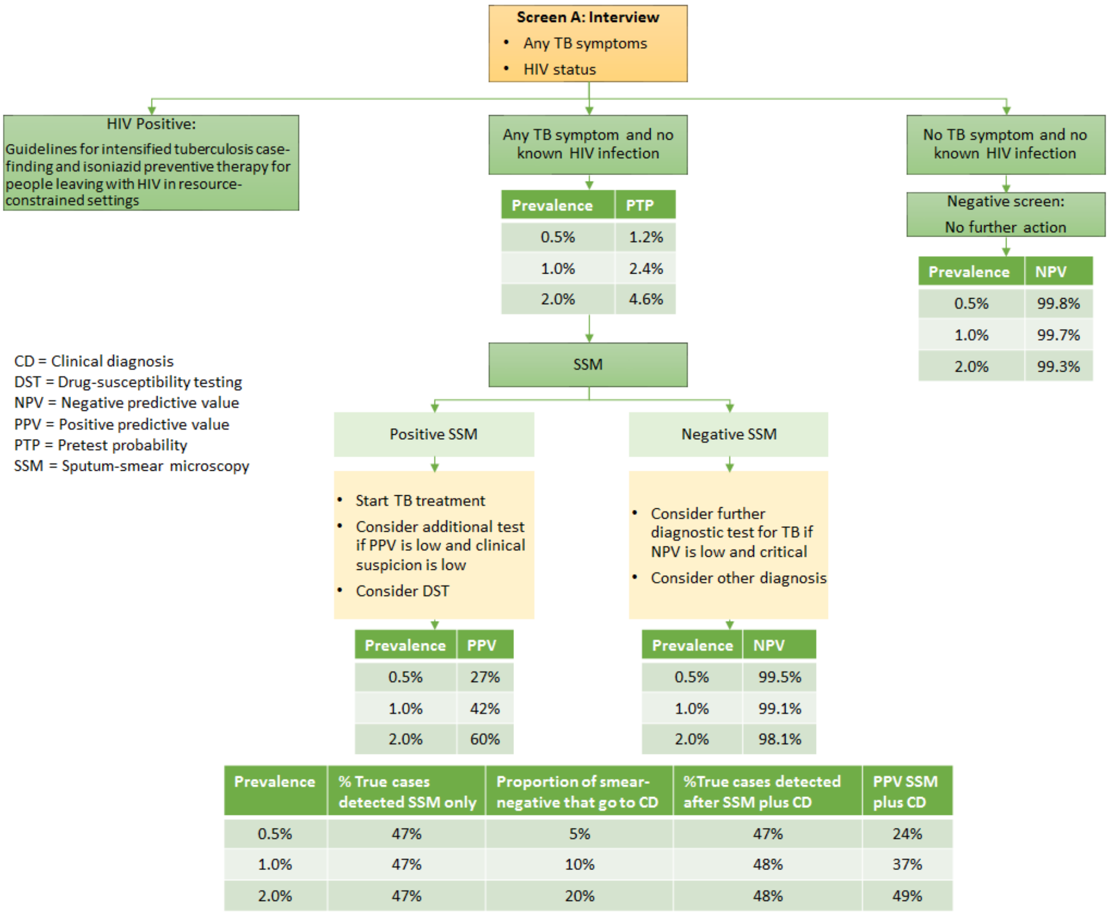
***Figure 3. Tuberculosis natural history and diagnostic tools***

*Following Mtb-exposure a heterogeneous range of clinical outcomes can take place, ranging from clearance of the infection through innate responses leaving no trace of Mtb exposure to symptomatic and infectious active disease. Mtb exposure or latent infection is inferred by detecting host's reactivity to microbial antigens using either the tuberculin test (TST) or the IFN-$\gamma$ release assay (IGRA). However, a positive response will be shared with individuals that have cleared the infection through adaptative response reactions. Additionally, patients with subclinical disease and active infection will have a positive response to these tests. Conversely, patients with disease-induced immunosuppresssion may have negative test results. Active tuberculosis is diagnosed by detecting Mtb through sputum smear and culture. Positive results are highly dependent on bacillary burden and therefore sensitivity is highly variable. The pathophysiology complexity of tuberculosis limits the performance of the current available diagnostic tools and significantly contributes to underdiagnosis and transmission (9, 57).*

The diagnosis of the active disease relies on the detection of the *Mtb* bacilli or direct products of the pathogen. Sputum smear microscopy remains the most common diagnostic test for the active disease in low- and middle- income countries, which represent over 90% of the worldwide TB burden (46, 58). Smear microscopy has particularly variable sensitivity between 32% and 97% and it is unable to distinguish drug-resistant strains (59). Confirmation of TB diagnosis requires culture of bacilli, which can take over six weeks. The World Health Organization recommended in 2010 the implementation of the *Xpert MTB/RIF* test for real-time PCR identification of *Mtb* and rifampicin resistance (60). This method is robust, simple and fast, although it requires electrical supply and a high initial investment in machines, consumables and infrastructure, which often is not available in high TB-burden countries (9, 61, 62). However, the sensitivity of the Xpert assay displays variable performance according to the clinical settings (63). Delayed or missed diagnosis and deficient access to high quality healthcare lead to suffering, sequelae, catastrophic financial consequences, higher risk of death and critically sustained transmission of infection.

Various screening algorithms have been developed for children and adults (www.who.int/tb/tbscreening). Initial screening for pulmonary tuberculosis includes screening for symptoms or screening with chest radiography. These algorithms have different performance and depend upon the disease prevalence in the screened population. The screening algorithm recommended by the WHO in cases when chest radiography or Xpert MTB/RIF are not available is completely based on symptomatology assessment and sputum smear positivity (Figure 2) (64). The risk of false-positive diagnosis increases as the prevalence declines, thus accuracy is a crucial when the prevalence of tuberculosis is less than 1% in the target population.

The screening algorithm presented in Figure 4 relies on passive case-finding which is the most common strategy in low- and middle-income countries, prioritising treatment success among detected cases. However, patients typically only seek treatment when the symptoms have worsened and during the time they were unwell prior to diagnosis they are infectious and extensive transmission has occurred. Targeted active case-finding and early initiation of treatment are essential for epidemic control of tuberculosis, strategies such as FAST (“Finding TB cases Actively, Separating safely, and Treating effectively”) has facilitated health-care facilities to implement procedures for reducing the duration and risk of exposure to tuberculosis for both health-care workers and patients (65).


Screen A: Interview
Any TB symptoms
HIV status
HIV Positive:
Guidelines for intensified tuberculosis case-finding and isoniazid preventive therapy for people leaving with HIV in resource-constrained settings
Any TB symptom and no known HIV infection
No TB symptom and no known HIV infection
Negative screen:
No further action
SSM
Positive SSM
Negative SSM
Start TB treatment
Consider additional test if PPV is low and clinical suspicion is low
Consider DST
Consider further diagnostic test for TB if NPV is low and critical
Consider other diagnosis
CD = Clinical diagnosis
DST = Drug-susceptibility testing
NPV = Negative predictive value
PPV = Positive predictive value
PTP = Pretest probability
SSM = Sputum-smear microscopy

| Prevalence | PTP |
|---|---|
| 0.5% | 1.2% |
| 1.0% | 2.4% |
| 2.0% | 4.6% |

| Prevalence | NPV |
|---|---|
| 0.5% | 99.8% |
| 1.0% | 99.7% |
| 2.0% | 99.3% |

| Prevalence | PPV |
|---|---|
| 0.5% | 27% |
| 1.0% | 42% |
| 2.0% | 60% |

| Prevalence | NPV |
|---|---|
| 0.5% | 99.5% |
| 1.0% | 99.1% |
| 2.0% | 98.1% |

| Prevalence | % True cases detected SSM only | Proportion of smear-negative that go to CD | %True cases detected after SSM plus CD | PPV SSM plus CD |
|---|---|---|---|---|
| 0.5% | 47% | 5% | 47% | 24% |
| 1.0% | 47% | 10% | 48% | 37% |
| 2.0% | 47% | 20% | 48% | 49% |

*Figure 4. Screening algorithm in cases where chest radiography and Xpert MTB/RIF are not available*

*The sensitivity of this particular algorithm is limited by the use of sputum-smear microscopy as the principal diagnostic tool. The specificity of sputum-smear microscopy varies depending on the prevalence of non-tubercular mycobacteria, gold standard used for assessment, case definition, and the quality of slide preparation and reading. Adapted from (64)*

Rapid biomarker-based tests that do not depend on detection of the bacilli in sputum could increase the accuracy and speed of diagnosis. Therefore, there is an urgent need for developing new diagnostics for tuberculosis as it has been estimated that fast and widely available tests, highly sensitive (≥85%) and specific (97%), could save around 400,000 lives per annum (4). The ideal diagnostic test should be accessible in the point-of-care, giving fast results, working without requiring electricity, refrigeration or clean water, and should be easy to operate with minimal training (4, 55, 66). Carefully searching, qualifying and validating biomarkers and new signatures is a pivotal task required to meet the needs in tuberculosis diagnosis. However, differentially expressed host molecules in different clinical phenotypes are not necessarily qualified biomarkers. A high number of immunological markers have been described as differentiating markers on the basis of general exploratory data, but have not been qualified properly (10).

### 1.1.2 Urgency for Novel Diagnostic Tools

In 2016 the global gap between the new and relapse cases notified and the estimated incidence of tuberculosis was 39%, meaning about 4.1 million people were undiagnosed or managed in informal/private sectors (1). This alarming figure has a critical impact on the transmission cycle of tuberculosis since most of these individuals remain driving ongoing transmission in their communities. The WHO has identified three main factors that determine this gap (1):

- Under-reporting of detected cases: especially relevant in countries lacking of mandatory policies for notification of cases in both public and private sectors.
- Uncertainty about the levels of tuberculosis incidence: estimation of incidence for 54 countries presented in the latest WHO report was based on expert opinions rather than direct data from surveillance or surveys.
- Under-diagnosis of tuberculosis cases: poor access to healthcare; absent or mild symptomatology that delay individuals to seek healthcare; failure to test for tuberculosis when people contact health system; and limited performance of current diagnosis tools.

Noticeably, the contribution of each one of these factors to the gap of missing cases must considered as context-dependent. For instance, in countries with state-of-the-art national surveillance systems gaps between the number of notified cases and the tuberculosis incidence could be attributed to failure to detect the infection suggesting under-diagnosis. Conversely, in countries where major financial or geographical barriers limit the access to healthcare, discrimination of the particular contribution of underreporting and under-diagnosis to the gap of missing cases is more challenging to define. This epidemiological complexity implies that the effective impact of point-of-care diagnostic tools reducing the gap of missing cases might be maximised in contexts where under-diagnosis is the major responsible for missing cases.

Although significant developments have occurred in diagnostic technologies, many high burden countries depend upon antiquated sputum smear microscopy. Worldwide, only 57% of tuberculosis cases are confirmed with this tool, which requires a high bacilli burden. Unfortunately, tuberculosis is strongly correlated with low-income settings and only low-cost tools available at the decentralised level for the communities will efficaciously address the gap of undiagnosed tuberculosis (67, 68).

Diverse strategies have been explored in order to develop an ideal diagnostic test, which might compromise factors such as: immediate results, suitable for point-of-care and widely available for any level of the health-system care; ranging from hospital wards, peripheral health posts to outreach teams visiting remote locations and within patient’s home. Although standalone diagnostic tools are already available for infection diseases such as HIV, malaria, and Chagas’ disease, tuberculosis imposes the greatest challenges. The wide range of clinical manifestations of the infection that can be produced by even low bacilli loads and occurs in potentially any anatomic site hampers the development of this necessary test (69).

In addition to Xpert MTB/RIF, there are various tests under development and a limited number of them have some potential for use in point-of-care in restricted conditions. Loop-mediated isothermal amplification (LAMP), a simplified manual molecular assay for laboratory-based with visual colorimetric readout is aimed for resource-limited settings since it only requires a water bath for amplification. However specificity was insufficient to be recommended as replacement of microscopy (70).  Detection of *Mtb* antigens are considered a promising diagnostic tool since it may reflect mycobacterial burden and it does not relies on the particular immune host's response.

Mycobacterial antigens can be detected in urine and the most promising of these is the cell wall lipopolysaccharide lipoarabinomannan (LAM). There are already two assays commercially available for this antigen: an ELISA based test and a lateral flow test strip for urine. Multiple studies have shown that the target population for this assay is HIV-associated TB in patients with advanced immunodeficiency, since the underlying mechanism of LAM presence in urine is renal involvement with TB following haematogenous seeding (69). Recently, hydrogel technology was used to capture LAM in urine of HIV-negative tuberculosis patients allowing for the detection of this antigen at very-low abundance levels (71). On the other hand, the prospect for a rapid, simple, low-cost and non-instrumental assay suitable for all levels of the health system and community has made the serologic test very attractive. However, in 2011 WHO issued a negative recommendation against current available test as they exhibited a limited accuracy and therefore were of no clinical value (69, 72). Nevertheless, serological testing cannot be discouraged and active research is taking place. A better understanding of the humoral response to the tuberculosis infection, larger prospective studies and better methods for discovery and validation of new target candidates are required for developing new tests.

## 1.2 Biomarkers and Tuberculosis Diagnosis

Implementation of biomarkers on translational medicine has the potential of radically improving diagnosis, prognosis, treatment and follow-up of disease, which would reduce mortality and morbidity burdens and healthcare costs (73, 74).  Additionally, biomarkers can be used as a powerful and dynamic approach in randomized clinical trials, analytic and observational epidemiology and screening (75).

### 1.2.1 Biomarker Definitions and Classification

It is relevant to distinguish among biomarkers, clinical endpoints and surrogate end points, considering their different utility and application in the clinic. A biological marker or *biomarker* was defined by the Biomarkers Definition Working Group (2001) as *"a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention"*(76). From a more practical perspective this definition is augmented to emphasise its clinical utility. A marker is any physical sign or laboratory measurement which substitutes a clinically relevant end point and is expected to predict the effect of therapy (77). Biomarkers, then, include a wide selection of molecules that ranges from gene expression products, gene variants, and single nucleotide polymorphisms to proteins, polysaccharides, hormones and metabolites.

On the other hand, a *clinical endpoint* is defined as a variable or characteristic that reflects how a patient functions or feels, or how long is the survival expectancy. A *surrogate endpoint* is a biomarker that intends to substitute for a clinical endpoint. In is expected that if pathophysiological, therapeutic or epidemiological or other evidence is used to select a clinical endpoint, this may predict clinical benefits harm, or lack of effect (76, 78).

Classically biomarkers were classified into three groups for the purpose of distinguishing them from clinical endpoints, thus enabling debate on application and validation of surrogate end points (78). These three groups are:

- *Type 0 biomarkers*: Indicate the particular stages of the natural history of a disease and correlate longitudinally with established clinical signs, such as symptoms.
- *Type I biomarkers*: Enable monitoring and prediction of the effects of a therapeutic intervention depending upon the pharmacologic mechanisms and properties of the drug.
- *Type II biomarkers*: Considered surrogate endpoints because changes in the marker state allows to predict clinical benefits (75-78).

Despite the efforts made to standardise the terminology describing characteristics, classification and application of biomarkers (76, 78), some confusion remains in the scientific literature particularly driven by diverse concepts of translational and personalised medicine. However, six categories has been proposed from a biomarker-driven decision-making process during disease management (74). Figure 5 describes these six groups: stratification, efficacy, differentiation, toxicity, screening and prognostic. Suitable diagnostic tools are critical to deliver the potential of clinical biomarkers in the clinical practice.

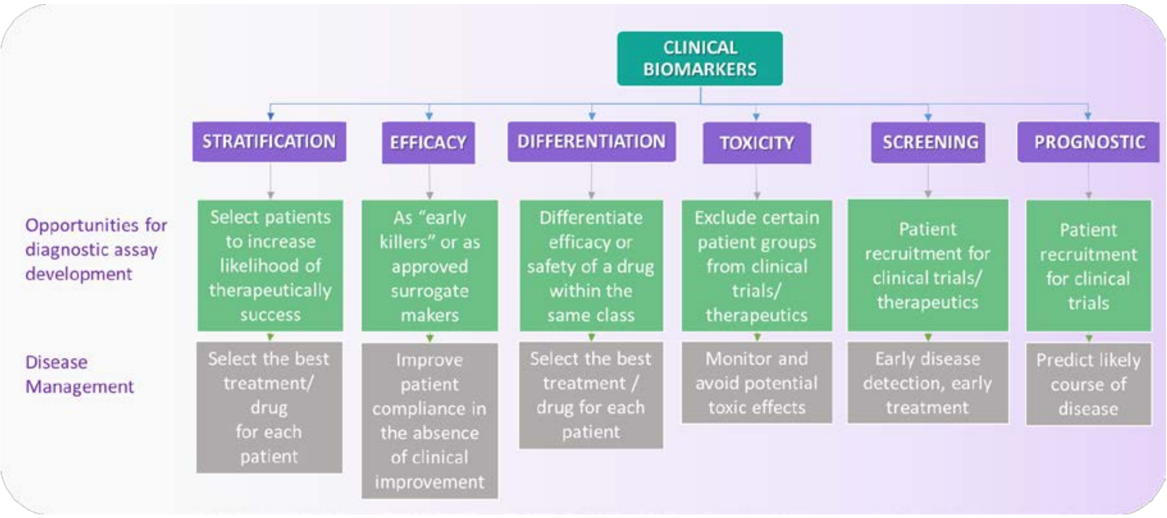

CLINICAL BIOMARKERS
STRATIFICATION
EFFICACY
DIFFERENTIATION
TOXICITY
SCREENING
PROGNOSTIC
Opportunities for diagnostic assay development
Disease Management
Select patients to increase likelihood of therapeutically success
As "early killers" or as approved surrogate makers
Differentiate efficacy or safety of a drug within the same class
Exclude certain patient groups from clinical trials/ therapeutics
Patient recruitment for clinical trials/ therapeutics
Patient recruitment for clinical trials
Select the best treatment/ drug for each patient
Improve patient compliance in the absence of clinical improvement
Select the best treatment / drug for each patient
Monitor and avoid potential toxic effects
Early disease detection, early treatment
Predict likely course of disease

*Figure 5. Overview of biomarker categorisation*

*Diverse biomarker classification groups have been proposed according to their characteristics or applicability. However, six categories can be defined considering clinical biomarkers utility in a biomarker-driven decision-making process during disease management. Diagnostic tools encompassing biomarker functions are required to deliver their potential in the clinics. Adapted from (74)*

### 1.2.2 Current Screening Biomarker Pipeline

Although most of the current pipeline of screening biomarker development has been driven by the needs and advances in cancer biomarker research, the same rationale can be fully incorporated to any other translational need. Simplistically, this pipeline can be described in terms of five phases, which intrinsically involve pre-analytical, analytical and post-analytical challenges (79, 80). Accordingly, each of these five phases might encompass different validation and verification procedures in order to ensure that the required efforts and funding for biomarker development are invested in the most promising candidates. Figure 6 outlines the biomarker pipeline structure necessary to develop biomarkers from bench-based research to routinely tests at the bedside (79-82). Additionally, it presents the main objectives of each phase and some requirements in terms of validation and sampling.

Taking into consideration the main aims of this research, which involve discovery and validation of screening biomarkers for tuberculosis, it is necessary to discuss in further detail the phases of biomarker development.  Special emphasis will be placed on *Phase I*, which directly implicates the discovery strategies.

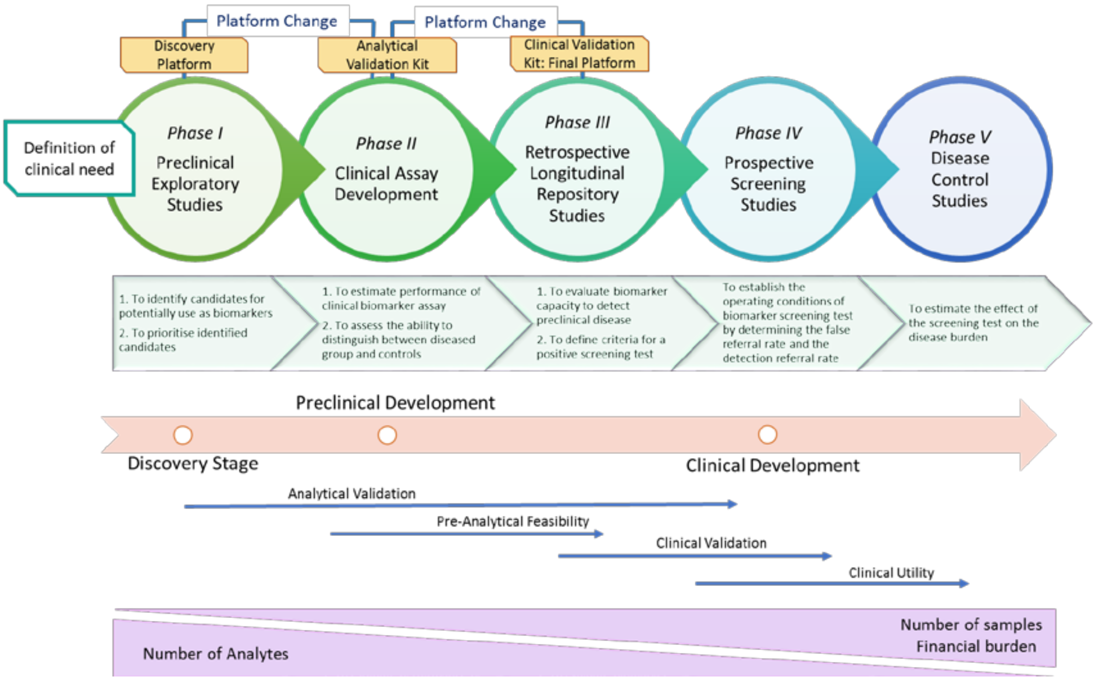

Platform Change
Platform Change
Discovery Platform
Analytical Validation Kit
Clinical Validation Kit: Final Platform
Definition of clinical need
Phase I
Preclinical Exploratory Studies
Phase II
Clinical Assay Development
Phase III
Retrospective Longitudinal Repository Studies
Phase IV
Prospective Screening Studies
Phase V
Disease Control Studies
1. To identify candidates for potentially use as biomarkers
2. To prioritise identified candidates
1. To estimate performance of clinical biomarker assay
2. To assess the ability to distinguish between diseased group and controls
1. To evaluate biomarker capacity to detect preclinical disease
2. To define criteria for a positive screening test
To establish the operating conditions of biomarker screening test by determining the false referral rate and the detection referral rate
To estimate the effect of the screening test on the disease burden
Preclinical Development
Discovery Stage
Clinical Development
Analytical Validation
Pre-Analytical Feasibility
Clinical Validation
Clinical Utility
Number of samples
Financial burden
Number of Analytes

***Figure 6. Biomarker development pipeline***

*This figure summarises the phases of biomarker development and the objectives of each one. Additionally, it shows a potential regulatory scenario for a single test accounting for an FDA Co-Developed Products document. Adapted from (79, 80, 83)*

- *Phase I – Preclinical Exploratory Studies*: Once a clinical need has been identified, sufficiently documented and evaluated, the screening biomarker pipeline is initiated with preclinical exploratory studies. *Phase I* integrates the discovery stage and mainly aims to identify new candidates and prioritise them according to its ability to distinguish the diseased condition from healthy controls. There are two complementary biomarker discovery approaches considering the underlying research paradigm: (1) 'Knowledge-based' (deductive method) and (2) the 'unbiased' (inductive strategy) (84). However the term *'Discovery'* is frequently used in the biomarker literature to refer an unbiased and semiquantitative process by which the differential expression of specific analytes (gene products, gene variants, proteins, metabolites, etc.) between different states is first established (81). High-throughput technologies, when adequately performed, enable the simultaneous unbiased assessment of thousands of analytes. Typically the discovery phase is executed on a limited number of samples, therefore strictly-defined Standard Operating Procedures (SOPs) for sample selection, collection, storage, handling, analysis and data mining (73, 74) are critical to define and prioritise the most promising candidates for further validation and qualification.

*Phase II – Clinical Assay Development for Disease Testing:* A clinical assay based on a specimen, typically the same type used in the discovery phase, is developed at this stage (79). Additionally, an analytical validation must be implemented in order to determine the test performance distinguishing the diseased condition from healthy controls and its operating conditions. On the other hand, the platform used for validation should be reasonably simple ensuring reproducible results within and between laboratories.

The participating cohorts must be carefully stratified and representative of the screening target population. The sample size is determined by the requirements of validation and may be large enough to allow testing of the null hypothesis $H_0$: that operating characteristics are below target values. A biomarker with high level of discrimination, despite random variation, might have a high probability of rejecting the $H_0$ (79). The power calculation will depend on the objective of the test. For instance, if the objective of the diagnostic test is to determine whether (or not) a specific biomarker can be used as a screening tool; then the validation has to ensure that it has a sufficiently-high degree of sensitivity, but a lower degree of specificity can be tolerated. On the other hand, if the test is developed as a specific tool to be used as a diagnostic tool, then the validation will usually have to target for a high degree of both sensitivity and specificity (85).

- *Phase III – Retrospective Longitudinal Repository Studies:* During this stage the biomarker's ability for detecting preclinical disease is evaluated as a function of time. Algorithms for screen positivity based on multiple markers can be developed. Effects of cofounding variables including demographics and other relevant clinical information on biomarker discriminatory properties are described in this stage as well (79). Consequently, the final analytical validation platform for screening purposes is designed and its clinical feasibility evaluated.

- *Phase IV – Prospective Screening Studies*: Importantly this phase involves screening the relevant population leading to diagnosis and subsequent treatment, therefore important ethical considerations are an integral part of this stage. The operating characteristics of the screening biomarker test are evaluated by determining the false referral rate and the detection rate (79).
- *Phase V – Disease Control Studies:* This phase aims to establish the net benefit of a screening biomarker test on the population. There are various reasons why a good biomarker test might not represent an overall benefit for the tested population. Some of them may include political or economic reasons such as *(a)* limited compliance with screening programs or difficulties for implementing it in terms of practicality, *(b)* prohibitive economic costs of the screening itself and the diagnostic workup of individuals falsely screened positive (79).

### 1.2.3 Biomarker Validation and Qualification

Emergent high-throughput technologies such as DNA microarrays, RNA-seq, and proteomics continuously deliver a myriad of new biomarker candidates annually, notwithstanding only a considerably reduced number are routinely utilised in the clinical practice. In 2011, it was estimated that from 150000 papers documenting thousands of new biomarkers just fewer than 100, around 0.07%, were properly validated for its use in the clinics (73). This dramatic figure illustrates the highly challenging process of biomarker validation, which should involve strict standardised selection, handling, and analysis of specimens as well as large scale studies for validation.

The biomarker path to routine clinical application is a stepwise process. It involves two general phases: *Validation* or assessment of biomarker performance characteristics and operational conditions under which reproducible and accurate data will be obtained. In addition, *qualification* establishes the clinical utility of the biomarkers through an evidentiary process that links biomarkers with biological processes and clinical endpoints (39, 40).

The validation approaches are defined by the type of biomarker that is assessed. *Type 0* can be validated using longitudinal studies using a well-defined patient population against a gold standard validator. *Type I* must be assessed in parallel with the drug candidate and *Type II* should link both pathophysiology of the disease and the mechanisms of the drug (77). This section is focused on diagnosis biomarkers for screening (*Type 0*) which are relevant for this work.

Many analytical parameters may be rigorously controlled in the biomarker discovery phase. The number of biomarkers that make it through the biomarker pipeline from the bench to the bedside is extremely limited. Several factors associated to limitations and pitfalls in biomarker discovery have been reported (39, 40, 74, 80, 84, 86-90) and some of them have been summarised in the Figure 7.

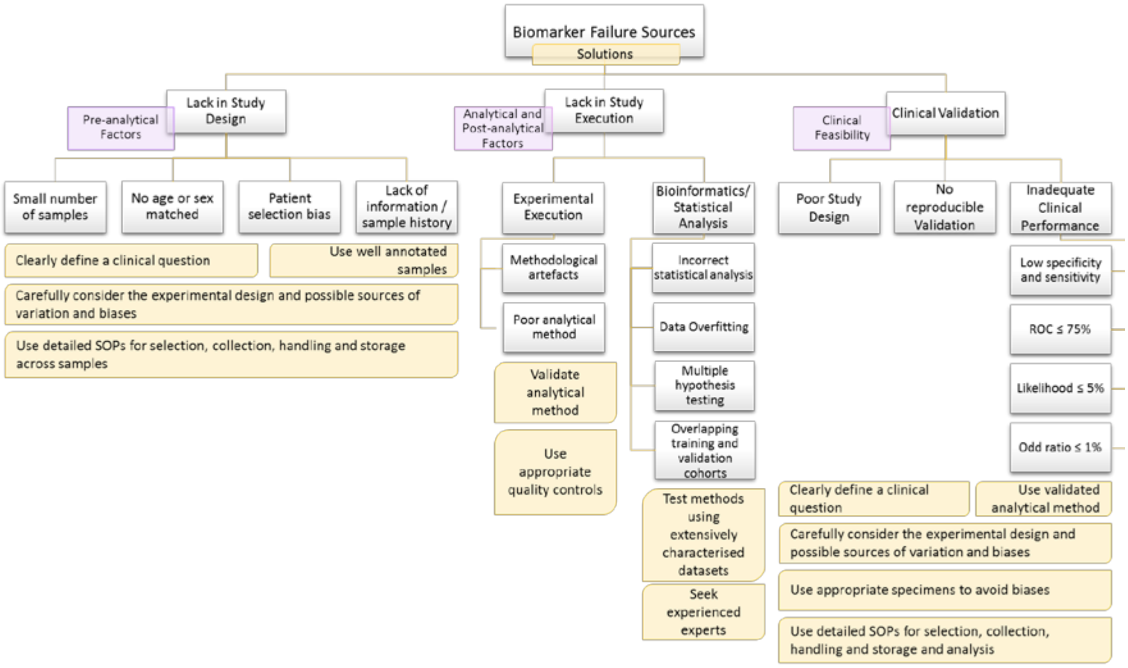

Biomarker Failure Sources
Solutions
Pre-analytical Factors
Lack in Study Design
Analytical and Post-analytical Factors
Lack in Study Execution
Clinical Feasibility
Clinical Validation
Small number of samples
No age or sex matched
Patient selection bias
Lack of information / sample history
Clearly define a clinical question
Use well annotated samples
Carefully consider the experimental design and possible sources of variation and biases
Use detailed SOPs for selection, collection, handling and storage across samples
Experimental Execution
Methodological artefacts
Poor analytical method
Validate analytical method
Use appropriate quality controls
Bioinformatics/ Statistical Analysis
Incorrect statistical analysis
Data Overfitting
Multiple hypothesis testing
Overlapping training and validation cohorts
Test methods using extensively characterised datasets
Seek experienced experts
Poor Study Design
No reproducible Validation
Inadequate Clinical Performance
Low specificity and sensitivity
ROC ≤ 75%
Likelihood ≤ 5%
Odd ratio ≤ 1%
Clearly define a clinical question
Use validated analytical method
Carefully consider the experimental design and possible sources of variation and biases
Use appropriate specimens to avoid biases
Use detailed SOPs for selection, collection, handling and storage and analysis

*Figure 7. Common biomarker failure sources and possible solutions*

*Overview of the main reasons of potential biomarker failure in achieving adequate specificity and specificity required in clinical settings. The most relevant issues are related to discovery and validation pitfalls. Additionally, some potential solutions are presented to address these issues. Adapted from (74, 80)*

In terms of the pre-analytical factors, an adequate experimental design is pivotal, which in turn is determined by a clearly defined clinical question. Careful consideration must be given to factors such as; type of specimen and its pathophysiological relevance to diagnosis, statistical power calculation for determining number of samples according to an estimation of the biological and technical variability. On the other hand, possible sources of biases must be examined in order to control confounding variables and other variability sources. Consequently, comprehensive clinical and technical information of patients and samples should be kept, as well as clear criteria of inclusion and exclusion of patients defined. As illustration of the effect of these confounding variables on the stability of biomarker candidates, several publications have reported that factors such as longitudinal variability (aging), sexual dimorphism, genetic background (*i.e.* race) and environmental conditions influence the expression of plasma proteins and metabolites subpopulations (74, 91, 92).

Although this scenario stresses the importance of detailed SOPs for sample collection, selection, handling and storage, it has been reported that failures at this stage are worryingly frequent in discovery studies and substantially impairs successful biomarker development (73, 74, 80, 87).

The study execution stage comprises *(1)* analytical factors related to methodological performance and *(2)* post-analytical factors that mainly include statistical and bioinformatic analysis approaches. A wide range of discovery platforms based on high-throughput technologies and sensitive detection devices have been developed, aimed to analyse molecular composition to large-scale (77). Typically, the methods associated to these platforms implicate a considerable number of different steps that consequently have associated technical variability. Therefore, sufficiently standardised methods and quality control processes on critical steps are required to minimise and account for both systematic and random errors.

Bioinformatics and statistical assessment of the data is a crucial stage. The methods of quality control and normalisation of data as well as significance assessment of the differential expression will dictate the biomarker candidates for validation and prioritise them through the development pipeline. Figure 7 summarises some of the problems frequently found in the data analysis. Often the selection of individual candidates into panels is based on significance of fold-changes, and conventional statistical approaches such as t-test, ANOVA and Kolmogorov-Smirnov are required in conjunction with more recent comprehensive approaches such as forward stepwise multivariate/logistic regression modelling and support vector machine analysis (SVN) (77). Additionally, software packages and tools for data visualisation, correlation (linear and non-linear), integration, retrieval and storage are essential for mining data generated from high-throughput technologies such as proteomics platforms.

Once a multi-marker panel has been set up for validation, its performance is evaluated in terms of sensitivity (SN: ability of identifier to identify true positives 'true-positive rate or TPR') and specificity (SP: ability of identifiers to detect the absence of the disease). The correlation between these factors and therefore the test performance identifying the disease can be studied using a Receiver-Operating Characteristics (ROC) curve plotted against the clinical 'gold standard' test (84).

The ROC curve method plots sensitivity (TPR) in function of the false-positive rate (FPR) (1 - Specificity) at different threshold settings. The area under the ROC curve ($0 < AUC < 1$) provides a statistical summary that allows to evaluate the performance of a classifier and is equivalent to its probability to rank a randomly chosen patient higher than a randomly given control which is equivalent to the Wilcoxon test of ranks (93). Another important tool for performance evaluation is the positive and negative predictive values, which usually are expressed as percentages and describe the probability that those individuals testing positive and negative are true hits, respectively (84). In Table 1, some of the more frequently considered parameters for performance evaluation are presented.

*Table 1. Clinical meaning of parameters of biomarkers performance.*

*Taken and adjusted from (84)*

| *Result* | *Decision* |
|---|---|
| True positive (TP) | Correct hit |
| True negative (TN) | Correct rejection |
| False positive (FP) | Type I error (false alarm) |
| False negative (FN) | Type II error (true miss) |
| True positive rate (TPR) (Sensitivity - SN) | TPR = TP/P = TP/(TP + FN) |
| False positive rate (FPR) | FPR = FP/N = FP/ (FP + TN) |
| Accuracy (ACC) | ACC = (TP + TN)/(P + N) |
| Specificity (SP) | SP = TN/(FP + TN) =1 – FPR |
| Positive predictive value (PPV) | PPV = TP/(TP + FP) |
| Negative predictive value (NPV) | NPV = TN/(TN + FN) |

### 1.2.4 Biomarkers in Tuberculosis Diagnosis

Despite the policies and efforts addressed to control tuberculosis, the global incidence rate is just decreasing 1.5% annually (28). There is an urgent need for finding the millions of cases that are missed each year due to the lack of suitable accurate point-of-care diagnostic tests, particularly in locations with the highest disease burden (94).

The targets for 2025 and 2035 proposed by the WHO in the EndTB initiative (29) are highly unlikely to be met without novel diagnostic tools with high performance and suitable for using close to the patients in affordable diagnostic algorithms (25). Biomarkers of early tuberculosis disease diagnosis and vaccine-induce protection against tuberculosis were recognised as the type of markers representing the largest impact on eradication strategies (95). Furthermore, in a more recent meeting summary reported by Denkingher, *et al.* in 2015, four high-priorities for diagnostic needs in tuberculosis control were established: (1) A sputum-based substitutive test of smear-microscopy; (2) A non-sputum-based biomarker test for the different forms of disease differential diagnosis; (3) a simple, low cost triage test as a rule-out test suitable for community health workers; and (4) a rapid test for drug susceptibility (25). This study comprised a survey to stakeholders including representatives from national tuberculosis programs, clinical/clinical laboratory experts and researchers from high, middle and low-income countries in conjunction to literature review. Consequently, enabling new biomarkers is a pivotal task to make the strategies of control and elimination of tuberculosis realistically feasible.

It is generally accepted that a tailored combination of biomarkers will exhibit an improved performance when comparing to a single marker. When a multiple panel is being designed, it is important to consider not only the performance of each predictor but also its biological independency. In other words, the multiplexed test will be more likely to give a better performance if the included markers are not biologically related (74, 96).

Especially relevant in this context are the high-throughput technologies, which ideally enable the unbiased screening of thousands of molecules (gene products, RNAs, proteins, metabolites, etc.) simultaneously. Such combination of predictors is typically referred in the literature as *'biosignature'* (5, 96, 97). New biosignatures for tailoring novel high performance diagnostic rapid-tests have a pivotal importance for tuberculosis control strategies. Various platforms have been employed in the quest for new biomarkers for tuberculosis such as whole genome sequencing (WSG), transcriptomics, proteomics and metabolomics.

- *Gene signatures for tuberculosis*

Nucleic acid amplification technologies (NAATs) and WGS have been used for diagnosis confirmation, outbreak identification and information about antibiotic resistance (98-101). Studies have proposed WGS for *M. tuberculosis* identification from liquid culture and uncultured isolates expediting the diagnosis between 1 to 3 days (99, 102). Despite of recent technologies facilitating sample preparation and gene expression such as isothermal amplification (LAMP) with fluorescent endpoint detection and manual cross-priming amplification for *M. tuberculosis* identification in sputum, some important obstacles persist for point-of-care application particularly related to sensitivity (98). Firstly, DNA-based diagnostic tools may not distinguish between cleared infections and the active disease due to DNA from dead bacteria remains detectable (103). Secondly, these approaches still require a minimal infrastructure and trained personnel which preclude their application as screening tools.

- *Transcriptomics*

The transcriptome is defined as the wide collection of RNA transcripts including mRNAs and non-coding small-RNAs expressed under given conditions. The analysis of the genome-wide gene expression through RNA sequencing or gene chip microarrays leads to the identification of the expressing genes and the measurement of transcript abundance (104). Particularly relevant to tuberculosis, the host transcriptional response resulting from the diverse *M. tuberculosis* infection stages has been examined in both blood and tissue. Nevertheless, only studies conducted in blood samples will be referred considering their diagnostic relevance. Over 20 papers exploring the transcriptome signatures derived from the diverse host-pathogen interactions have been published in the last 10 years, however no diagnostic tool resulting from this approach exists. Haas, C. T. *et al.* (2016) present a review of the available literature concerning the blood transcriptome signatures for tuberculosis. Among the limited number of studies aimed to discriminate tuberculosis from healthy controls, a very poor overlap between transcriptional signatures was found. Among the causes of discrepancies, differences in study design, variations in patient demography and profiling methodologies were pinpointed. Additionally the use of whole, depleted or fractionated blood was highlighted as an important confounding factor. However functional annotations such as FCGR signalling, interferon signalling, and complement pathways were observed common in the active tuberculosis signature (103).

Important efforts are being made to reduce the cost of gene expression based-tests improving efficiency by reducing the cost of PCR multiplexing. A recent study established a 4-gene signature able to distinguish tuberculosis patients from healthy individuals reaching a sensitivity of 88% and specificity of 75% (105). However, this test cannot outperform the Xpert MTB/RIF test specificity of 100%. Additional efforts have been made to explore RNA blood profiles for tuberculosis disease risk signatures, a recent study identified a 16-gene signature for tuberculosis progression (66.1% sensitivity and 80.6% specificity) (106) which open up the possibilities for targeted interventions.

Notwithstanding, transcriptomic studies generate critical large-scale information for understanding the complex pathophysiological course of the infection, assuming events in the periphery relate to those at the site of disease, but its application to point-of-care diagnosis is limited in terms of performance, infrastructure and training hampering full translation to clinics.

- *Metabolomics*

The metabolome is described as broad collection of small molecules or metabolites (sugars, lipids, nucleotides, amino acids, *etc*.) present in a clinical sample. These diverse analytes are typically explored using mass spectrometry and magnetic resonance. Most metabolomics studies have aimed to ascertain tuberculosis pathogenesis rather than verify diagnostic value (103, 107-110). Metabolites in plasma, urine, breath, sputum and cerebrospinal fluid have been examined. The signatures generated compromise host and pathogen derived candidates. Particularly relevant to point-of-care diagnosis an active tuberculosis signature encompassing 42 features, mostly related to a dysregulated tyrosine - phenylalanyl metabolism was obtained with an AUC of 0.85 (111). More recently, a panel of 4 metabolites for urine testing: sialic acid, diacetylspermine, neopterin, and N-acetylhexosamine exhibited ROC AUCs >80% in a blinded validation cohort, providing a potential non-invasive signature for tuberculosis (112).

However, heterogeneity of the chemical functional groups of metabolites results in a wide variety of physicochemical properties, and this imposes a significant analytical challenge to translation of metabolic signatures to POC settings. Additionally, the metabolome is particularly susceptible to a diversity of factors such as medication, diet, stress, comorbidities and environmental factors, therefore study design and rigorous validation are crucial.

- *Proteomics*

Typically proteomics is referred as the quantitative analysis of the protein composition at given time and conditions. The proteins are considered as the ultimate biological effectors and therefore examination at protein level, rather than transcripts and genes, reflects cellular functions and pathophysiological processes. On the other hand, a low correlation between gene number copy, transcripts and protein expression indicates that analysis of genome and RNA level not necessarily is a direct measurement of active biological functions (113). Unbiased screening of protein content in biological samples is performed usually using mass spectrometry preceded by fragmentation of proteins/peptides. A detailed examination of this approach is presented in section 1.4. The first identified protein fingerprint capable to discriminate tuberculosis patients included; serum amyloid A, transthyretin, neopterin and C reactive protein with specificity and sensitivity of 74% and 88%, respectively (114).

Many studies have been conducted in the field of tuberculosis proteomics, nonetheless, proteomics imposes many analytical challenges (36, 114-121). Protein biomarker translation to clinics has been hampered by (1) variability in reported biomarkers and (2) biased candidate validation. Differences in biomarker identification can be caused by limited coverage of the proteome due to the wide dynamic concentration range of proteins in biological matrixes, variability in proteomics techniques as well as differences in study design and statistical analysis. Validation is biased for availability of antibodies or ELISA kits and arbitrary inclusion/exclusion criteria of candidates (103). However, there are some common proteins significantly dysregulated in active tuberculosis; selected examples include CD14, S100A proteins, apolipoproteins, fibrinogen, orosomucoid and serum amyloid A (103). A recent study generated two biosignatures for active tuberculosis (AUC 0.96) and HIV co-infected patients (AUC 0.95) using a nested co-validation procedure (117).

Although, significant improvements are required in terms of method standardisation and candidate qualification to achieve successful translation to the clinic, proteomics is a promising approach for diagnostic biomarkers discovery and validation not only because proteomics data correlates better to biological processes but also because it captures post-translational processes such as protein turnover. Critically for tuberculosis diagnosis, protein signatures are highly conducive to rapid test devices and ongoing work is being conducted on this field, including colorimetric gold nanoparticles on paper-based devices and label-free biosensors (122-124).

## 1.3 Plasma Proteomics

As was briefly described in section 1.2.4, proteomics can be defined as the comprehensive study of a proteome and usually is focussed on identifying and quantifying a diverse collection of proteins expressed as a function of time and cellular localisation using mass spectrometric techniques. Proteins are the direct functional effectors of gene expression in organisms and therefore complex processing and interactions at protein level precisely define the state of a living system at a given point of time (31, 125).

Defining a proteome is a highly challenging task considering biological diversity and analytical complexity. Features such as post-translational modifications, isoforms, turnover, dynamic abundance and interactions define molecular diversity.  In addition, biological samples are highly complex analytical matrixes, which exhibit a wide dynamic concentration range comprising a broad spectrum of proteins with diversity of physicochemical properties as molecular size, hydrophobicity, and isoelectric points (35, 126, 127).

The abundance distribution of a proteome typically exhibits a nearly Gaussian distribution on the logarithmic copy number scale as is shown in the Figure 8. Examining the abundant portion of the proteome is a very productive task since sensitivity directly increases with every order of magnitude in number of copies per cell: less sample is required for identifying about a thousand additional proteins. Conversely, the low abundant proteome identification would require either larger amounts of sample or striking innovation in the methods (128).

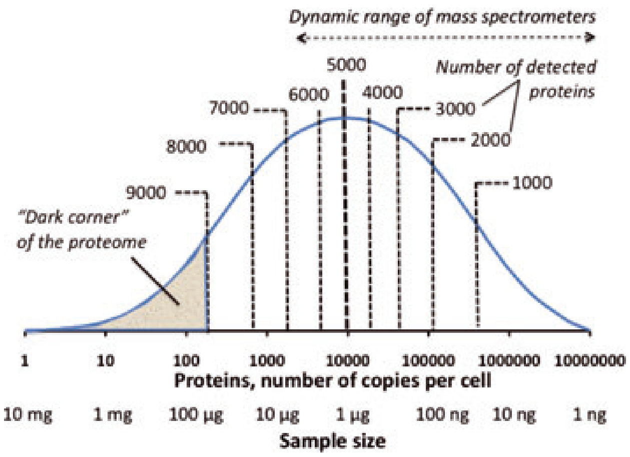


***Figure 8. Proteome abundance distribution.***

*Abundance of the proteins exhibits a bell-shape distribution. Current dynamic range of instrumentation allows efficient exploration of the abundant portion of the proteome. However, a wider coverage of the proteome requires considerably larger amounts of material. The analysis of the 'dark corner' is a particularly challenging task that will require notable improvement in both methods and instrumentation. Taken from (128).*

One of the most critical steps in both biomarker discovery and proteomic experimental design is the selection of the sample. Typically, cell lines and tissue samples exhibit lower complexity than specimens such as proximal fluids and blood-derived samples. Such complexity increases the dynamic range, which directly influences the efficiency of identification. There is a whole field of proteomics research devoted to profiling and quantifying diverse biological matrixes with the aim to define novel biomarkers for diagnosis, prediction, pharmacodynamics and surrogate endpoints (76). Plasma/serum is a common choice, considering the ease of collection, high concentration of protein and its relevancy as indicator of the overall physiological state of an individual since it contains proteins secreted, shed and released from all tissues and cells (33).

Although both plasma and serum are part of the blood, there is an ongoing debate about selection of plasma or serum as matrix for biomarker discovery. Serum is defined as the liquid fraction resulting from whole blood clotting usually under glass/silica-based activation, centrifugation and collection of the supernatant. On the other hand, plasma is obtained when blood is treated with an anticoagulant (129). Serum is recognised as a more heterogeneous matrix than plasma, since its collection depends upon a biochemical process that is regulated for many parameters as temperature, time for clotting and medication. Furthermore, some proteins can be bound to the clot or be released/activated during aggregation in an uncontrolled fashion, which is the case of some tissue inhibitors of metalloproteinases (TIMPs) and matrix metalloproteinases (MMPs) (130). These conditions are very difficult to standardise, therefore plasma is preferred for biomarker discovery.

The complexity of human plasma, which contains over 20000 proteins with a wide dynamic concentration range spanning 12 orders of magnitude, is one of the greatest obstacles for plasma proteomics studies (131). Moreover, only 22 proteins account for the 99% of the total protein content in plasma where serum albumin represents 50% of the overall protein composition (132). Figure 9 illustrates that low abundance proteins exhibit a higher diversity, being originated from different tissues in comparison to the most abundant proteins, making this protein subpopulation very promising for biomarker discovery (34). Importantly, Figure 9 highlights that proteins clinically relevant as biomarkers are expressed in low concentrations (μg/mL to pg/mL), however, most of profiled proteins in regular plasma proteomics studies are expressed in the range of mg/mL to ng/mL (133). This figure clearly depicts the considerable challenge of improving limits of detection and quantification of the proteomic platforms and methods to capture the low abundance proteome.

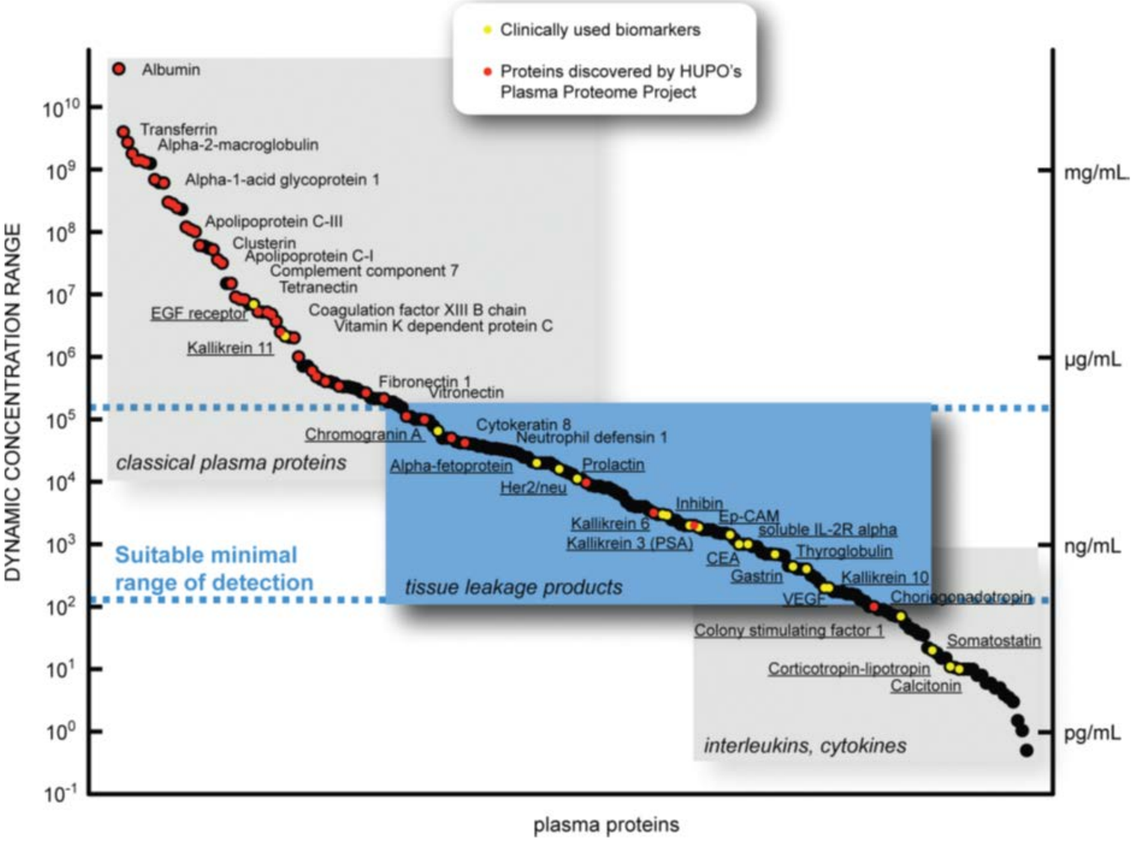


*Figure 9. Protein dynamic range in human plasma*

*The large dynamic range of proteins in the plasma, spanning from mg/mL to pg/mL is shown and proteins are grouped into three categories: classical plasma proteins, tissue leakage products, hormones and interleukins/cytokines. Proteins discovered by HUPO's Plasma Proteome Project are indicated in red and* *clinically used biomarkers in yellow. Used biomarkers are physiologically expressed in the order of µg/mL to ng/mL, some of them under the minimal range of detection of common proteomic platforms. Taken from (34, 133)*

A critical consideration for biomarker discovery based on plasma proteomics is the abundance stability of plasma proteins associate to host factors. A longitudinal study conducted on plasma samples collected from monozygotic and dizygotic twins at intervals of 2–7 years, demonstrated the patterns of abundance of plasma proteins are under regulation of genetic, environmental and longitudinal factors, suggesting that plasma biomarkers require calibration against temporal and genetic features (91).  Furthermore, Al-Daghri, N. M, *et al.* (2014) demonstrated differential proteomic profiles in nondiabetic overweight and obese women and men. Therefore, the experimental design for biomarker discovery should consider to host factors such as ethnicity, age, body mass index and smoking status.

Mass-spectrometry based proteomics is continuously developing powerful platforms for the analysis of complex matrixes and the profiling of thousands analytes in single experiments. The more recent instruments can cover up to 5 orders of magnitude achieving a resolution up to 500,000 FWHM (Full Width at Half Maximum), with isotopic fidelity up to 240,000 FWHM at m/z 200 (133, 134). However, the presence of peptides originated from highly abundant proteins induces an ionisation suppression effect on the peptides from the lower abundance proteins, thus handicapping the detection of the latter. Moreover, at a given time-dependent MS analysis event the amount of ions being stored, transmitted and detected is strongly determined for the stoichiometric relation among different ionic species. Consequently, in plasma proteomics, it is critical to reduce the sample complexity by depletion or fractionation prior to MS analysis in order to maximise the proteome coverage.

Despite the complexity of human plasma, most of clinical diagnoses (>70%) are blood informed and almost half of chemical pathology is dominated by proteins (135). Therefore, plasma biomarker discovery is an urgent need in the pursuit of improved/novel diagnostic tools.

### 1.3.1 The Depletion Dilemma

The multiple strategies developed to circumvent the analytical challenges imposed by the large dynamic concentration range of plasma can be categorised into three main methods: *depletion, equalisation* and *hyper-fractionation.*

*Depletion* of plasma using immunoaffinity capture is certainly the most common method used routinely on samples before interrogation by MS (31). This approach uses columns modified with antibodies that retain specific proteins. Depletion up to twenty proteins has been conducted (136) and a variety of formats are commercially available including spin-columns (137), and online-LC columns (137). Additionally new methods for depletion such poly(N-isopropylacrylamide-acrylic acid) hydrogel particles were recently explored (131). Nevertheless, removal of highly abundant proteins leads to co-removal of less abundant proteins due to mutual interaction through non-covalent forces and non-specific interactions to the antibodies can increase this effect (92, 136, 138). It is well recognised that albumin binds a variety of ligands including small molecules, peptides and proteins. Additionally, different species of high abundant proteins may be biologically informative, for instance albumin can be modified at both concentration level and oxidation state in response to disease (137). Consequently, the 'albuminome' offers an additional level of complexity to the plasma proteome analysis.

Multiple reports suggest that proper quantification of proteins for biomarker discovery purposes should additionally include the bound fraction resulting from depletion methods. Scumaci, D. *et al.* (2011) identified 67 protein ligands of human serum albumin using an affinity approach. Yadav, A. K., *et al.* (2011) quantified 101 proteins with high confidence (<1% FDR) in the bound fraction from three different multi-affinity removal systems and Koutroukides, T. *et al.* (2011) profiled 170 proteins in the depleted fraction using a top-20 immunodepletion column (139). Although the study of the 'albuminome' or 'depletome' may play an important role in clinical proteomics, depletion introduces significant biases to the plasma proteome profile and therefore to the biomarker discovery pipeline.

One alternative to depletion is the *equalisation* of low abundance proteins using combinatorial peptide ligand libraries, which is the basis for ProteoMiner Technology$^{TM}$ (136). Briefly, this strategy utilises a large and highly diverse bead-based library of combinatorial hexapeptides ligands, which simultaneously dilutes the high abundance proteins, when the ligands of this particular population are rapidly saturated. Conversely, low abundant proteins are concentrated on their specific ligands (136, 140, 141). Zhao, Y. *et al.* (2016) conducted a comparison of five analytical strategies for plasma proteome profiling. Proteome equalisation showed limited low-abundance proteins enrichment and the obtained proteome was biased toward low molecular weight and basic proteins (142). Additionally, the commercially available kits required a relative large volume of sample (>1.0mL) which in many cases can represent an important limitation.

Additionally to the depletion and equalisation strategies, a relatively simple approach is the *hyper-fractionation* at both protein and peptide level which minimises the influence of the high abundant proteins and reduces the high complexity of this biological matrix. Fractionation methods exploit the physico-chemical properties of the proteins/peptides to simplify the sample, thereby generating multiple simpler fractions. There is a wide repertoire of methods to fractionate protein mixtures with a variable degree of throughput and coupling; ranging from SDS-PAGE electrophoresis, off-gel Isoelectric Focusing (IEF) to various liquid chromatographic techniques. Recent studies have proved that extensive multidimensional chromatographic prefractionation leads to greater coverage of the plasma proteome over other strategies (35, 92, 142). The multidimensional strategy used in this study will be discussed in further detail in the next section.

### 1.3.2 MudPIT strategy

Garbis S. *et al* (2011) first published the multidimensional method applied in this work. This strategy comprised three different but complementary liquid chromatographic chemistries: size exclusion chromatography (SEC), zwitterion-ion hydrophilic interaction chromatography (ZIC-HILIC) and reversed-phase nano-ultra performance chromatography (RP-nUPLC). Notably, when samples pre-fractionated using SEC were compared to depleted samples, the first approach demonstrated a significantly higher serum proteome coverage. 1955 proteins were identified (FDR ≤ 5%) compared to 563 and 499 proteins identified from depleted samples analysed using ZIC-HILIC and strong cation exchange (SCX), respectively. More recently in 2014, this approach was adjusted to conduct a quantitative study using iTRAQ. SEC followed by offline high pH reverse phase (RP) chromatography (C8 chemistry) for labelled peptide prefractionation and RP-nUPLC. 2472 proteins were identified (FDR ≤ 5%) in serum samples and 248 were significantly modulated. This technique has proved to be a powerful tool for plasma/serum proteomics. The chromatographic principles of each technique involved will be described in further detail in the next sections.

#### 1.3.2.1 Size Exclusion Chromatography (SEC)

Size exclusion chromatography is a general term used to refer to the separation process of molecules according to their size, more exactly their hydrodynamic ratio, when a mobile phase flows through a packed bed of porous material. Separation is achieved by differential pore permeation; the effective accessible pore volume is greater for small molecules than for larger analytes. Figure 10A-B depicts this principle: the largest molecules elute first from the column since they present the shortest retention times in the pores of the packing bed (143).

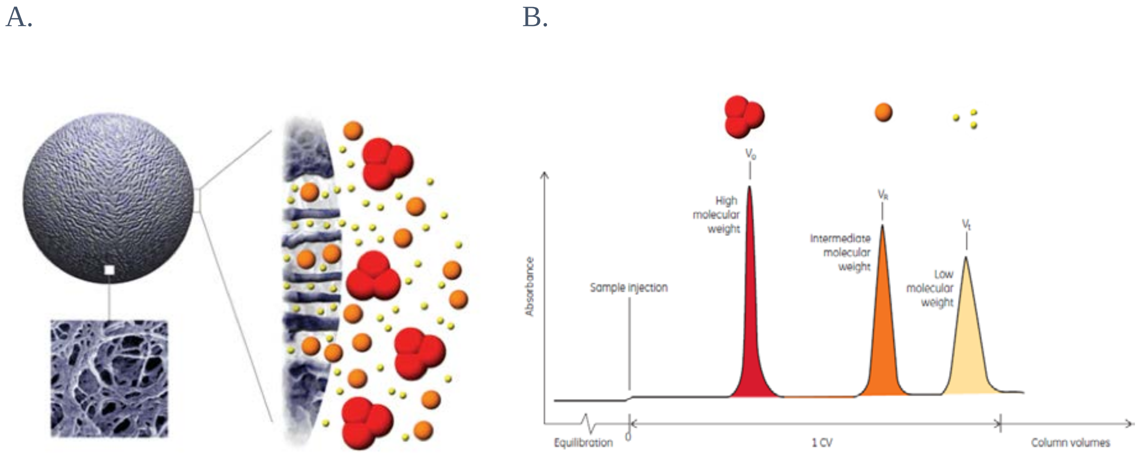


***Figure 10. Size Exclusion chromatography principle.***

***A.*** *Schematic structure of the particles of SEC medium with an electron microscopic magnification, the architecture of the pores is visible.* ***B.*** *Schematic chromatogram illustrating differential elution according to the molecular size. Adapted from (144)*

SEC packing typically consists of a porous matrix of inert spherical beads with physical and chemical stability. The packed medium is equilibrated with the eluent to fill the space between particles and the pores of the matrix. Samples are eluted isocratically with a final wash step with the solvent to remove the molecules that might have been retained and prepare the column for the next run. The resolution in SEC depends mainly on selectivity of the medium and its efficiency to achieved minimal peak broadening. The selectivity of a SEC medium depends on its pore size distribution, some common media for SEC are: silica, superdex (cross-linked dextran and agarose), sephadex (dextran) or agarose (144).

Since SEC is performed isocratically, the pH, composition or ionic strength of the eluent and sample buffer does not directly affect resolution as long as these conditions do not affect the molecular size and stability of the packing material. Chaotropic agents such as urea and guanidine hydrochloride can be used to increase the solubility of proteins and as main component of the eluent allows separation under denaturing conditions, thereby disrupting all the non-covalent interactions and hydrogen bridges (144).

SEC is widely used to separate therapeutic proteins, encapsulated drug-loaded liposomes from free drugs, extracellular vesicles, polymer chemistry and it has proved a good performance to isolate exosome-derived proteins (145). Although SEC provides a limited resolving power compared to other chromatographic techniques used in proteomics, it is considered a comprehensive technique for preparative purposes conducive to classify heterogeneous mixtures of biopolymers according to a general property: size. Additionally, SEC is relatively simple and robust. These particular characteristics make SEC uniquely useful as first step of a multidimensional pipeline for plasma/serum proteomics.

In the particular workflow applied in this study, plasma/serum samples are prefractionated using SEC under denaturing conditions (6M guanidine hydrochloride), then the fractions are desalted before trypsin digestion. Tryptic peptides are iTRAQ labelled for quantification. Pooled labelled peptides are fractionated using high pH reverse phase HPLC and each fraction is further online separated using RP-nUPLC coupled to ESI-MS/MS analysis. (Further details in Methods)

#### 1.3.2.2 Reverse Phase Chromatography RP-HPLC

RP-HPLC is by far the most extensively used chromatographic mode for biomolecule analysis, its wide range of mobile and stationary phases and online coupling to sample injection and detection systems, particularly MS, make it ideal for shotgun proteomics. RP-HPLC comprises a polar mobile phase and a nonpolar stationary phase, which typically is composed of spherical silica particles derivatised with hydrocarbon moieties. The variety of starting silica and methods of surface derivatisation provide a wide variability in separation and retention properties (146). Proteins/peptides differentially bind the stationary phase and the elution time differs according to their hydrophobicity.

The most common hydrophobic ligands are $C_{18}$, $C_8$, $C_6$ or $C_4$, subscripts indicating the number of carbons in the aliphatic chain. The mobile phase normally contains a mixture of water and a water-miscible organic solvent, a pH modifier such as trifluoroacetic, and acetic or formic acid is added to promote positive ionisation of the peptides/protein limiting unwanted ionic interaction with the stationary phase. High-resolution separation diminishes co-elution reducing ion suppression in MS.

### 1.3.2.3 Orthogonality in Multidimensional separations

Analysis of highly complex mixtures requires techniques that provides the maximum separation possible. In the case of iTRAQ quantification, which involves pooling the trypsinised protein content of eight different samples, it is key to use a multidimensional strategy. The combination of two or more different chromatographic techniques within the same system significantly increases the peak capacity. In order to effectively exploit the high peak capacities, the mechanisms of separation in each dimension must be independent from each other or chromatographically orthogonal (147). For example, a high orthogonal system could involve a hydrophilic separation in the first dimension and a hydrophobic separation for the second dimension. These two opposite physicochemical properties would increase the separation space among the components of the peptide mixture.

To demonstrate this principle Gilar, M. *et al* (2012) compared the orthogonality of different liquid chromatographic chemistries and RP-HPLC using a C18 column (pH 2.6) which is the more widely used online technique coupled to MS analysis in shotgun proteomics. 196 tryptic peptides were separated in a LC x LC mode, including a phenyl reverse-phase column, a pentafluoro phenyl (PFP) reversed-phase column, a C18 column at high pH and a hydrophilic interaction chromatography (HILIC) column. Figure 11 illustrates the different extent of orthogonality in these four different systems. As expected, when two opposite chemistries in terms of hydrophobicity are coupled, *i.e.* HILIC vs. $C_{18}$, the orthogonality is maximised. Conversely, when two modes with similar hydrophilicities are used, the orthogonality between dimensions is significantly reduced. Interestingly, the peptide separation increases when the fractionation is conducted using different pH conditions (alkaline and acidic) but the same column chemistry.

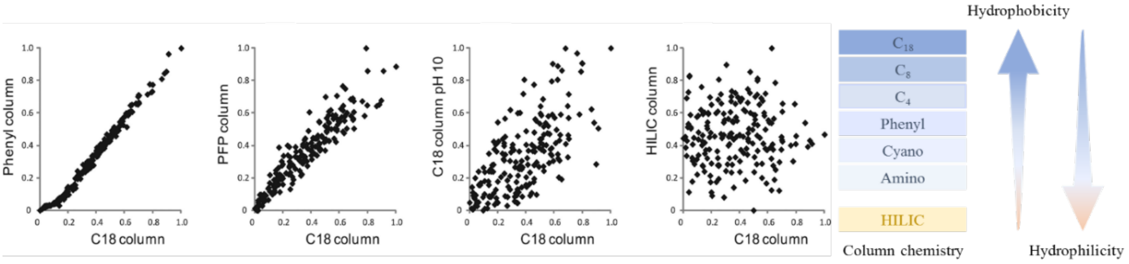


***Figure 11. Graphical representation of orthogonality in an LC x LC system.***

*Normalised sets of data from two-dimensional data. 196 tryptic peptides were separated using four LC x LC configurations. The highest orthogonality is achieved when modes of chromatography in each dimension are based on different chemical interactions. Adapted from (148)*

### 1.3.3 Current Status of Plasma/Serum Proteomics in Tuberculosis

The minimum specifications for a rapid point-of-care diagnostic test for adults that should guide research and assay development are summarised in Table 2. A standalone test for pulmonary tuberculosis that satisfies these requirements would be ideal, however the particular features of this disease such as the wide spectrum of clinical outcomes resulting from the infection caused by a very low bacillary burden, hinders this goal. An alternative option might include a high-sensitivity, low-specificity screening test followed by referral for a definitive confirmatory test (69).

***Table 2. Minimum requirements for the ideal point-of-care diagnostic test for tuberculosis***

Data taken and adapted from (69)

| **Parameter** | **Minimum specification required** |
|---|---|
| Outcome of testing | Initiation of treatment |
| Sensitivity in adults (irrespective HIV status) | Smear-positivity, culture positive PTB: 95%<br>Smear-negativity, culture positive PTB: 60%-80%[a]<br>EPTB: preferred but not required |
| Specificity | Compared with culture: 95% |
| Throughput | 20 tests per day by a single operator |
| Waste disposal | Environmentally friendly disposal feasible |
| Storage and stability | No cold chain required; stable at 30°C for 2 years |
| Instrumentation | Maintenance free, robust in tropical climates, acceptable replacement cost, portable (*e.g.*, in backpack), can be battery operated, shock resistant |
| Operation | Requires minimal instruction |
| Cost | Less than $10 per test |

[a] No consensus about this limit

PTB: Pulmonary Tuberculosis, EPTB: Extrapulmonary Tuberculosis

Although some progress has been made regarding novel biomarker discovery in plasma/serum particularly using various proteomic platforms, the methods used so far provide a very limited proteome coverage. Additionally, poor validation of proposed biomarkers is a common limitation of most biomarker discovery studies. A systematic literature research conducted in 2017 indicates that from 399 biomarkers (non-DNA) reported as tuberculosis biomarkers between 2010 and 2015 only 12 were validated in a prospective study, from which only one has been reviewed by the WHO (LAM in urine) (149). Table 3 summarises the serum/plasma proteomic studies for new biomarker discovery for diagnosis of the active disease published in the last 10 years.

*Table 3. Serum proteomic studies for novel biomarkers in active pulmonary tuberculosis*

| Authors | Year | Genetic Background | Proteomic Strategy | Proteome Coverage | Proposed Candidates | Performance Reported |
|---|---|---|---|---|---|---|
| Agranoff, D., *et al.* **(114)** | 2006 | African | ProteinChip CM10 Array<br>SELDI-ToF MS<br>MALDI-ToF MS | 219 peak clusters from m/z spectra in the range<br>2000–100000 m/z | SAA1<br>TTR<br>CRP<br>Neopterin | Sensitivity 88%,<br>Specificity<br>74%<br>No HIV co-infection |
| Liu, J., *et al.* **(121)** | 2013 | Chinese | SELDI-ToF MS | 251 protein peaks<br>1500 - 15,000m/z | Four peaks (2554,6; 4824,4; 5325,7; and 8606,8 Da) One of them identified as Fibrinogen | Sensitivity 83.3%,<br>Specificity<br>84.2%<br>No HIV co-infection |
| Song, S. H., *et al.* **(150)** | 2014 | No information - Samples collected in Seoul | RP-HPLC Quadrupole ToF, label-free quantification | 518 Proteins | SERPINA1<br>SERPINC1 | AUC 0.947 for alpha-1-antitrypsin (SERPINA 1) |
| Xu, D. D., *et. al*. **(36)** | 2014 | Chinese | iTRAQ-coupled 2D LC-MS/MS | 434 Proteins | APOCII<br>CD5L<br>HABP2<br>RBP4 | Sensitivity 93.42%,<br>Specificity<br>92.86% |
| Xu, D., *et al.* **(119)** | 2015 | Chinese | iTRAQ 2D LC-MS/MS | 434 Proteins | S100-A9<br>SOD3<br>MMP9 | Sensitivity 92.5%,<br>Specificity<br>95% |
| Achkar, J. M., *et al.* **(117)** | 2015 | Asian/Black/ Hispanic/ Caucasian (NYC) | LC-Q-ToF<br>MRM-MS validation | Total coverage not reported 165 proteins modulated | CD14, SEPP1, SELL, TNXB, LUM, PEPD, QSOX1, COMP, APOC1 (HIV-) | AUC 0.96 for HIV−TB<br>AUC 0.95 for HIV+ TB |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | CD14, SEPP1, PGLYRP2, PFN1, VASN, CPN2, TAGLN2, IGFBP6 (HIV+) | |
| Li, C., *et al.* **(151)** | 2015 | Chinese | iTRAQ<br>LC-ToF/ToF | 434 Proteins | SHBG | Sensitivity 75.6%<br>Specificity 91.5% |
| Jiang, T., *et al.*,**(152)** | 2017 | Chinese | iTRAQ LC-Triple ToF | Total coverage not reported 79 proteins modulated | SAA<br>PROZ<br>C4BPB | Sensitivity<br>97.08%<br>Specificity<br>95.45% |
| Chen, C., *et al.* **(153)** | 2018 | Chinese | iTRAQ 2D LC-MS/MS | 716 Proteins | ENG (HIV+) | Not reported |

Most of the biomarkers reported as upregulated in active tuberculosis are associated to phagocyte migration, neutrophil and granulocyte activation, inflammation, stress response, innate immune response and acute response. In contrast, proteins downregulated during active tuberculosis are associated with lipid processing and transport. Limited knowledge of the differential protein expression during tuberculosis infection course is available. Recently, Scriba, T. J., *et al.* (2017) enrolled a cohort of 6,363 adolescents and followed them for at least 24 months to conduct a longitudinal study for tuberculosis progression. Whole blood transcriptomic analysis by RNA sequencing and plasma proteome analyses using DNA aptamers (SOMAscan) was performed to interrogate progression makers in tuberculosis. Transcriptomic and proteomic data suggested that the complement cascade modules were upregulated earliest during progression simultaneously with IFN responses. Subsequently, changes in myeloid inflammation modules and blood coagulation appeared around 200 days before the tuberculosis diagnosis. Other modules associated with tissue remodelling, platelet activation and haemostasis emerged within 200 days before tuberculosis diagnosis (154). This study suggested that tuberculosis progression involves a slow progression through stages of inflammatory perturbations that results in strong inflammatory responses emerging proximal to the active disease. Furthermore, Hamilton, K., *et al.* (2018) in a systematic review of active-finding strategies for tuberculosis in a homeless cohort found that around 30% of asymptomatic or minimally symptomatic had smear-positive disease and 15% were cough aerosol-positive (155). These recent studies highlight the complexity of the tuberculosis natural history that it is not accurately represented by the clinical manifestations, which usually determine the initiation of the diagnostic algorithms.

Therefore, more precise biosignatures that better encompasses this complexity are urgently required for active case finding strategies in tuberculosis.

Despite early studies, which investigated if serum protein biosignatures can be used to discriminate pulmonary tuberculosis patients, and several studies since, a universal proteomic profile remains elusive. Variable resolution of the utilised techniques and methods, case definition, statistical assessment of the differences and limited or absent validation of the diagnostic performance have contributed to these discrepancies. On the other hand, proposed biosignatures are not consistently: (1) evaluated for their diagnosis potential; (2) cross-validated, or (3) evaluated with external databases (103). Improvements in standardisation and validation of proteomic platforms and methods are essential to increase accuracy and reproducibility of biomarker panels and facilitate their translation to clinical settings.

### 1.4 Shotgun Proteomics

The most widely used strategy for in-depth protein profiling is known as shotgun proteomics and, as a peptide-centric method, is conducive to biomarker discovery allowing the analysis of thousands of proteins from complex samples. Proteomic strategies can be categorised by the molecular level at which the analysis takes place. Figure 12 depicts the three main approaches. The analysis of intact proteins so-called *top-down proteomics* provides information on protein mass and amino acid sequence. The proteins are separated and subjected to diverse fragmentation methods such as electron capture dissociation (ECD), collisional-activated dissociation (CAD), or electron-transfer dissociation (ETD). Among the advantages of this method are; reduced sample preparation and connectivity information of different posttranslational modifications (PTMs) and isoform occurrence. However, this method presents a low sensitivity since ionisation and fragmentation in gas phase of proteins is considerably more challenging than peptides, which is precisely the rationale of bottom-up proteomics (31, 156).

The vast majority of proteomic studies involve the digestion of the proteins and the subsequent analysis of the derived peptide mixture by MS. This approach is so-called *bottom-up proteomics*, and when this strategy is applied to a mixture of proteins is referred as *shotgun* proteomics (157). The extent of proteolysis will define if the experiment is middle- or bottom-proteomics. Typically a bottom-up experiment will analyse peptides among 500 to 3000Da which means a sequence no longer than ~26 residues (156). Multiple advantages of working at the peptide rather than protein level include; more versatile chromatographic separation, fewer charge states, lower molecular mass, and more efficient ionisation and fragmentation in gas phase. These features have placed the bottom-up strategy as the preferred for discovery purposes (31, 156, 157).

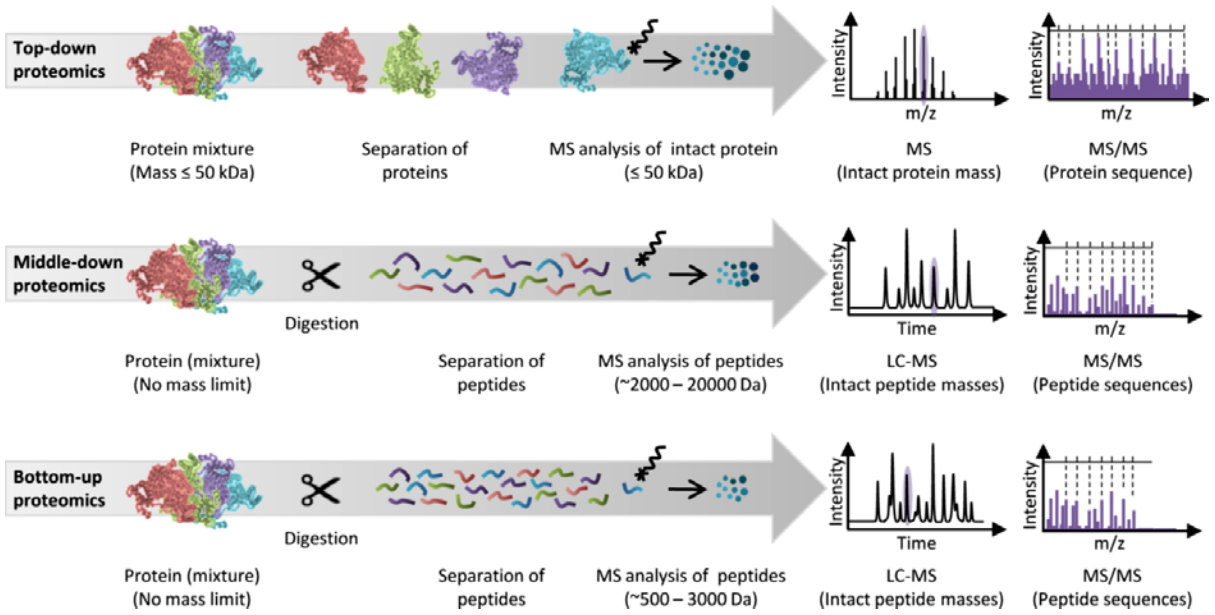

Top-down proteomics
Protein mixture (Mass ≤ 50 kDa)
Separation of proteins
MS analysis of intact protein (≤ 50 kDa)
Intensity
m/z
MS (Intact protein mass)
Intensity
m/z
MS/MS (Protein sequence)
Middle-down proteomics
Digestion
Protein (mixture) (No mass limit)
Separation of peptides
MS analysis of peptides (~2000 – 20000 Da)
Intensity
Time
LC-MS (Intact peptide masses)
Intensity
m/z
MS/MS (Peptide sequences)
Bottom-up proteomics
Digestion
Protein (mixture) (No mass limit)
Separation of peptides
MS analysis of peptides (~500 – 3000 Da)
Intensity
Time
LC-MS (Intact peptide masses)
Intensity
m/z
MS/MS (Peptide sequences)

***Figure 12. Proteomic strategies.***

*Schematic representation of the principles of the three strategies in proteomics. Bottom-up proteomics is a peptide-centric approach. Proteins are extensively hydrolysed and peptides are used to infer the identity of the proteins present in the sample. In middle-down proteomics, the proteins are digested generating longer sequences. Conversely, top-down proteomics rely on the direct analysis of protein mixtures Taken from (156).*

Shotgun proteomics involves several handling steps from the enzymatic digestion to the MS analysis, therefore well optimised protocols are essential to generate reproducible results. Once the peptide mixture is obtained, it is subjected to liquid chromatographic separation coupled to a mass spectrometer where tandem mass spectrometry is required to sequence and quantified the proteome. The main features of this strategy will be discussed in detail through the next sections.

### 1.4.1 Enzymatic Digestion

Trypsin has been recognised as the standard enzyme in the proteomics field and has been heavily engineered to maximise its specificity, activity and stability, minimising autolysis which must be avoided since it results in chymotrypsin-like activity (156). Considering that shotgun proteomics is a peptide-centric approach and therefore the proteolysis step plays a central role, a clear understanding of trypsin is required to understand the current strategies of protein inference in the context of complex peptide mixtures.

Trypsin belongs to the group of extracellular serine proteases and structurally is characterised by two six-stranded beta-barrels exhibiting the classic Greek-key architecture (158). Figure 13A depicts the trypsin fold. The catalytic site is located in the interface of these barrels. Trypsin cleaves at carboxy-terminal of Lys (K) and Arg (R) and the residues responsible of this catalytic activity are the nucleophilic Ser (S) 195 and Asp (D) 105, which are coordinated through electrostatic interactions by the imidazole group of the His 57 (158). Considering that Lys and Arg are relatively abundant residues in the human proteome (Figure 13B) and typically well distributed throughout a protein

(159), the trypsin hydrolysis pattern yields an average of 61 peptides per protein with an average length of 9 residues and a standard deviation of 15. However, if one miscleavage is allowed the average length increases to 14 residues with a standard deviation of 20 (156, 158, 159). Although engineered trypsin exhibits a high specificity, commercially available sequencing grade trypsin achieves <95% of specificity, there are well-recognised causes of miscleavage (Keil rules, 1992). The most widely known occurrence is the miscleavage taking place when Arg or Lys is followed by a Pro (P), explained by the steric hindrance imposed by the Pro on these residues. Two additional configurations can lead to restriction of the hydrolytic activity of trypsin: when two or more positively charged residues follow each other and when there are negatively charge residues in close proximity to Lys or Arg  (158).

A typical bottom-up experiment involves the digestion *in silico* of the proteome under study using algorithms that include some of these rules to ensure the most accurate fit to the experimental trypsin hydrolysis activity on that particular proteome. These theoretical peptides are, then, fragmented *in silico* to generate a theoretical spectra set. Peptide sequencing is achieved through peptide-to-spectrum matching algorithms which compare theoretical and experimental spectra (31, 158). This process will be discussed in further detail in the section 1.4.4.

One important feature of the trypsin digestion is that the generated peptides usually fall within the detection range of the spectrometer, except for the smallest and biggest tryptic peptides. This can be clearly visualised in the Figure 13C where the blue histogram represents the probability of the length distribution of tryptic peptides derived from a digestion *in silico* of the human proteome of UniProtKB/SwissProt allowing one miscleavage. In comparison, the red histogram represents the peptide-length found in a typical shotgun experiment extracted from the PRIDE database repository (158). Hence, the enzymatic properties of the trypsin are critically exploited in two senses; firstly, to hydrolyse a given proteome into complex peptide mixture; and secondly, to simulate *in silico* the digestion of this proteome for inferring the protein identities.

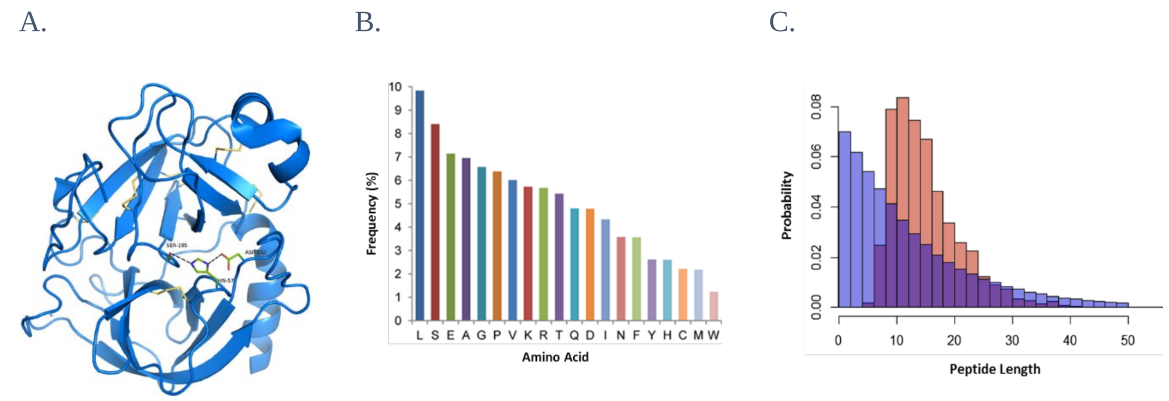


*Figure 13. Trypsin.*

*A. Molecular structure of the trypsin; the residues responsible for the catalytic activity: Ser 195, Asp 105 and His 57, are shown in the interphase of two beta-barrels. B. Amino acid frequency distribution in the human proteome. C. Comparison between probabilities of histograms of peptide length distribution identified from human samples obtained from PRIDE database (red) and peptide length distribution of a trypsin digest of the human complement of UniProt/SwissProt (blue). Taken from (156, 158)*

In addition to the trypsin properties previously discussed such as high specificity, stability and length of tryptic peptides, there is an important additional related benefit; the balance in basicity between the free amine of the N-terminus and the presence of a basic residue in the C-terminus (Arg or Lys) in every tryptic peptide. This characteristic enhances ionisation (positive mode) and fragmentation as explained through the mobile-proton hypothesis (158). Overall, these characteristics have positioned trypsin as the preferred protease for shotgun proteomics.

Alternatively to the use of trypsin alone, a few studies have been conducted exploring combination of two or more proteases with the aim of increasing the proteome coverage and capture more information of posttranslational modifications (160). Alternative proteolysis workflows can include proteases with different specificity to trypsin such as LysC, LysargiNase, ArgC or proteases with a broad specificity such as pepsin and subtilisin. Although the inclusion of multiple proteases into shotgun workflows has proved to increase the proteome coverage from cellular lysates (161, 162), the results for more complex specimens such as plasma can be more discreet. Additionally, some proteases may yield peptides with atypical charge states and lengths that fall outside of the detection limit of the mass spectrometers or coverage of the chromatographic retention capabilities of C18 columns (162).

Experimentally, the shotgun workflow starts with the solubilisation of the protein content using buffers containing detergents such as SDS (sodium dodecyl sulphate) or SDC (sodium deoxycholate) and chaotropic agents as guanidinium hydrochloride or urea. Importantly, these reagents might impair trypsin activity, therefore the working concentrations must be well optimised. Trypsin maintains most of its enzymatic activity at 2.0M guanidine HCl, 2.0M urea, 0.1% SDS, or 2-5% SDC (163, 164). Frequently, this chemical treatment is accompanied by physical disruption of protein and membrane complexes using trituration methods and sonication. Prior to trypsin digestion, two reactions must occur to disrupt tertiary structures in the proteins, thus maximising trypsin access to the hydrolysis sites: reduction and alkylation of disulfide bonds.

- *Reduction of disulfide bonds*

Oxidative folding is crucial for stabilising tertiary structures in proteins. This process is driven by oxidation of sulfhydryl groups only present in cysteine (Cys) residues, resulting in the formation of a disulfide bond so-called *cystine*. Disruption of high structures at protein level is important to ensure availability of the hydrolytic residues to the trypsin and reduce miscleavages.

The first step is the reduction of the disulfide bonds to thiol groups using reducing agents such as 2-mercaptoethanol, dithiothreitol (DTT) or tris(2-carboxyethyl)phosphine (TCEP). The first two reagents contain thiol groups and act by exchanging the thiolate ion ($XS^-$) with the cystine groups. However, this mechanism can be inconvenient when subsequently reacting protein sulfhydryl groups with thiol-reactive probes. Sulfhydryl groups from the reducing agent in excess will compete with those from the protein, and therefore must be eliminated before labelling (165). Trialkylphosphine based reducing agents such as TCEP irreversibly and quantitatively reduce organic sulphides in aqueous solution (166) according to the reaction in Figure 14. This reagent has shown to be significantly faster and stronger reductant than DTT at pH below 8.0 and more stable than DTT at pH below 7.5; additionally it is not necessary to eliminate prior to utilising sulfhydryl reactive labels (165). Although TCEP is widely used in shotgun proteomics, it is charged in aqueous solution and so must be avoided when isoelectric focusing is used.

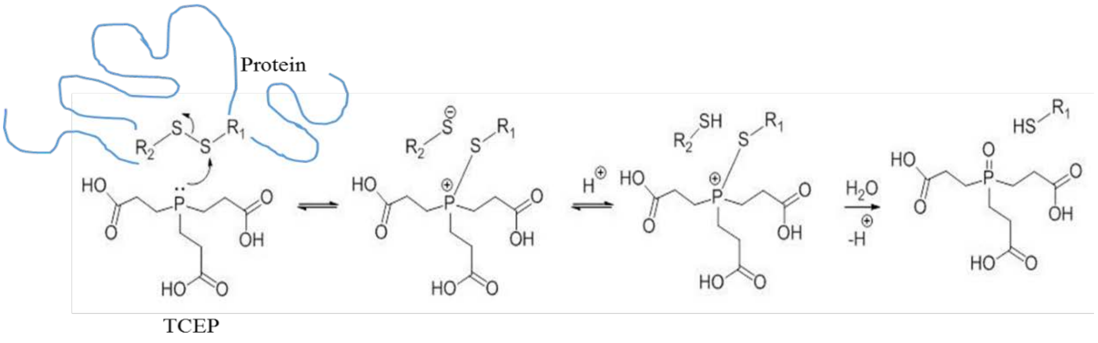


***Figure 14. TCEP reaction mechanism.***

*Schematic representation of reduction of cystine to cysteine by TCEP reaction. Adapted from (167).*

- *Alkylation of sulfhydryl groups*

Once the sulphide bonds are reduced to sulfhydryl groups, these must be blocked to avoid the reformation of new bonds in a random fashion. Thiol groups can be modified by a variety of reagents such as iodoacetic acid (IAA), iodoacetamide (IAN), 4-vinylpyridine and methyl methanethiolsulfonate (MMTS). Although IAA and IAN are by far the most popular for proteomic applications, MMTS is widely used due to its reversibility and is recommended when iTRAQ labelling is attempted (168). Figure 15 presents the sulfenylation mechanism of the MMTS, which attacks the reduced sulfhydryl groups resulting in their alkylation to dithiomethane (-S-S-$CH_3$). This modification must be considered in the workflows for peptide sequencing from spectral information.

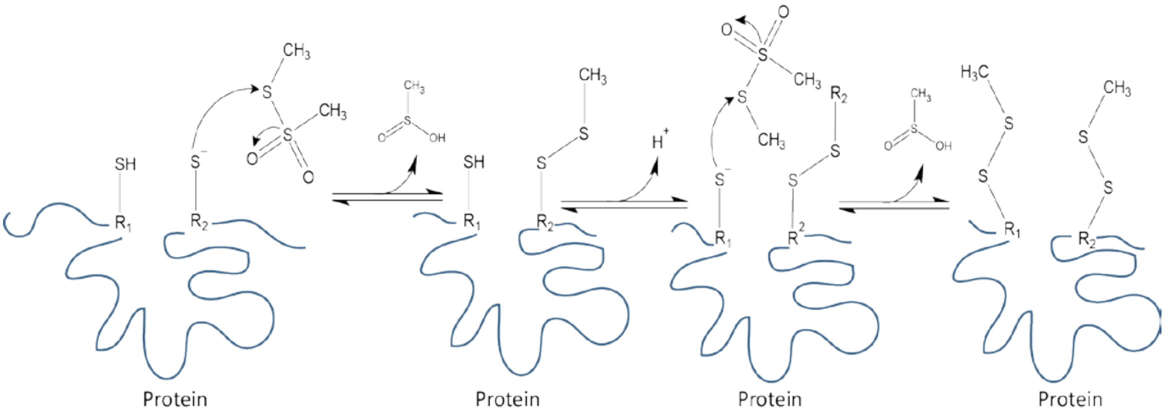


***Figure 15. MMTS reaction mechanism***

*Alkylation of cysteine by reaction with MMTS. Sulfhydryl groups are blocked by the addition of a thiomethane group. Adapted from (169).*

Once the cystine residues are blocked, trypsin is used in an optimised enzyme – substrate ratio between 1:30 to 1:50 (158, 170). This ratio is particularly important since autolytic rate of most exocrine proteases is concentration dependent; an optimised concentration reduces frequency of nonspecific cleavages and partial enzymatic digestion (170). The mixture is typically incubated overnight (~16h) at 37°C to achieve a complete protein digestion.

### 1.4.2 Mass Spectrometry and Tandem Mass Spectrometry ($MS^2$)

The emergence of new technologies and resources such as the completion of the Human Genome Project, advanced bioinformatics and mass-spectrometry (MS)-based profiling have driven the development of high-throughput methodologies for proteome analysis (32, 171). Mass spectrometry has emerged as the central tool for large-scale proteome analysis. By definition, a mass spectrometer consists of an ion source, a mass analyser that determine the mass to charge ratio (*m/z*) that is used as a molecular identifier and a detector which registers the number of ions at each m/z value (172). Figure 16 illustrates the general workflow of MS-based proteomics including the different instrumentation available (173, 174). MS-based proteomics involves generation and detection of charged peptide (shotgun proteomics) or protein (top-down proteomics) ions in the gas phase (mainly in positive ion mode *via* protonation).

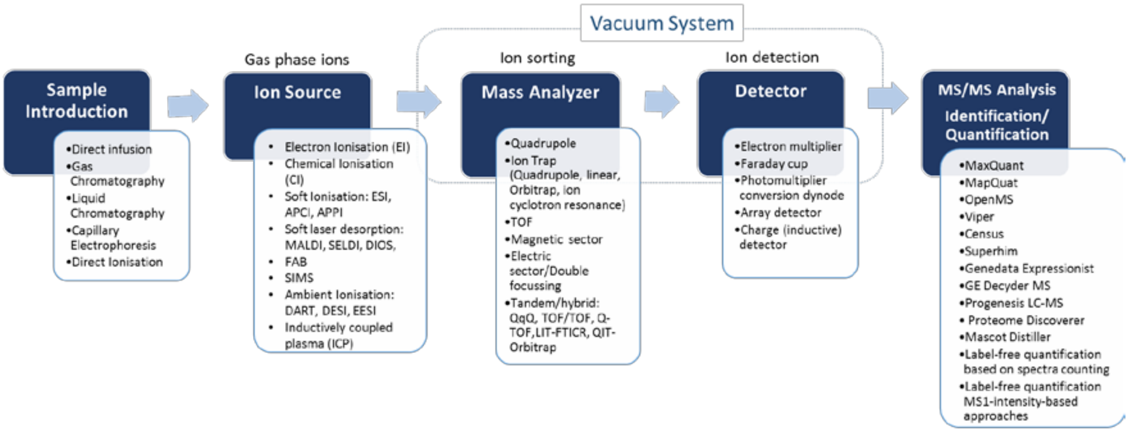

Vacuum System
Gas phase ions
Ion sorting
Ion detection
Sample Introduction
•Direct infusion
•Gas Chromatography
•Liquid Chromatography
•Capillary Electrophoresis
•Direct Ionisation
Ion Source
• Electron Ionisation (EI)
• Chemical Ionisation (CI)
• Soft Ionisation: ESI, APCI, APPI
• Soft laser desorption: MALDI, SELDI, DIOS,
• FAB
• SIMS
• Ambient Ionisation: DART, DESI, EESI
• Inductively coupled plasma (ICP)
Mass Analyzer
•Quadrupole
•Ion Trap (Quadrupole, linear, Orbitrap, ion cyclotron resonance)
•TOF
•Magnetic sector
•Electric sector/Double focussing
•Tandem/hybrid: QqQ, TOF/TOF, Q-TOF,LIT-FTICR, QIT-Orbitrap
Detector
•Electron multiplier
•Faraday cup
•Photomultiplier conversion dynode
•Array detector
•Charge (inductive) detector
MS/MS Analysis
Identification/ Quantification
•MaxQuant
•MapQuat
•OpenMS
•Viper
•Census
•Superhim
•Genedata Expressionist
•GE Decyder MS
•Progenesis LC-MS
• Proteome Discoverer
•Mascot Distiller
•Label-free quantification based on spectra counting
•Label-free quantification MS1-intensity-based approaches

***Figure 16. MS-based proteomics: General workflow and available instrumentation.***

*Description of the mass spectrometry workflow and available instrumentation for each component of the system. Abbreviations-ESI: Electrospray ionisation, APCI: Atmospheric-pressure chemical ionisation, APPI: Atmospheric pressure photoionisation, MALDI: Matrix assisted laser desorption ionisation, SELDI: Surface enhanced laser desorption ionisation, DIOS: Desorption ionisation on silicon, FAB: Fast Atom Bombardment, SIMS: Secondary Ion Mass Spectrometry, DART: Direct analysis in real time, DESI: Desorption electrospray ionisation, EESI: Extractive electrospray ionisation mass spectrometry, IC: Ion chromatography, TOF: time-of-flight mass analyser, QqQ: triple quadrupole, TOF/TOF: tandem mass spectrometer composed of two TOF analysers, Q-TOF: hybrid mass spectrometer composed of a transmission quadrupole mass spectrometer (Q) coupled to an orthogonal acceleration time-of-flight mass spectrometer (TOF), LIT-FTICR: hybrid mass analyser between linear ion trap and Fourier transform ion cyclotron resonance mass spectrometer, QUIT/Orbitrap: Quadrupole ion trap-Orbitrap. Adapted from (173).*

Once the sample is introduced into the system, the molecular ions are produced in the ion source and then they are transferred to the mass analyser through several ion optics (skimmer, focussing lens, multipoles, etc.), which basically focuses and stabilises the trajectory of the ion stream. The mass analyser sorts and separates the ions according to their charge and mass ratio (*m/z*) values and finally, detection systems are used for measuring abundance displayed as mass spectra. High vacuum ($10^{-3}$ to $10^{-6}$ Torr) is required since ions in gas phase are very reactive (175).

In tandem mass spectrometry experiments, a first analyser is used to scan and selectively transfer an ion into another reaction area where excitation and fragmentation occurs. Then a second analyser is used to record the *m/z* of products of dissociation. The specific patterns of dissociation depend upon various factors such as amino acid composition, peptide length, charge state, and excitation method. The information extracted from the MS/MS spectra is used for both identification and quantification of the peptides (176). The main steps required in a MS/MS experiment workflow are conducive to reporter ion quantification. Reporter ion based quantification (iTRAQ/TMT) requires $MS^2$ analysis, which involves fragmentation of a fixed number of precursor ions, selected from $MS^1$ scans.

### 1.4.2.1 Electrospray Ionisation (ESI)

Mass spectrometry requires the transference of analytes from the condensed state to gas phase followed by ionisation. However, proteins/peptides are thermally labile, non-volatile and polar molecules, which hamper their ionisation, and techniques circumventing structural destruction are necessary. Electrospray ionisation was first introduced in 1989 as a major advancement by J. B. Fenn, this soft ionisation method ionise intact molecules without undergoing fragmentation by multiple charging (177). An analyte undergoes three major processes during ESI, Figure 17 depicts the process of ESI.

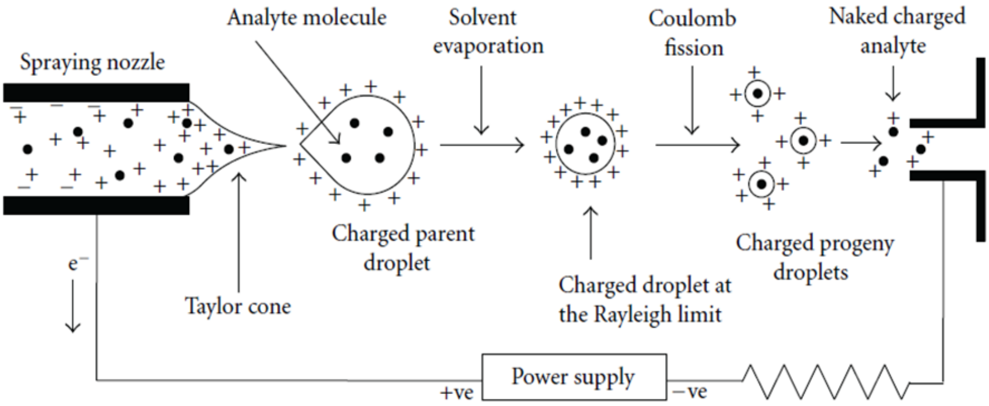


***Figure 17. Representation of electrospray ionisation process.***

*Application of high voltage to the emitter creates charged droplets containing the ionised analyte. Spraying is produced when the charges start to accumulate creating the Taylor cone, which ejects the droplets towards the heated ion transfer tube. As the droplets travel, the solvent is evaporated and the charge density increases at a critical point when the column fission occurs leading to naked charged analytes. Taken from (175)*

- *Production of charged droplets*

An electron flow is caused when the analyte is pumped through the high voltage capillary (emitter). Redox reactions produce positive or negative ions depending upon the polarity of the emitter electrode. Charges start to accumulate and are repelled by the high-voltage capillary destabilising them creating the Taylor cone, an area of high turbulence where a high field ejects a fine spray of liquid from its apex towards a counter electrode (175).

- *Coulomb explosion and droplet disintegration*

There are two competing forces acting in the formed charged droplets: surface tension and the Coulomb repulsion forces between like charges on the surface of the droplet. As the solvent evaporates when the droplets travel from the spraying nozzle to the heated capillary, its size is reduced and the repulsive forces increase causing a Coulomb explosion or fission, disintegrating the droplets. This process occurs repeatedly generating nanodroplets from which the gas-phase ions are formed (175).

- *Gas-phase ion formation*

The gas-phase analyte ions formed from precursor droplets have been explained by two hypothesis: (a) Charge residual mode (CRM) and (b) Ion evaporation model (IEM). Evidence suggest that molecular ions of large macromolecules like proteins are mainly formed following CRM (178). Basically, CRM proposes that analyte molecules retain the charges (protons in positive mode) after desolvation of the charged droplets (175).

#### 1.4.2.2 Fragmentation Methods

Typically shotgun experiments are conducted in data-dependent mode (DDA) in which a preset number of the most intense peptide ions (typically between 5-12 precursor ions) from a full scan MS are selected for fragmentation and MS/MS analysis (179). Analysis of complex mixtures of peptides has driven new mass spectrometric technologies to improving measurement accuracy of both precursor ions and fragment ions using different fragmentation methods. Currently, the most popular instruments for proteomics are the linear ion trap (LIT), Orbitrap, Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR), quadrupole (Q) and time of flight (ToF). Additionally, hybrid instruments such as LIT-Orbitrap have demonstrated to be powerful tools for shotgun experiments (31, 179). This last configuration is used in this work therefore will be discussed in further detail in the section 1.4.4.3.

Excitation of the precursor ion is usually achieved by energetic collisions with inert gasses such as nitrogen or helium. The most robust and common fragmentation method is collision-induced dissociation (CID). CID implicates acceleration of the ions to promote energetic collisions with a target gas (He) causing conversion of kinetic energy to internal energy resulting in fragmentation. For trapping instruments such as LIT or LTQ, resonant excitation creates waveforms specific to a given *m/z* value to accelerate a precursor ion, consequently only one precursor ion is activated (31).

Under CID the peptides are mainly fragmented along the peptide backbond, often transferring one or two hydrogens to yield a stable ion structure. Informative sequence ions are generated involving *b* and *y* ions, which are produced by the amide bond cleavage (Figure 18). Amino acid sequence can be determined from the mass difference of successive fragment ions of the same type. *b* and *y* ions are particularly informative since they usually retain the side chain of each residue (180). Depending upon energy of collision, fragmentation can occur at any bond along the peptide structure, generating different series of ions. Fragment ion series *a*, *b* and *c* are produced when the charged is carried at the N-terminus, while *x*, *y*, and *z* when the charge is retained at the C-terminus.

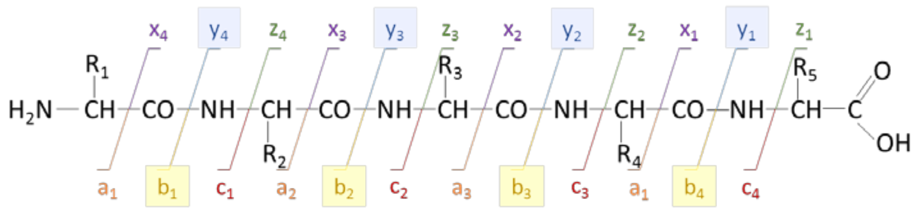


*Figure 18. Schematic representation of CID fragmentation*

*Adapted from (176).*

Ion trap CID cannot trap low mass fragment ions (up to 1/3 of the *m/z* of the precursor ion), restricting the analysis of product ions with *m/z* values below 25%-30% of the precursor ion. This implies that information about the termini might be lost and on the other hand, quantification using iTRAQ/TMT is highly limited since the reporter ions exhibit low *m/z* values. Hybrid instruments such as LTQ-Orbitrap introduced a complementary fragmentation method: higher energy collision-induced dissociation (HCD). HCD fragments the ions in a collision cell rather than an ion trap, once fragmented product ions are transferred back through the C-trap for analysis at high resolution at the Orbitrap. HCD utilises higher energy dissociation resulting in more diverse fragmentation pathways. This approach allows investigators to overcome the 1/3 rule and acquire high quality MS/MS spectra due to high resolution detection and increased ion fragmentation yield (181).

### 1.4.2.3 LTQ-Orbitrap Technology

Since its first introduction in 2005, the Orbitrap-MS technology has been continuously improved and hybrid instruments are widely used in shotgun proteomics. In this section, some of the main characteristics of the Orbitrap Elite (Thermo Scientific) will be briefly described considering that this instrument is used in this work. Figure 19 depicts a scheme of the instrument configuration. In addition to the linear ion trap for ion trapping, selection, CID fragmentation and low resolution detection and a HCD collision cell, this instrument combines the sensitivity and speed of the LTQ with the high resolution and mass accuracy of the Orbitrap mass analyser.

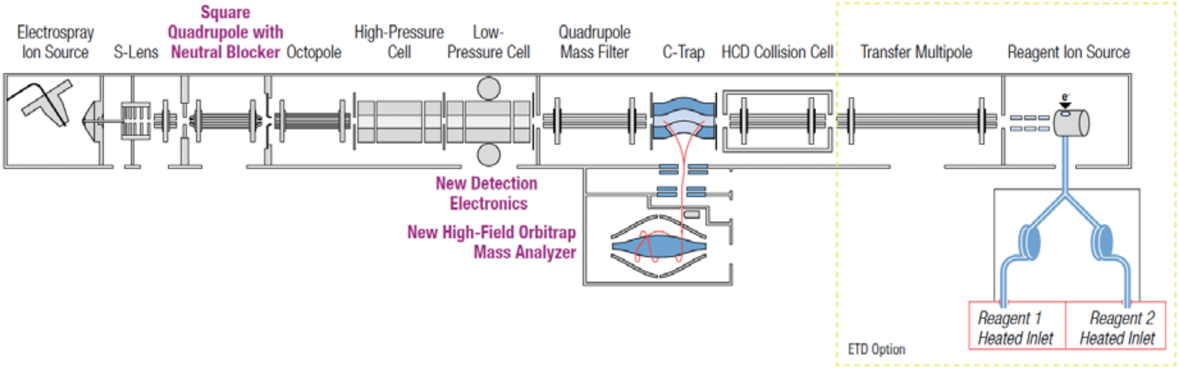


***Figure 19. Schematic representation of Orbitrap Elite.***

*The Orbitrap Elite is a hybrid instrument combines a linear ion trap and a high-field Orbitrap mass analyser. Additionally, it includes a HCD collision cell. Taken from (182)*

The Orbitrap analyser consists of a central spindle and an outer barrel-like electrode and employs electrostatic fields for trapping and analysing the ions. Thus, ions injected to the Orbitrap from the C-trap are electrostatically trapped while oscillating and rotating along the spindle. Image current signals are induced on the outer electrodes by oscillating ions. Fourier transformation is conducted in order to convert the current signals into frequencies specific for a given *m/z* value and ultimately generate a mass spectrum (179). Additionally, the Orbitrap Elite comprises a compact high field and an enhanced algorithm for signal deconvolution, features that improve both resolution (15000 to 240000 at *m/z* 400) and scan speed. This instrument has been successfully used for shotgun and top-down proteomics (183).

### 1.4.3 Quantitative Proteomics

Accurate measurement of small changes in protein and peptide abundance in response to an altered status is a challenging area in proteomics. The bottom-up approach is the preferred for quantification due to its high-throughput capabilities and sensitivity. However, mass spectrometry for peptides is not inherently quantitative, as peptides exhibit a wide spectrum of physicochemical properties including: molecular mass, hydrophobicity and charge leading to a highly variable mass-spectrometric response. Consequently, the MS signal intensity and the concentration of the analytes does not exhibit a linear relationship that can be used as an accurate quantitative measurement (174, 184).

In the last 15 years a range of approaches that allow relative and absolute quantification using label-free or label-based strategies have become available (185). The main approaches will be briefly described below.

#### 1.4.3.1 Label-free Quantification

This strategy aims to quantify peptides and proteins without using any chemical modification for labelling and compromises two general quantification methods: (a) Spectral counting and, (b) Peptide peak intensity as result of the peptide response in the mass spectrometer (174). Spectral counting indirectly infers the amount of a protein based on the observation that the number of MS/MS spectra or peptide-to- spectrum matches (PSMs) correlates to the protein abundance. This protein-centric approach requires a sufficient peptide yield in terms of number and abundance, therefore quantification of small (<20kDa) and low abundant proteins tends to be more variable (174, 185). On the other hand, quantification based on peak intensities comprises integration of the peak areas extracted from the ion chromatograms of every peptide. This approach requires multiple replicates and consistent standardisation of chromatographic and mass spectrometric conditions as it is particularly sensitive to technical variability. A major benefit is that label-free approaches are inexpensive and applicable to an unlimited number of samples (185).

#### 1.4.3.2 Stable-isotope Labelling

The premise of any stable-isotope labelling approach is that labelled and natural occurring peptides will share the same physicochemical properties including chromatographic behaviour and MS signal response. Therefore, differences in peptide and protein expression among samples can be quantified by comparing the intensity of different isotope coded peaks distinguished by their MS spectra (174). Excluding isobaric mass tags, there are a range of strategies for labelling which typically involve the addition of small mass differences that can be measured in the MS1 spectra. The labelling can be incorporated very early *in vivo* through metabolic activity or during sample preparation by chemical modification of peptides (186). Two of the most common stable-isotope labelling strategies are described below.

- *Metabolic Labelling*

In the case of metabolic labelling, the protein labelling is conducted *in vivo* during cell growth and division. The most popular metabolic labelling strategy is SILAC (stable isotope labelling by amino acids in cell culture) first introduced in 2002 (187). Basically, SILAC incorporates light and heavy isotopes ($^{13}C$, $^{15}N$) of arginine and lysine *in vivo* and combines heavy and light samples prior to sample processing, reducing the technical variability and allowing the quantification of small protein expression alterations. Some SILAC refinements have been introduced to increase accuracy and extend its use to protein turnover studies (188, 189). In 2010 a new strategy called super-SILAC was developed for human tumour proteome quantification. A mixture of stable SILAC-labelled cell lines serve as internal standards when combined with samples of interest (190). Although super-SILAC extended the use of this strategy to some clinical samples, SILAC is mostly limited to cell lines, primary cells and few small animal models.

- *Isobaric-based Labelling*

Isobaric-based quantification was first introduced in 2003 and is achieved by peptide chemical derivatisation using an array of isobaric isotopologue tags. The difference in protein expression is determined by labelling each sample with a specific tag, then the samples are pooled together, co-fragmented and analysed simultaneously which represents a clear advantage in terms of minimising technical variability (191). Importantly, a given peptide present in different samples, carrying a different isobaric tag in each sample, will exhibit the same chromatographic properties and will appear as a unique composite peak at the $MS^1$ level. Fragmentation of this derivatised precursor ion during MS/MS will provide two pieces of information: (a) reporter ion peaks used for relative quantification and (b) peptides fragment ion peaks for peptide sequencing (186). Additionally to multiplexing, isobaric labelling-based quantification has advantages such as broad dynamic range profiling of both low- and high-abundant proteins with variated physicochemical properties and even increases MS/MS fragmentation efficiency (186). There are two main isobaric reagents commercially available: iTRAQ, which allows multiplexing up to 8 samples (AB Sciex), and TMT (Thermo Scientific) up to 10 samples. iTRAQ characteristics will be described in further detail since it is the approach used in this study.

- *Isobaric Tags for Relative and Absolute Quantification (iTRAQ)*

One of the most widely used methods for multiplex quantification is iTRAQ, a $MS^2$-based method which allows simultaneous relative quantification of up to eight samples within a single run (192). Labelling is based on *N*-hydroxysuccinimide (NHS) chemistry and consists of three functional motifs: a unique isotopic reporter (N-methylpiperazine), a cleavable mass balancer (carbonyl group), and an amine-reactive group (186). Figure 20 illustrates the general structure of iTRAQ reagents. Briefly, peptide mixtures from different biological samples are labelled with different isobaric tags by reacting the NHS-ester activated group with the N-terminus and ε-amine groups of lysine residues.

The reporter is released during the peptide fragmentation process in the mass spectrometer by cleaving the linker group. The quantification is achieved by directly correlating the relative intensity of the reporter ions, which represent the relative abundance of the original peptides (31, 186).

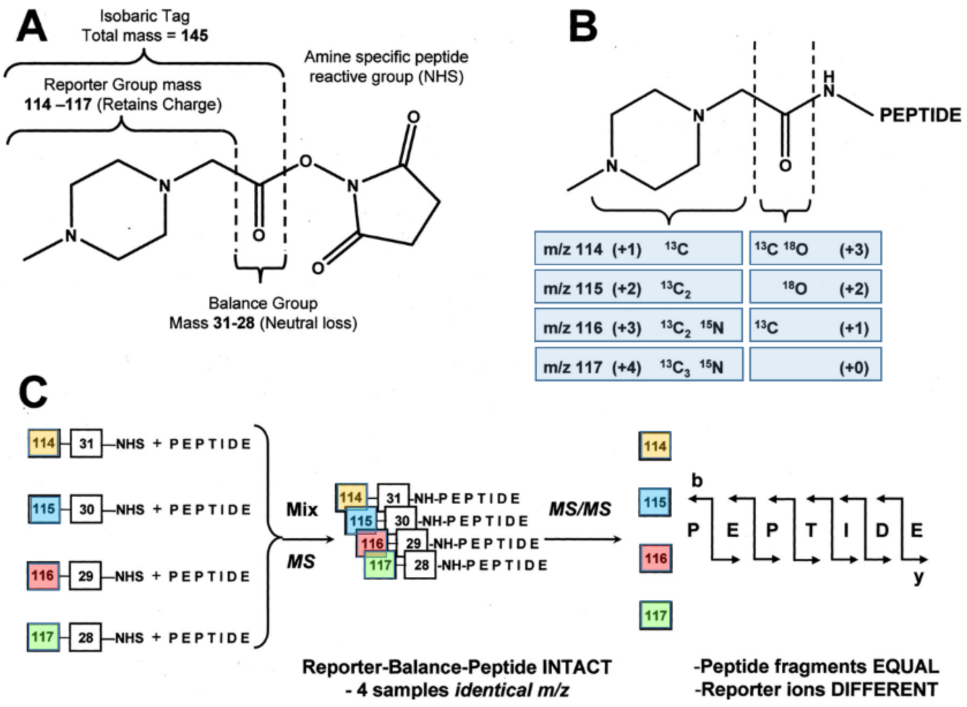


*Figure 20. General chemical structure of iTRAQ reagents.*

*A. The molecule consists of a reporter group (based on N-methylpiperazine), a mass balancer group (carbonyl), and a peptide-reactive group (NHS ester). B. Isobaric mass tags have identical overall mass but vary in terms of the distribution of heavy isotopes along their structure. C. An illustration four isobaric combinations for a mixture of four identical peptides. Each peptide labelled with a different isobaric tag appears as single, unresolved precursor in the MS1 scan. Following CID fragmentation, the four reporter group ions appear as distinct masses (114–117 Da). The relative concentration of the peptides is thus deduced from the relative intensities of the corresponding reporter ions. Adapted from (193)*

Figure 21 presents an illustration of the MS analysis of 8-plex iTRAQ reporter ions. The full MS scan registers the precursor ions detected at a given retention time, the same peptide originated from different samples labelled with a specific iTRAQ label will appear as one peak in the MS scan. Then, the most abundant ions are selected for fragmentation. The HCD spectrum shows the backbone fragmentation and the reporter ions in the low mass region. High resolution is required to achieve baseline separation of the reporter ions. The abundance of each reporter ion is used for calculating the relative abundance of a particular peptide in the different samples.

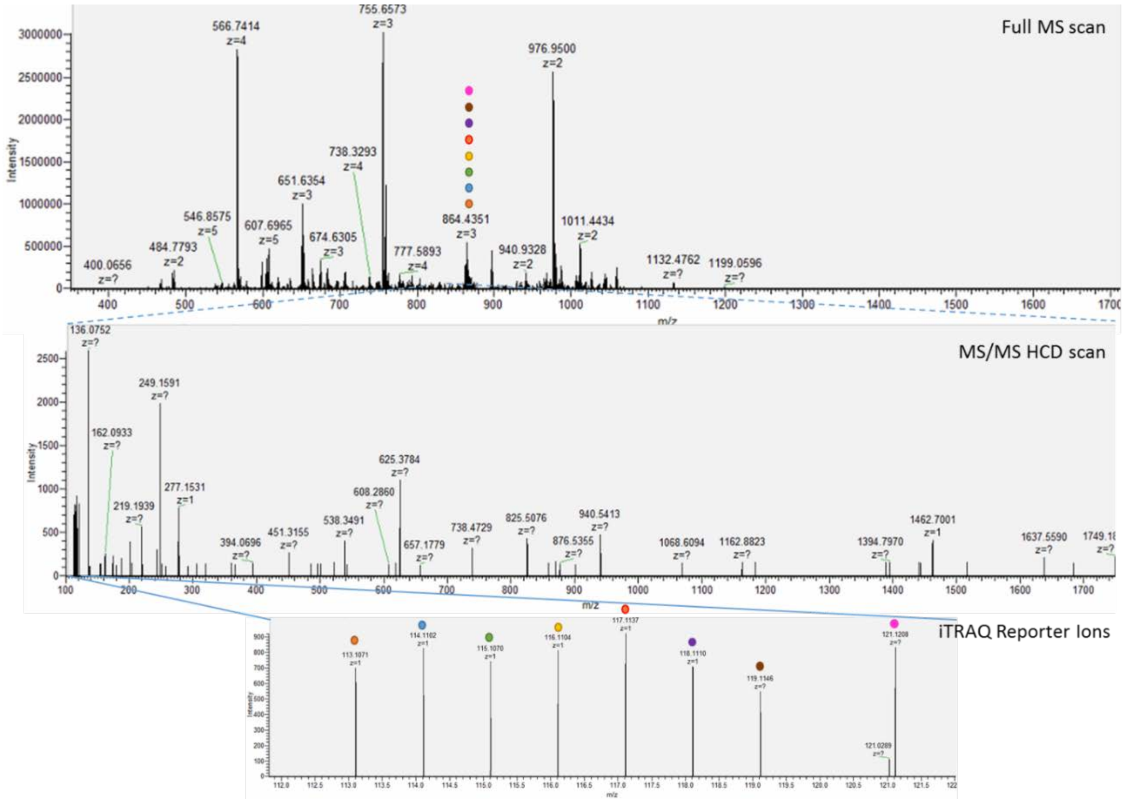


*Figure 21. Illustration of iTRAQ quantification using MS/MS.*

*The first spectrum shows a total ion chromatogram of the precursor ions at a specific retention time, then the most abundant ions are selected for fragmentation. HCD fragmentation method yields masses in the low m/z region required for reporter ion analysis. The abundance of the reporter ions is used for quantification of the same peptide originated from different samples.*

Although iTRAQ is an effective approach for relative quantification, which surpasses metabolic labelling in terms of reproducibility and precision (194), there are some analytical challenges that may affect the performance of the method. Briefly, some of these challenges will be discussed.

a. Isotopic Purity and Correction

iTRAQ and TMT reagents compromises a set of labels with variable isotopic composition: $^{13}$C and $^{15}$N atoms. However, the label reagents are not 100% isotopically pure as a result of the synthesis reactions. As consequence the observed intensity of a given peak or label will be smaller than the one expected if the composition was completely pure. Additionally, the impurities from other labels will contribute to the intensity of adjacent peaks. Figure 22 illustrates this effect. As an illustration, the label 117 will not only comprise molecules having three $^{13}$C atoms but might also contain molecules with one, two or even five $^{13}$C atoms therefore this reporter ion will report additional peaks at positions -2, -1, +1 and +2 Da from the nominal reporter ion mass. Impurities from other labels will contribute to the observed intensity of the peak at 117m/z. These shifts might confound the observed changes of protein expression and bias quantification.

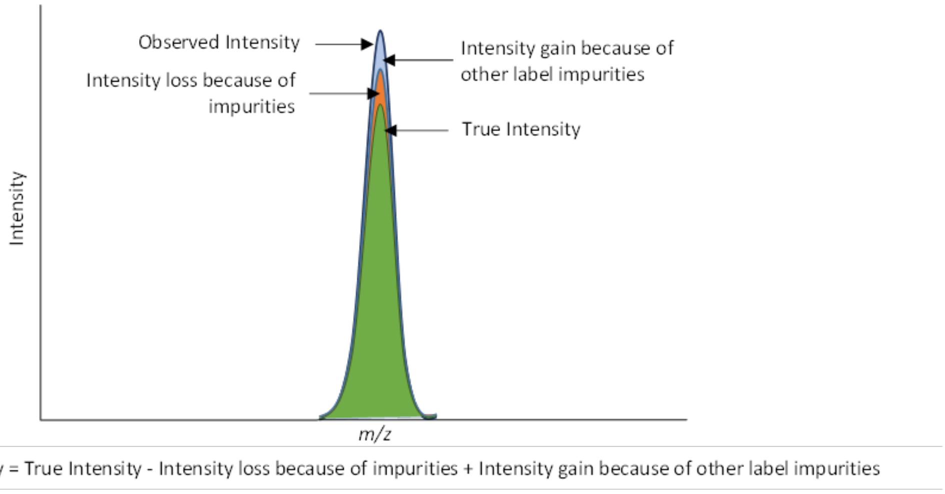


***Figure 22. Effect of isotopic purity on intensity.***

*The observed intensity of a given reporter ion comprises a composed composition and differs from the abundance of the nominal tag. The true intensity of the tag results from the intensity loss because of impurities of such tag and the intensity gain because of other label impurities. An accurate quantification requires the correction of the reporter ion intensities considering the isotopic composition.*

Manufacturers provide information about isotopic composition indicating the percentage of the isotopic variants from each reporter ion. These values can be used for correction and some algorithms have been developed to recalculate the peak areas accounting for the isotopic distribution. A system of linear equations is required for correction. Shadforth, I. *et al.* (2005) applied Cramer's rule to solve simultaneous equations and calculate true peak areas (195). However, current software tools for identification and quantification, such as Proteome Discoverer (Thermo Scientific), include an option for isotopic correction.

b. Ratio Compression and Correction

Typically to create a fragment spectrum, one precursor mass is selected, isolated and subjected to fragmentation within a preselected mass window, then the product ion masses are recorded. However, in practice not solely one mass is selected, but the precursor ions present in a specified window (1 – 2Da) around the isolation mass. Consequently, coeluting ions falling within this window will be co-isolated and fragmented together. This *co-isolation* effect can reduce the identification confidence and it is inherent to the reporter ion quantification.

On the assumption that most of proteins in biological studies do not change significantly, peptides from these proteins co-fragmented with the peptide of interest will result in a compression towards one of the ion reporter ratios expressed as fold changes (186, 195, 196). There are some experimental and computational approaches to minimise and correct this co-selection phenomenon. Since the interference from coeluting peptides is a function of sample complexity, extensive fractionation on the protein and peptide level partially alleviates the ratio compression (197). Concurrently, other approaches can be implemented: a high resolving power, $m/\Delta m$ > 15000, in the low mass region facilitates discrimination of contaminant from reporter ions minimising interference (198); narrow

MS/MS isolation window, gas purification (199) and multinotch isolation for MS$^3$ mass analysis (200, 201). Additionally in the data processing pipeline, quantitative information can be rejected having a percentage of isolation interference above a specified threshold.

### c. Estimation of Fold Changes

A protein is considered significantly modulated if the measured fold change upon perturbation exceeds a specified cut-off point. Fold changes calculated using reporter ion quantification have shown to be a function of protein abundance and mass, with low abundant and small proteins showing the largest variance (202). Threshold determination must encompass most of the technical and biological variation among replicates. Levin (2011) applied the function pwr.2p.test() in the statistical package R to calculate the statistical power for proteomic studies accounting for the total variation (biological and technical) and encompassing significance and fold changes. Therefore, a clear fold change cut-off can be determined considering the number of biological replicates and expected statistical power (203).

### 1.4.4 Data Analysis

Since each particular proteomic approach involves a specific statistical framework and bioinformatics, this section presents a general description of data analysis for reporter ion quantification. Protein identification and quantification requires independent but parallel analysis processes until the last steps of the workflow. Figure 23 illustrates a schematic of the data processing for protein quantification. The spectral data representing peptides is initially pre-processed to select peaks and remove noise, importantly quantification requires minimal filtering of data, which includes baseline correction and signal-to-noise threshold, thus profile data is recommended at this stage. On the other hand, identification benefits from processed and clean spectra and therefore centroid data is preferred (197).

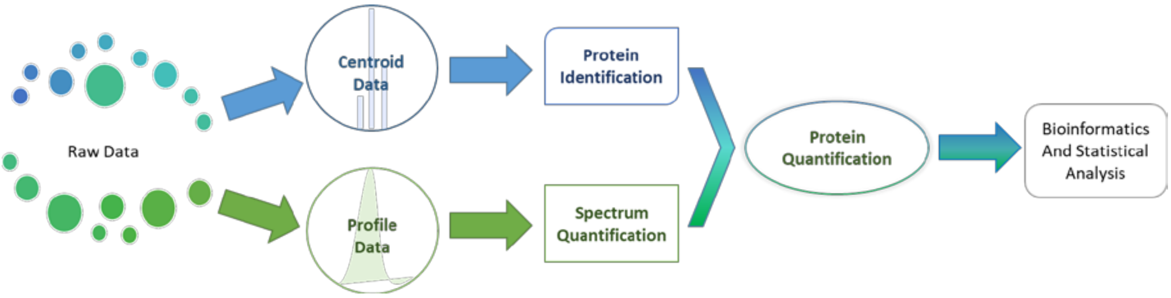


*Figure 23. Reporter ion quantification workflow.*

*Raw data is acquired in centroid and profile modes that are used for protein identification and quantification. Protein quantification is subjected to statistical assessment to validate the confidence of the protein inference and quantification. Adapted from (197)*

Protein identification implicates matching of each MS/MS spectrum with a database of simulated MS/MS spectra generated by *in silico* digestion of protein sequences from organisms with sequenced genomes. The extent of matching between theoretical and experimental mass spectra is ranked and filtered, the sequence with the best fit above a pre-set threshold is generally considered correct and will be included in a list of peptide spectrum matches or PSM and peptide identifications. Identified peptide sequences are assigned by inference to create a list of proteins. Peptide ratio measurements assigned to a given protein are then averaged to quantify the relative protein abundance (31, 204).

#### 1.4.4.1 Protein Inference

One of the main challenges of shotgun proteomics relies precisely on its peptide-centric rationale. Since peptides are used as surrogate representations of the proteins, the correct inference of proteins from peptide sequences require complex processing specially for isoforms and redundant proteins. Moreover, the correct assignation of peptides to proteins have become more complex with the improved capabilities of new proteomic platforms and continuously increasing databases. Among the main challenges for protein inference are particularly relevant: (a) stochastic sampling, (b) identification of redundant proteins (proteins which share tryptic peptides) and (c) identification of unique peptides for homologous proteins (protein isoforms) (31).

Unique peptides are required for a confident identification of both redundant and homologous proteins, in fact, one single high confidence unique PSM can provide enough evidence for the identification of a protein, particularly in conjunction with other high confidence non-unique peptides. Additionally, for reporter ion quantification the most accurate method is based exclusively on unique peptides. However, shared peptides by definition are more abundant than the unique peptides and consequentially these latter are more problematic to detect and identify (31). Multiple strategies has been proposed and used to address this problem, Li, Y. F. *et al.* (2012) have categorised these various computational approaches into three groups (205):

a. *Rule based strategies*: protein assignation is conducted relying only on the most confident identified unique peptides.
b. *Combinatorial optimisation algorithms*: these methods are based on constrained optimisation formulations of the protein inference problem to provide a list of proteins that optimise certain criteria, generating, for example, minimal protein lists comprising some or all confidently identified peptides. Usually these methods apply approximation algorithms.
c. *Probabilistic inference algorithms*: assign identification probabilities for each protein in the database and implicate usually two steps: (1) a pre-processing step which converts PSM scores into PSM probabilities. (2) Protein inference is conducted using the assumed probabilistic model, involving posterior error probabilities (PEP) calculation.

Importantly the assessment of the method performance for protein identification remains as an additional problem. However two different approaches can be applied: calculation of FDR at the protein level using decoy protein sequences (*e.g.* reversed or randomised) and the use of standard samples (mixture of known proteins) (205).

### 1.4.4.2 Statistical Significance of Protein Identification: FDR and PEP

False discovery rates (FDR) and posterior error probabilities (PEP) are two complementary statistical methods for assessing the significance of peptides assignments. Considering that a large number of peptide matches are required to be simultaneously validated for correct and incorrect assignments, multiple testing is essential. In this scenario, using even extremely small *p*-values by random chance alone can allow a significantly high number of false positives, which is detrimental when the protein inference is based on the peptide identifications. There are available various strategies to address multiple testing corrections, the most classical being the *Bonferroni* correction which restricts the *p*-value in proportion to the total number of matches being validated. This is considered as a conservative approach that not only significantly reduces the number of false positives but at the cost of the true discoveries (206).

A more moderate approach has been adopted for proteomic studies; the false discovery rate (FDR) defined as the '*expected proportion of incorrect assignments among the accepted assignments at the global level*' (206). If a FDR threshold of 1% is set it means that a list of PMS comprising 99% correct and 1% incorrect matches will be accepted. The most widely used strategy to calculate FDRs is the target-decoy dataset search (TDS), which implies to search a set of spectra against a decoy database of peptide sequences (shuffled, reversed or Markov-chain generated) and assign a particular score: a *q*-value, to every PSM. The underlying assumption here is that false positives should occur with similar likelihood in both the decoy and target databases. This idea can be applied to calculate local or global FDRs and validate assignations at peptide and protein level (207, 208). Some further improvements to FDR calculation are necessary when the large MS experiments comprising hundreds to thousands LC/MS runs the basic assumption of TDS is compromised. Recently, Savitski, M. *et al.* (2015) have proposed an alternative FDR approach so-called picked target-decoy strategy (picked TDS) to prevent decoy protein overrepresentation. Picked TDS treats target and decoy sequences as a pair rather than as individual factors. The sequence either target or decoy with the highest score is selected (208).

A complementary strategy is the posterior error probability (PEP) calculation, which indicates the probability that an observed PSM is incorrect. If the PEP associated with a given sequence is 5%, this means that there is a 95% chance that the peptide was in the mass spectrometer when the spectrum was generated (207). The relationship between FDR and PEP is presented in the Figure 24. In terms of distribution areas, FDR is defined as the ratio of the number of incorrect PSMs with score >x (B) to the total number of PSMs with score >x (A + B). On the other hand, the PEP is the defined in terms of the heights of distribution: the number of incorrect PSMs with score = x (b) to the total number of PSMs with score = x (a + b) (207).

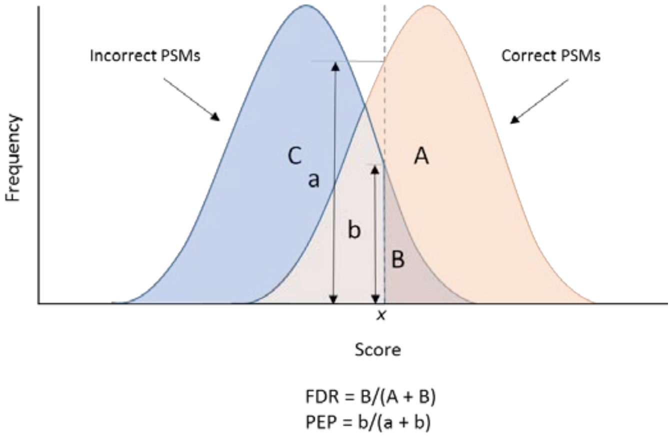


***Figure 24. FDR and PEP for statistical assessment of PSMs.***

*These complementary methods are used for assessment of PSM assignments. FDR is the expected proportion of incorrect assignments among the accepted assignments at the global level. PEP indicates the probability that a PSM is incorrect. Adapted from (207)*

Current software tools for proteomics data such as Proteome Discoverer (Thermo Scientific) allows calculation at peptide and protein levels, and provides PEPs and additional assessment tools for assignment confidence like XCorr and ΔCN, which are measurements of the correlation between the experimental and theoretical spectra. This particular software was used for the initial processing of the raw data in conjunction with statistical pipelines designed in this work to assess the changes in relative abundance of proteins and thus, determining novel candidates for tuberculosis diagnosis.

# HYPOTHESIS

Unbiased in-depth quantitative proteomic analysis of plasma from healthy controls and active pulmonary tuberculosis patients will identify novel biomarkers for active TB. Validation of these candidate biomarkers will lead to a multi-marker panel that can distinguish TB patients from relevant controls.

In order to test this hypothesis I have proposed four main aims:

1. Perform an unbiased in-depth plasma proteomic analysis of healthy controls and active pulmonary tuberculosis in two groups of different ethnic origin using an optimised quantitative MudPIT (Multidimensional Protein Identification Technique) approach.

2. Identify a common set of biomarkers suitable for validation for active pulmonary tuberculosis through appropriate bioinformatic and statistical analysis, utilising peptide intensities to protein expression levels.

3. Validate the multi-marker panel for active tuberculosis using appropriate plasma samples groups.

4. Determine multi-marker panel's performance considering possible confounding variables in order to evaluate its applicability for clinical use.

# CHAPTER 2

# Materials and Methods

## 2.1 General Experimental Design

The general workflow for this study is presented in the Figure 25. Briefly, novel biomarkers for active tuberculosis diagnosis were explored using plasma samples from healthy controls and tuberculosis patients. Since the ultimate goal of this work is to identify a more universal biosignature that can distinguish active tuberculosis patients in diverse genetic backgrounds, the proteomic profile was conducted using samples from two different ethnic ancestries through an optimised MudPIT approach and iTRAQ quantification. Raw data generated by mass spectrometry were assessed by quality control and then normalised. Subsequently, the protein expression levels were determined and statistical differences between groups established. A diverse set of analyses was performed such as those based on principal component analysis, biological network and pathway characterisation, and gene ontology assessment, approaches to identify the top most significantly modulated proteins as markers for tuberculosis. Structure and characteristics of clinical cohorts for validation were established. A subset of those candidates were evaluated in various validation groups using Luminex and ELISA and performance of multi-markers panel was assessed.

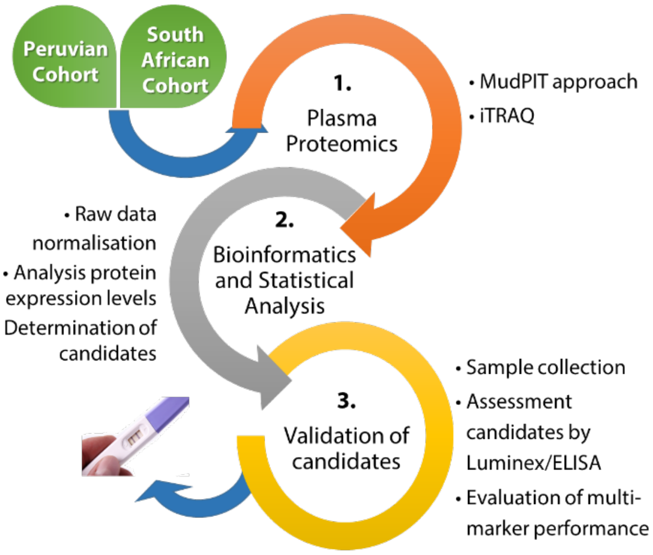

Peruvian Cohort
South African Cohort
1. Plasma Proteomics
• MudPIT approach
• iTRAQ
• Raw data normalisation
• Analysis protein expression levels
• Determination of candidates
2. Bioinformatics and Statistical Analysis
3. Validation of candidates
• Sample collection
• Assessment candidates by Luminex/ELISA
• Evaluation of multi-marker performance

*Figure 25. General workflow for TB diagnosis biomarkers discovery.*

*Proteomics of plasma samples from two different ethnic ancestries: Black-African ancestry and Amerindian ancestry was performed using MudPIT technology and iTRAQ relative quantification. Bioinformatics and statistical assessment allow determination of a multi-marker panel for active pulmonary tuberculosis.*

*Subsequently, validation and performance assessment of this panel was performed using suitable groups of plasma and serum samples.*

This chapter presents general methods used in this work, details for specific experiments such us optimisation (Chapter 3) or plasma proteomic profiling (Chapter 4 and 5) are described in each chapter methods section.

### 2.1.1 Sample Size

In order to minimise the effect of host variability on clinical features of tuberculosis infection, only samples from male individuals were selected and two groups of different ethnic origin were considered: Black–African ancestry (Cape Town) and Amerindian ancestry (Peru). A reported simulation designed for proteomic studies assuming 40% variation and using the function for R pwr.2p.test() was considered to determine the number of plasma samples required for this study (203). Figure 26A shows a simulation which indicates that a study including 7 individual biological replicates per group with a total variation of 40%, the minimum size effect required to achieve statistical power over 0.9 would be approximately two-fold change. This calculation is consistent with Cohen, G., *et al.* (2013) power calculation for a discovery iTRAQ set up comparing two conditions.

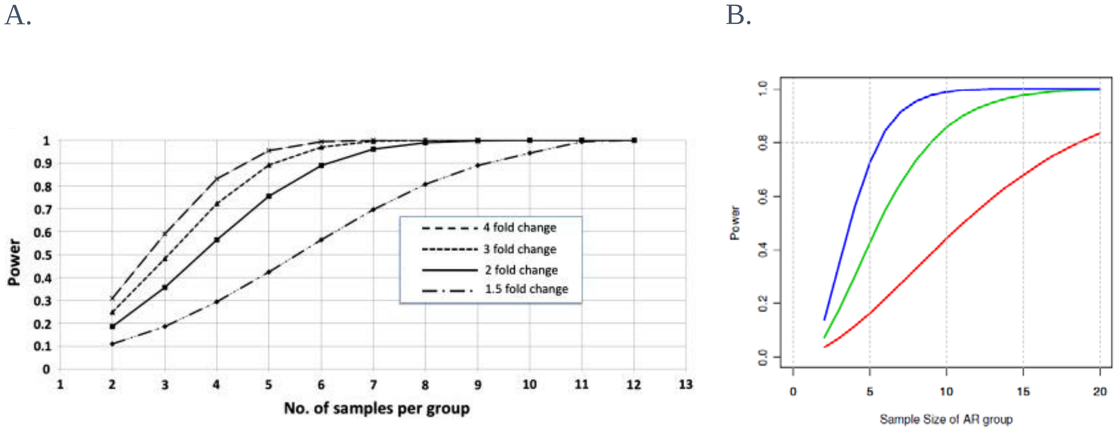


***Figure 26. Power analysis simulation for proteomic analysis***

*A. Calculation of number of samples vs. power for different expression differences using the function pwr.2p.test() in the statistical package R. Figure 2B taken from Levin,Y., et al. (185). B. Power calculation based on iTRAQ ratio quantification estimated considering a standard deviation of 0.25 on logged iTRAQ ratios. The red, green and blue lines indicates logged-2 effect sizes of 1.0, 1.5 and 2.0, respectively. Supplementary figure 1A taken from Cohen, G., et al. (209).*

In total, six 8-plex independent iTRAQ experiments were performed on SEC prefractionated samples, each set included seven samples and one master pool to control for batch effect. This work included two experimental approaches for discovery as illustrated in the Figure 27. The general structure of the work presented in this thesis compromises the optimisation of the proteomic multidimensional method for plasma profile which is described in chapter 3. A complete plasma proteomic profile based on one single set of 8 samples and four iTRAQ experiments (one per SEC segment) presented in chapter 4. And finally, the detailed analysis of the most informative segment which involved two additional 8-plex experiments.

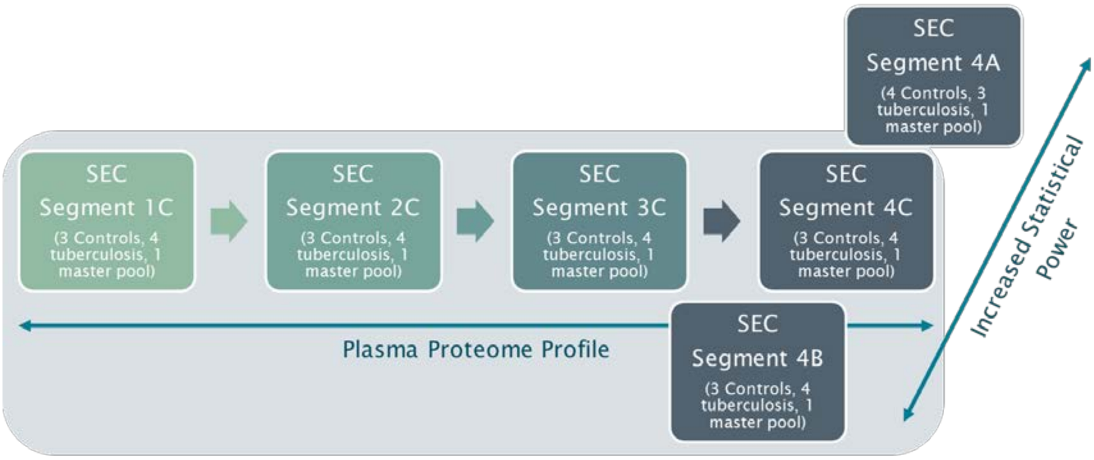

SEC
Segment 4A
(4 Controls, 3 tuberculosis, 1 master pool)
SEC
Segment 1C
(3 Controls, 4 tuberculosis, 1 master pool)
SEC
Segment 2C
(3 Controls, 4 tuberculosis, 1 master pool)
SEC
Segment 3C
(3 Controls, 4 tuberculosis, 1 master pool)
SEC
Segment 4C
(3 Controls, 4 tuberculosis, 1 master pool)
Increased Statistical Power
Plasma Proteome Profile
SEC
Segment 4B
(3 Controls, 4 tuberculosis, 1 master pool)

***Figure 27*. Experimental strategy for discovery**

*The discovery stage of this work involved six 8-plex iTRAQ sets. A first group of samples, four tuberculosis and three control were used to profile the plasma proteome across all 4 segments; the remaining tag was used to label the master pool. Sample size was increased for the most informative segment, which involved two more iTRAQ experiments analysing segment 4 reducing the false positive rate for discovery.*

The whole plasma proteome described in chapter 4 was profiled using the one set of samples, four tuberculosis and three control samples. Secondly, the sample size for the most informative segment was increased to a total of 11 tuberculosis and 10 control samples (chapter 5). In total, 29 samples from South African and Peruvian male controls and tuberculosis patients were available for this study, the master pool was prepared pooling together 20µL of each one of all this samples. From this cohort 21 samples were selected for proteomic profiling. The clinical data and general allocation in iTRAQ sets is described in Table 4.

In general, three or four plasma samples per group from two different ethnic ancestries were included in each iTRAQ set. iTRAQ label allocation was block-randomised, excepting the master pool which was consistently labelled with the tag 113, thus reducing bias resulting from isotopic tag derivatisation. Considering the power calculation previously mentioned, at least ten samples per group were utilised for discovery, which, allows a statistical power near to 1.0 when the size effect cut-off is at least 2.0-log2 iTRAQ ratio. Table 4 describes the general iTRAQ experiment structure and sample allocation used for profiling the plasma proteome in active tuberculosis. Age, BMI and smoking status of the individuals included in this work is included. Additionally, the chapter where the data is presented and discussed is indicated.

*Table 4. General description of iTRAQ experiments for comprehensive plasma profile*

| Experiment | Profiled SEC segment | Chapter | Group | Sample ID | iTRAQ Tag | Age | BMI | Smoking |
|---|---|---|---|---|---|---|---|---|
| Preliminary Exploration | 4 | 3 | Active tuberculosis | TB17 | 113 | 38 | 23.98 | No |
| | | | | TB35 | 114 | 25 | 17.32 | Ex |
| | | | | TB39 | 115 | 24 | 19.86 | Smoker |
| | | | | TB55 | 116 | 25 | 18.86 | No |
| | | | Control | HC14 | 117 | 28 | 21.02 | Smoker |
| | | | | HC24 | 118 | 36 | 22.7 | Smoker |
| | | | | HC29 | 119 | 34 | 29.37 | No |
| | | | | HC58 | 121 | 21 | 19.38 | Smoker |
| Set A | 1, 2 ,3, 4 | 4 (Profile from segment 4 used in chapter 5 as well) | Active tuberculosis | TB03A | 114 | 42 | 19 | Ex |
| | | | | TB04A | 121 | 30 | 22.27 | No |
| | | | | TB10P | 118 | 27 | 21.9 | Current |
| | | | | TB01P | 119 | 30 | 21.1 | Ex |
| | | | Control | HC03A | 115 | 22 | 22.21 | Ex |
| | | | | HC04A | 116 | 32 | 22.31 | Current |
| | | | | HC04P | 117 | 34 | 25 | No |
| | | | | MP01 | 113 | N/A | N/A | N/A |
| Set B | 4 | 5 | Active tuberculosis | TB01A | 117 | 28 | 17.87 | Ex |
| | | | | TB06A | 119 | 35 | 17.63 | Current |
| | | | | TB03P | 115 | 44 | 25 | No |
| | | | Control | HC02A | 114 | 27 | 25.42 | No |
| | | | | HC07A | 116 | 30 | 21.86 | Current |
| | | | | HC02P | 118 | 26 | 25.1 | No |
| | | | | HC03P | 121 | 26 | 24.1 | No |
| | | | | MP05 | 113 | N/A | N/A | N/A |
| Set C | 4 | 5 | Active tuberculosis | TB05A | 116 | 25 | 20.6 | Current |
| | | | | TB02A | 118 | 42 | 20.96 | Ex |
| | | | | TB02P | 114 | 25 | 18.8 | Ex |
| | | | | TB05P | 121 | 21 | 22.3 | Ex |
| | | | Control | HC05A | 119 | 35 | 22.94 | Current |
| | | | | HC01P | 115 | 24 | 23.2 | No |
| | | | | HC07P | 117 | 27 | 25.3 | No |
| | | | | MP04 | 113 | N/A | N/A | N/A |

### 2.1.2 Sample Collection

Ethical approval for sample collection with informed consent was provided for both South African and Peruvian samples. University of Southampton ERGO approval for transporting samples to the United Kingdom was granted (Approval 17758). The criteria for inclusion and exclusion are summarised in Table 5. Samples were collected and stored at -80°C.

Blood collection and plasma processing were performed according to the recommendations of Standard Operating Procedure Integration Working Group (SOPIWG) (210). To maintain consistency in the proteome, all blood samples were collected in tubes with sodium citrate. Plasma were collected in the Ubuntu HIV/TB clinic in Cape Town, South Africa and community health clinics in Lima, Peru.

*Table 5. Criteria for inclusion and exclusion of participating individuals*

| *Inclusion Criteria* | *Control Group* | *TB Group* |
|---|---|---|
| 1. Male | ✓ | ✓ |
| 2. Age range from 20 to 35 | ✓ | ✓ |
| 3. Body Mass Index from 18.5 to 24.9 | ✓ | ✓ |
| 4. Non-smoker | ✓ | ✓ |
| 5. No drug treatment before enrolment | ✓ | ✓ |
| 6. Same ethnicity | ✓ | ✓ |
| 7. HIV negative | ✓ | ✓ |
| 8. Available standard clinical information | ✓ | ✓ |
| 9. Quantiferon negative (Only Peru) | ✓ | X |
| 10. Smear positivity | X | ✓ |
| 11. Sputum culture positivity | X | ✓ |
| 12. Chest X-ray abnormalities | X | ✓ |
| 13. Extensive pulmonary infiltrates | X | ✓ |
| *Exclusion Criteria* | *Control Group* | *TB Group* |
| 1. Female | ✓ | ✓ |
| 2. HIV positive | ✓ | ✓ |
| 3. Diabetes mellitus | ✓ | ✓ |
| 4. Haemoglobin <8 g/dl | ✓ | ✓ |
| 5. Renal impairment with Creatinine >150µm/l | ✓ | ✓ |
| 6. Abnormal liver function with ALT >80i.u./l | ✓ | ✓ |
| 7. Use of any investigational or non-registered drug, vaccine or medical within 182 days preceding of study, or planned use during the study | ✓ | ✓ |
| 8. Enrolment in any other clinical trial | ✓ | ✓ |
| 9. Evidence of severe depression, schizophrenia or mania | ✓ | ✓ |
| 10. Unable to provide informed consent | ✓ | ✓ |
| 11. Principal investigator assessment of lack of willingness to participate | ✓ | ✓ |

## 2.2 Patient Recruitment and Ethics Statement

Participants from South Africa belong to black-African ethnicity were recruited at Ubuntu HIV/TB clinic in Cape Town, located in the southwest coast of South Africa. Written informed consent was obtained, HIV testing was offered, and chest radiographs were performed as per routine practice. The study was approved by the University of Cape Town Research Ethics Committee (HREC, REF 516/2011). The diagnosis of active tuberculosis was based on sputum smear positivity and x-ray chest examination. For the control group, all sputum samples were smear and culture negative for acid-fast bacillus (AFB). Samples from South Africa were retrospectively selected from a cohort collected and described by Walker, N. *et al.,* (2017) (211).

On the other hand, individuals from the Peruvian cohort were prospectively recruited at clinics in Lima, Peru to match individuals from South Africa. Written consent was obtained. The study was approved by the Peruvian University Cayetano Heredia Research Ethics Committee (Constancia 419-21-15). The diagnosis of active tuberculosis was made on tuberculosis symptoms questionnaire, sputum smear positivity, culture positivity using microscopic-observation drug-susceptibility (MODS) culture and chest X-ray. Healthy control individuals were Quantiferon negative.

Control individuals for both cohorts were recruited in the clinics when attending to the clinics with a friend/family member who was seeking healthcare due to respiratory symptomatology. Only blood samples from male individuals who were HIV negative were included in this study. Clinical data information is provided in relevant result chapters.

## 2.3 Sample Collection and Processing

Blood was collected on sodium citrate-treated tubes after consent and centrifuged in order to remove blood cells. Plasma samples were then frozen at −80°C in aliquots of 100µl to minimize freeze-thaw cycles prior to analysis. As soon as the samples were received at the Southampton General Hospital, aliquots of 121.2µl of plasma were liquid –fixed with 383.8µl of 7M guanidine hydrochloride and 10% ethanol and stored at -20°C until size exclusion chromatography fractionation.

## 2.4 Reagents and Chemicals

The chemical reagents acetonitrile, ethanol, isopropanol, methanol, acetone, triethylammonium bicarbonate, and formic acid (HPLC grade) were obtained from Sigma Corporation (Poole, Dorset, UK.). Guanidine hydrochloride, acetonitrile and methanol (MS grade) was obtained from Thermo Fischer Scientific, UK. The ultrapure HPLC grade water, utilised for the initial peptide fractionation with high pH RP and subsequent LC-MS analysis procedures, was generated from the Barnstead water filtration system (Dubuque, IW, USA). All iTRAQ reagents and buffers were obtained from Applied Biosystems (Warrington, Cheshire, UK.).

## 2.5 Plasma proteomics analysis

### 2.5.1 Multidimensional Protein Identification Technology (MudPIT) Analysis

Optimisation and use of the bottom-up MudPIT approach for separation of plasma proteins in this work was based on the method reported elsewhere (35, 92). The general workflow for the plasma proteomics analysis is illustrated in the Figure 28 and consists of five steps: (1) High-Performance Size Exclusion Chromatography (HP-SEC), (2) dialysis purification, (3) trypsinisation, (4) Stable isotope labelling, (5) peptide prefractionation using alkaline reverse-phase high performance liquid chromatography (RP-HPLC) and (6) on-line acidic RP-HPLC coupled to tandem mass spectrometry, HPLC-MS/MS. The method presented in Figure 28 describes the method fully optimised and used to profile the differential plasma proteome resulting from active tuberculosis infection. Chapter 3 describes optimisation steps to develop this particular method. Chapter 4 presents the application of this method to profile the plasma proteome from the four SEC segments in a set of 8 samples and finally, Chapter 5 describes a detailed analysis of one of the SEC segments 21 samples using the method depicted in Figure 28.

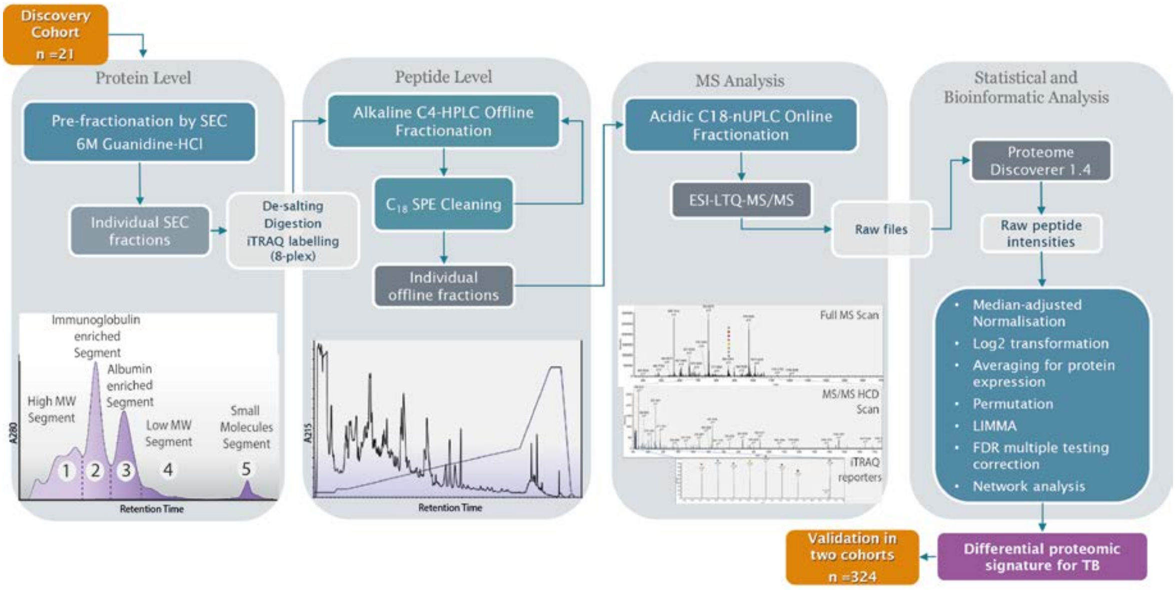


*Figure 28. Multidimensional protein identification technology (MudPIT) analysis*

*Identification and quantification of plasma proteins was performed using MudPIT approach which comprises a series of fractionation steps at both protein and peptide level. Highlighted in blue are specific steps that were developed and optimised in this work.*

### 2.5.2 High Performance Size Exclusion Chromatography

In general, SEC separations were performed on the Shimadzu HPLC system equipped with an inline membrane degasser (Model DGU-20A5), dual piston high pressure pump (Model LC-20AD), thermostatic column compartment (CTO-20A), and multi-wavelength UV detector (Model SPD-20A). Each sample was independently pre-fractionated. For each injection, 550µL of previously fixed plasma in 7M Guanidine hydrochloride was thawed on ice. The separations were performed with five serially connected Shodex KW-804 (5.0µm size particle, 8.0mm × 300 mm) and KW-302.5

(7.0µm size particle, 8.0mm × 300 mm) columns (Showa Denko America, Inc., NY, U. S. A.) under isocratic elution. Five retention time defined segments (assigned as Segments 1−5) were generated as shown in Figure 28 detected at 280nm.

### 2.5.3 Dialysis Purification

The protein segments were dialysis-purified using 3KDa MWCO Slide-A-Lyzer cassettes according to manufacturer's specifications (Thermo Fisher, Hemel Hempstead, Hertfordshire, UK). Four volumes of 4L of ultrapure water were renewed every 12h intervals in a cold room environment (4°C). The resulting dialysates were transferred into 15mL tubes and completely lyophilised using the Edwards Modulyo EF4-174 freeze dryer and Thermo Savant Micro Modulyo-115 benchtop freeze dryer. Protein extracts were stored at -80°C under argon atmosphere.

### 2.5.4 Protein Quantification

Total protein lyophilised extracts obtained from each SEC segment were reconstituted with 0.5M TEAB and 0.05% SDS and sonicated on ice. Protein extracts were then centrifuged for 10 minutes at 16000xg and 4°C. The supernatants were transferred to fresh tubes and the pellets were kept at -20°C. The protein content in the supernatants were quantified using the Nanodrop ND-1000 spectrophotometer (Thermo Scientific, Wilmington, USA) using the $A_{280}$ program. 120µg of protein from each sample were aliquoted in new tubes and the volumes were adjusted with ultrapure water to the same volume, setting the most diluted sample as maximum volume. All the samples were adjusted to a final concentration of ~4µg/µL of protein.

### 2.5.5 Trypsin Digestion

120µg of protein volume-adjusted were reduced with 2µL of TCEP (50mM tris-2-carboxymethyl phosphine) and incubated for 1h at 60°C. Reduced samples were then alkylated using 1µL of MMTS (200mM methylmethane thiosulphonate) and incubated 10 minutes at room temperature. 6µL of 500ng/µL Trypsin MS grade (Pierce, Thermo Fisher Scientific, UK) were added to each sample and incubated overnight for 16h at 37°C in dark.

### 2.5.6 iTRAQ labelling

iTRAQ 8-plex tags were equilibrated at room temperature and isopropanol was added accordingly to ensure >60% organic phase during labelling. Each tag was added to the appropriate trypsinised sample according to the particular experimental design, then the labelling reaction was conducted for 2h at room temperature. The reaction was stopped with 8µL of 5% ammonium hydroxylamine. Samples were dried and stored at -20°C until chromatographic separation.

### 2.5.7 Peptide Prefractionation with Offline alkaline RP-HPLC

Offline peptide fractionation was based on high pH Reverse phase (RP) chromatography using the Kromasil, C4 column (3.5µm, 2.1mm x 150mm) and on the Shimadzu HPLC system described in section 2.5.2. iTRAQ labelled tryptic peptides were analytically reconstituted and pooled together with 100µL of mobile phase, centrifuged at 16000xg at room temperature for 10 minutes, and the pellet was kept at -20°C. Supernatant was injected and separated at a flow rate 0.30mL/min and 30°C. The fractions were collected in a peak-dependent fashion detected at 215nm. The peptide fractions were dried at room temperature with speedvac concentrator for 4−5h and stored at −20°C until the LC−MS analysis.

### 2.5.8 LC-FT-Orbitrap MS Analysis

The LC−MS experiments were performed on the Dionex Ultimate 3000 UHPLC system coupled to the high resolution nano-ESI-LTQ-Velos Pro Orbitrap-Elite mass spectrometer (Thermo Scientific). Individual peptide fractions were reconstituted with 31μL of loading solution (2% acetonitrile, 1% formic acid). Injection mode using the loading pump was at 5μL/min flow rate for 5min.

Two separate analyses for HCD and CID fragmentation for each of the collected fractions were performed. For the analytical separation the AcclaimPepMap RSLC, 75μm × 25 cm, nanoViper, C18, 2μm particle column with trap cartridge retrofitted to a PicoTip emitter (FS360-20-10-D-20-C7) was used for multistep gradient elution. Mobile phase (A) was composed of 2% acetonitrile, 0.1% formic acid and 5% DMSO and mobile phase (B) was composed of 99.9% acetonitrile, 0.1% formic acid and 5% DMSO. The gradient elution method at flow rate 300nL/min gradually increased mobile phase B. The iTRAQ labelled peptides were fragmented in the axial electric field assisted higher energy collisional dissociation (HCD) cell. The mass spectrometer was set so that each full MS scan was followed by the ten most intense ions for MS/MS with charge +3 and +2. The normalised collision energy for MS2 was 35.0%. Full MS scans and MS/MS scans were acquired at a resolution of 30000 or 60000 for profile-mode and 15000 for centroid-mode, respectively, with a lock mass option enabled for the 445.120025m/z ion (DMSO). Data were acquired using Xcalibur software. The LTQ FT-Orbitrap system was externally mass calibrated every 3−4 days using the positive ion calibration solution (Thermo Pierce, Rockford, IL, USA). Ion tuning was verified on a weekly basis as recommended by the manufacturer. Conditions for ionisation, CID and HCD fragmentation and ion detection were reported in a previous work (212).

## 2.6 MS Data Processing

Processing of the acquired mass spectra was performed with the Proteome Discoverer 1.4 software followed the workflow illustrated in Figure 29A. SequestHT was used for the target decoy search for tryptic peptides, allowing two missed cleavages, a tolerance of 10ppm, and a minimum peptide length of 6 amino acids. A maximum of 2 variable (3 equal) modifications; oxidation (M), deamidation (N, Q) and phosphorylation (S, T, Y) were set as dynamic modifications. As static modifications were set: iTRAQ8plex (Any N-terminal), Methylthio (C) and iTRAQ8plex (K).

Fragment ion mass tolerances of 0.02Da for the FT-acquired HCD spectra and 0.5Da for the IT-acquired CID spectra. FDR was estimated with the Percolator (6.4Bit) and set to ≤0.01 and validation was based on *q* value at <0.01 for high confidence or <0.05 for moderated confidence. All spectra were searched against the reviewed UniProtKB SwissProt human proteome and the reference proteome (SwissProt and TrEMBL) for *Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv) both retrieved on 04 August 2017. All peptide spectrum matches (PSM) of reporter ions and iTRAQ ratios were exported to .txt at 1%FDR or 5%FDR peptide confidence and 50% co-isolation exclusion threshold. Protein grouping was allowed, maximum parsimony principle was applied and normalisation on protein median performed with minimum protein count set at 20. Only unique peptides were considered for quantification downstream analysis.

### 2.6.1 Statistical Pipeline for iTRAQ based quantification

A general description of the statistical approach used in this work is presented here, more specific details are included in the relevant result chapters. Firstly, box-and-whiskers plots and interquartile analysis were performed as quality control for the raw peptide intensities. As a result of this first inspection, the pipeline shown in the Figure 29B was designed to calculate the protein expression levels. Briefly, raw data produced by Proteome Discoverer were imported into R (version 3.3.1), an open source statistical analysis software, and using custom code provided by Dr. Cory White, median adjusted normalisation was performed on unique peptides. Median-normalised peptide intensities were log2-transformed and values were averaged to obtain the mean relative expression for each protein. Statistical analysis (significance testing) methods were applied to determine size effect based on quantitative iTRAQ data.

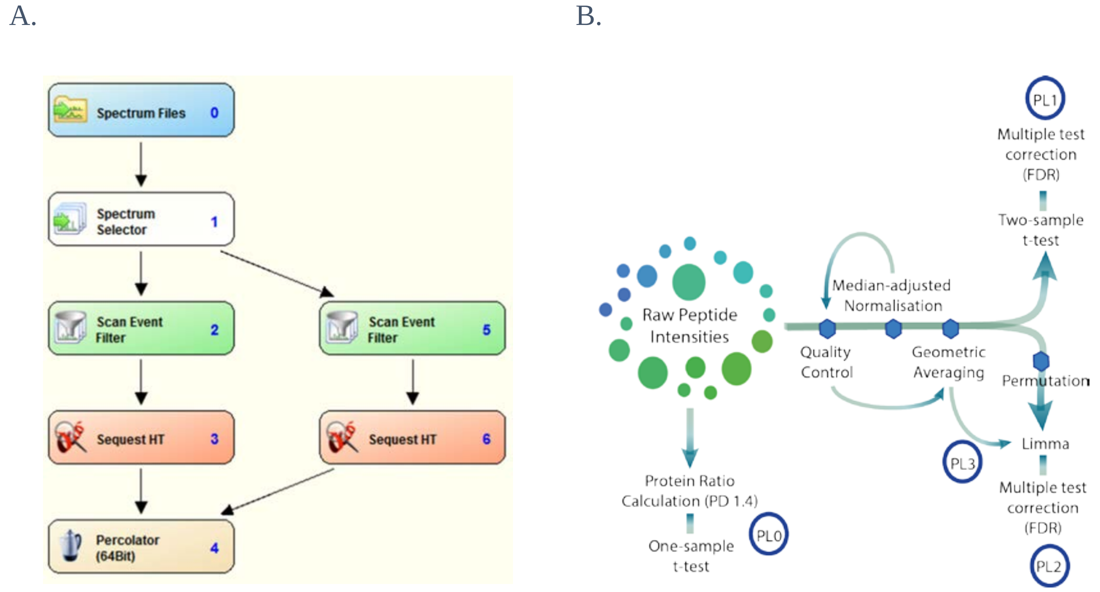


*Figure 29. Statistical and bioinformatic pipelines*

*A. Spectra analysis is performed using Proteome Discoverer 1.4. SequestHT is used for the target decoy searching for tryptic peptides. All spectra are searched against a customised fasta file containing UniProtKB SwissProt human proteome and Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv). Peptide spectrum matches (PSM) of reporter ions are extracted and rejected if any channels were absent. Workflow generated in Proteome Discovery 1.4. B. Statistical pipeline for raw intensities data. PL0 involves the calculation and direct extraction of protein ratios from Proteome Discoverer (PD 1.4). For pipelines P1, PL2and PL3 peptide intensities are inspected by quality control, then median-adjusted normalised and log-2 transformed.*

*Normalised peptide intensities are averaged using geometric mean, evaluated by quality control again and used for protein expression levels calculation. These values are used for significance assessment of effect size between groups using two-sample t-test (PL1), or Limma with or without previous ratio-based permutation (PL2 and PL3, respectively). Differences considered significant when p value < 0.05). PL1: Pipeline 1, PL2: Pipeline 2 and PL3: Pipeline 3.*

To determine an appropriate statistical method for assessing differential abundance, three different pipelines were designed and tested including; two-sample t-test comparing control and tuberculosis individuals and, permutation based on ratios and linear modelling using the R package (version 3.1.2) linear models for microarray data "LIMMA", followed by multiple test correction using FDR. RStudio (version 3.1.2) was used for data analysis, generation of box-and-whiskers plots, principal component analysis (PCA) or multidimensional scaling (MDS) plots, hierarchical clustering and heatmaps. GraphPad Prism 7 was used to draw volcano plots and chromatographic profiles of SEC and RP-HPLC separations.

### 2.6.1 Gene Ontology Analysis

ToppGene Suite (213) or the Cytoscape platform (version 3.1.2) was employed to process biological network visualisation and data integration. ClueGO (Biological Networks Gene Ontology tool) is a Cytoscape plug-in based for visualisation of non-redundant biological terms for gene clusters. CluGO network are generated using kappa statistics and reflects the association between the terms based on the similarity of their associated genes (214). Biolayout Express 3D was used as orthogonal bioinformatic strategy to *p*-value based statistics for data analysis, since Biolayout is based on co-expression network analysis (215)

## 2.7 ELISA and Luminex validation

Proteins selected for validation from the proteomic discovery experiments were measured in two different cohorts: a cross-sectional study with participants enrolled in South Africa and the Multifunctional Integrated Microsystem for rapid point-of-care TB IdentifiCation (MIMIC), a multi-centre cohort enrolled in the UK (211). Patient samples from the South African group included in the discovery were selected from this latter cohort. Both cohorts have suitable ethical approval for sample collection and informed consent was provided to participants.

### 2.7.1 ELISA

Plasma/serum samples from the South African and MIMIC cohort were analysed using commercially available ELISA kits from 2B Scientific Ltd, Upper Heyford, UK and Caltag Medsystems Ltd, Buckingham, UK. ELISA kits were performed according to manufacturer's directions. In brief, each ELISA involved pre-coated plates with primary antibody against the antigen of interest. A calibration curve was prepared by serial dilution of concentrated standards and 50μL to 100μl of each sample and standard were placed in the plates. The plates were covered and incubated for 30 minutes to 2 hours at 37 °C. Next, the plates were washed 3-5 times manually and 50μl of biotinylated antibody were dispensed to each well followed by an incubation for 2 hours at 37 °C. A further step of washing was performed and 50µl-100µl of HRP-avidin were dispensed to each well incubating for 1 hour at 37 °C. After a third step of washing, 90µl to 50µl of TMB substrate were added to each well incubating for 10 to 30 minutes at 37 °C. The reaction was stopped with 50µl of acidic solution. OD was determined using a microplate reader set at 450nm and wavelength correction at 540nm was not recommended. Volumes and incubation times vary between kits.

### 2.7.2 Luminex

Plasma/serum samples from the South African and Peruvian cohort were analysed on the Luminex Bio-Plex 200 platform to determine fluoresce intensities or concentration of proteins prioritised for validation. Luminex beads were custom-made and purchased from Protavio Ltd., Stevenage Bioscience Catalyst, UK. Bio-Plex manager software was used to construct standard curves and interpolate unknown concentrations. Briefly, standard curves for each protein were prepared by making eight three-fold serial dilutions from the corresponding analytes concentrated standard with sample diluent. Dilution factors were optimised for each protein. To run the assays, 50µl of beads mix were dispensed to each well of a pre-wet plate. 35µl of standards and samples were placed in each corresponding well. The plate was covered with foil and incubated for 90 minutes at 900rpm at room temperature. After incubation, the plate was washed twice and then, 20µl of diluted biotinylated antibody were added to each well incubating for one hour at 900rpm at room temperature. The plate was washed as before, and 50µl of diluted Streptavidin-PE was dispensed to each well incubating for 30 minutes. Finally, the plate was washed and the microparticles were resuspended with wash 65µL buffer followed by incubation for ten minutes on a shaker. The plate was read using the Luminex 200 platform and Bio-Plex manager software was used for data analysis.

GraphPad Prism 7 was used to visualise the data from ELISA and Luminex data. Group comparisons using Mann-Whitney test or Kruskal – Wallis test and Dunn’s multiple comparison correction to determine significant differences was performed on GraphPad Prism 7. Significant differences considered when $p < 0.05$.

## 2.8 Contribution statement

This PhD work was conducted by Diana J. Garay-Baquero under the supervision of Professor Paul Elkington, Dr. Spiro Garbis and Dr. Christopher Woelk. The work performed by Diana J. Garay-Baquero included optimisation of the MudPIT strategy, proteomic profiling of the South African and Peruvian samples, data analysis in R, Cytoscape, Biolayout Express 3D, Graphpad Prism and SSPS, validation experiments by ELISA and Luminex on the MIMIC and South African cohort and production of this thesis. Importantly, many people contributed to the development of this work, names and contributions are listed below.

- Dr. Cory White: provided bioinformatics input. Dr. Cory wrote and provided the R code for peptide normalisation, transformation, calculation of protein expression, LIMMA, permutation and multiple testing correction.
- Dr. Naomi F. Walker: conducted collection, clinical characterisation and shipping of South African samples for discovery and validation.
- Dr. Cesar Ugarte Gil: conducted collection, clinical characterisation and shipping of Peruvian samples for discovery.
- Dr. Marc Tebruegge: Principal Investigator of MIMIC study. Dr. Marc provided MIMIC samples for validation.
- Hannah Schiff: managed MIMIC cohort and assisted the validation experiments including ELISA and Luminex measurements.
- Dr. Ashley Heinson: provided initial training in R environment.
- Dr. Antigoni Manousopoulou: provided initial training in the proteomic platform and method for plasma profiling. During this training period the preliminary experiments presented in chapter 3 were generated.
- Dr. Liku B. Tezera: provided training in the Luminex platform.
- Dr. Andres Vallejo: provided bioinformatic input for network analysis, particularly Biolayout Express 3D and GWCNA. Dr. Andres provided the R code for ComBat normalisation and GWCNA analysis.

# CHAPTER 3

## A Preliminary Exploration of the Active Tuberculosis Plasma Proteome for Method Optimisation

### 3.1 Introduction

Plasma mass spectrometry based proteomics offers a unique opportunity to identify new biomarkers, since serologic tests based on antibody detection can be translated relatively easy to limited access clinical settings.  However, the proteomics analysis of plasma poses many challenges. These include the extensive protein concentration range of over 10-orders of magnitude, and limited availability of standardized or validated methods. Furthermore, over 90% of the total protein mass is comprised of mainly albumin and IgGs and thus mask the presence of the lower abundant and more clinically viable proteins. During the last 10 years, some serum proteomic studies have been conducted aimed at discriminate active pulmonary tuberculosis cases (103). Importantly, these studies were conducted using depletion approaches to reduce the complexity of the matrix and a limited coverage of the proteome was reached. The most extensive proteomes published to date have profiled between 518 and 716 proteins (117, 153).

A deeper proteome coverage is required to generate and validate a universal biosignature including different ethnic origins under a most robust experimental design, challenges that are addressed in this work.

The central hypothesis of this work predicts that our quantitative MudPIT plasma proteome will discriminate active tuberculosis patients from healthy individuals. This chapter presents a preliminary exploration of the differential plasma proteomes of active tuberculosis using the MudPIT approach with iTRAQ quantification to test this hypothesis. Two different experiments are compared using a set of eight plasma samples comprising four healthy individuals and four active pulmonary tuberculosis patients from South Africa. Only segment 4 derived from SEC prefractionation was used for profiling.

The original analysis methods reported by Garbis, S.D., et al (2011) and Al-Daghri, N. M., *et al* (2014) were used for these preliminary experiments (92). The resulting preliminary profiles guided the further development and optimisation of our MudPIT approach to reduce sample-processing biases and increase coverage and depth of the plasma proteome. This chapter presents further adjustment of SEC prefractionation, offline and online separation. Importantly, orthogonality between offline and online separation is a critical feature for multidimensional separations. Therefore in this study, high pH C4 chemistry separation is explored as an optional substitute of C8, considering that C4 is more hydrophilic and should increase orthogonality and therefore the more efficient offline prefractionation process prior to the online separation based on low pH C18 chemistry. This combination of reverse phase chemistries enabled the reduction of peptide co-isolation prior to their mass spectrometric analysis, which in turn allowed for their improved relative quantitation and increased proteome coverage. On the other hand, as a result of iTRAQ labelling, a broad peak containing underivatised iTRAQ is consistently eluted in the early retention times of the offline separation, masking the most hydrophilic peptides. I developed a method for cleaning of these fractions applying solid phase extraction to recover highly hydrophilic peptides. This chapter presents the development process which led to the final method described in Figure 28 and subsequently used in Chapters 4 and 5. One set of 8 samples from the South African cohort were used for a preliminary exploration of the plasma proteome in the context of active tuberculosis. A series of small experiments were undertaken to increase the analytical power of this MudPIT approach and presented in this chapter.

## 3.2 Methods

### 3.2.1 Patients Cohort

Recruitment and ethics for the South African cohort were described in the section 2.2. Healthy control individuals included in this preliminary study presented a mean age ± SD of 29.8 ± 6.8 (range 21-36 years) and BMI ± SD of 23.1 ± 4.4. In the case of active pulmonary TB patients, age ± SD was 28.0 ± 6.7 (range 24-38 years) and BMI ± SD was 20.0 ± 2.9. For these clinical characteristics, there was no significant difference between groups (at $p < 0.05$). Table 6 presents the clinical information of the individuals analysed in the pilot study. As previously stated in Table 4, this set of samples is uniquely used for technical refinement and method development presented in this chapter.

*Table 6. Clinical information of individuals participating in pilot study*

| *Variables* | *Healthy Controls* | *Pulmonary Tuberculosis* | *p value* |
|---|---|---|---|
| n | 4 | 4 | |
| Gender | Male (100%) | Male (100%) | |
| Mean age ± SD (years) | 29.8±6.8 | 28.0±6.7 | 0.725[(a)] |
| Age range (years) | 21-36 | 24-38 | |
| Mean BMI ± SD | 23.1 ± 4.4 | 20.0 ± 2.9 | 0.343[(a)] |
| Smoking History | | | |
| • Non smokers | 1 | 2 | 0.314[(b)] |
| • Current smokers | 3 | 1 | |
| • Ex-smokers | 0 | 1 | |
| Drug Treatment | | | |
| • None | 2 | 3 | 0.286[(b)] |
| • Amoxicillin | 1 | 0 | |
| • Vitamins | 1 | 0 | |
| • RHZE-Amoxicillin | 0 | 1 | |

[(a)] two-tailed *p*-value calculated by t-test

[(b)] two-tailed *p*-value calculated by Fischer's exact test

### 3.2.2 Sample Processing and Labelling

Plasma samples were prefractionated using size exclusion chromatography as generally described in section 2.5.2, dialysed and quantified. 100µg of protein were trypsin digested and labelled as defined in sections 2.5.4 to 2.5.6. The labelling scheme is presented in Figure 30.

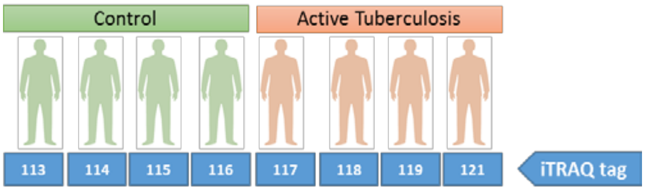


***Figure 30. iTRAQ labelling scheme of preliminary study.***

*Isobaric tags 113, 114, 115, and 116 were used for labelling healthy controls and tags 117, 118, 119, and 121 for active pulmonary tuberculosis patients.*

### 3.2.3 Peptide Fractionation

Offline peptide fractionation was based on high pH Reverse phase (RP) chromatography using the Waters, XBridge C8 column (3.5µm, 3.0mm x 150mm). Mobile phase (A) was composed of 0.10% ammonium hydroxide and mobile phase (B) was composed of 99.90% acetonitrile and 0.10% ammonium hydroxide. The pooled mixture of tryptic peptides was reconstituted with 100µL of mobile phase (98% mobile phase A and 2% mobile phase B) and separated, with gradually increasing mobile phase B according to the following program: 10 minutes isocratic 2% (B), for 10 minutes gradient up to 5% (B), for 60 minutes up to 20% (B), for 25 minutes up to 85% (B) and 10 minutes isocratic 85% (B) at a flow rate 0.30mL/min and 30° C. The fractions were collected in a peak-dependent fashion during the entire gradient elution phase. The peptide fractions were finally dried with speedvac concentrator for 4−5 h and stored at −20 °C until the LC−MS analysis at 30°C.

### 3.2.4 LC-FT-Orbitrap MS Analysis

The LC−MS experiments were performed on the Dionex Ultimate 3000 UHPLC system coupled with the high resolution nano-ESI-LTQ-Velos Pro Orbitrap-Elite mass spectrometer (Thermo Scientific). Individual peptide fractions were reconstituted in 31μL of loading solution (2% acetonitrile, 0.1% formic acid). Scans were conducted at 30000FWHM for $MS^1$ and 15000FWHM for HCD $MS^2$ spectra.

### 3.2.5 Bioinformatics and Statistical Analysis

Spectra were processed and protein expression levels calculated through the pipeline presented in section 2.6.1. Protein expression fold changes were calculated by $log_2 TB_x - \frac{1}{4}\sum_{i=1}^{4} log_2\, C_i$, where $TB_x$ is each tuberculosis sample and $C_i$ is healthy control sample.

### 3.2.6. Solid Phase Extraction Cleaning Protocol (SPE) and C4 Chromatography

A set of experiments were designed to develop a SPE-based cleaning protocol for iTRAQ/TMT labelled peptides and a method for offline peptide fractionation using C4 chemistry to increase orthogonality in comparison with the C8 chemistry used for the preliminary experiments. Particularly, TMT was used for developing the cleaning protocol and to establish the conditions for C4 separations. TMT presents a higher multiplexing degree allowing for allocation of 2 more conditions than iTRAQ which is beneficial in this optimisation experiment. Additionally, optimised parameters are easily translated from TMT to iTRAQ since the labelling chemistry is highly comparable. HEK cell protein extracts were obtained following the methodology described in section 2.5.4. 120µg of protein per isobaric tag was digested and labelled with TMT tags (Thermo Scientific, UK) according to manufacturer's instructions as illustrated in the scheme presented in Figure 31. The TMT tags were split into two groups; tags 126, 127N, 127C, 128C and 128N were used for developing the SPE cleaning protocol and tags 129C, 129N, 130C, 130N and 131 were used to develop the C4 separation method and compare it to C8 chemistry. The first set of tags was pooled together and cleaned using a Gracepure SPE C18-AQ 100mg/1ml cartridge (Grace, St. Neots, UK). Cleaned peptides were then fractionated using C4 HPLC separation. The second set of samples did not undergo SPE cleaning and was directly separated using C4 chromatography. This first comparison allows to estimate peptide losses resulting from the SPE protocol. The fractions collected from the C4 separation in this second set were then pooled together, dried and reconstituted for C8 separation. This second experiment compares C4 *vs.* C8 chemistry. All the fractions collected through these experiments were LC-FT-MS/MS analysed as described in section 2.5.8.

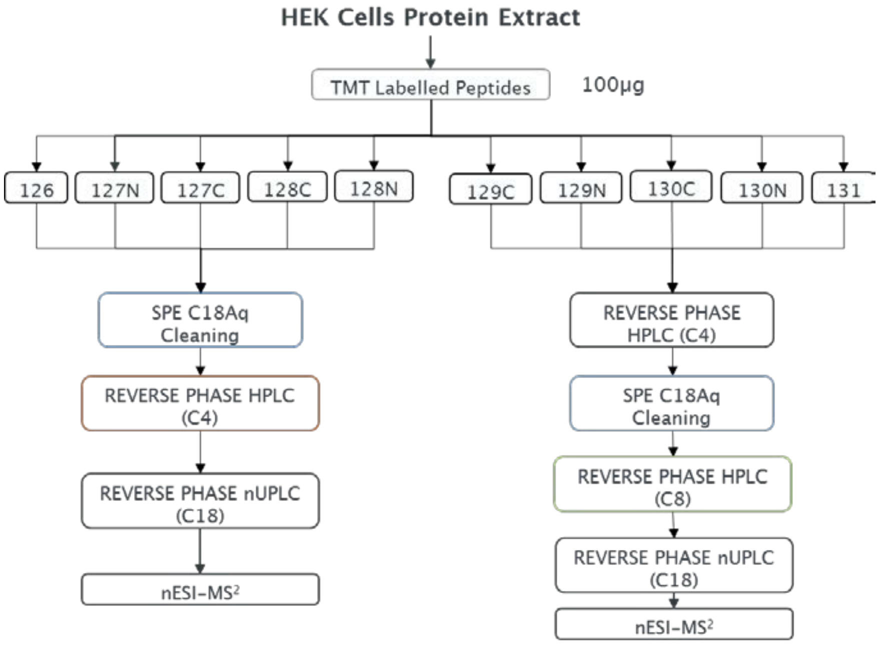


***Figure 31. Experimental design for SPE cleaning and C4 separation method development***

*HEK cell protein extracts were used to develop a SPE-cleaning method for TMT labelled peptides. Samples were TMT labelled and split into two groups. The first group of samples were subjected to SPE cleaning, followed by C4 HPLC separation and nUPLC-MS/MS analysis. On the other hand, the second set of samples were directly separated by C4 HPLC, then, the fractions were pooled together and separated using C8 chemistry followed by nUPLC-MS/MS analysis of the fractions.*

## 3.3 Results

### 3.3.1 Evaluation of Missingness

Raw data (peptide intensities) generated from tandem mass spectrometry analysis from the segment 4 were first evaluated at the peptide level to explore the incidence of missingness in the data set. This is a common issue in reporter ion based quantification due to the multiplex nature of the technology and it has a significant impact on the protein quantification. In fact, most of iTRAQ ratios are reported using uniquely the peptides that were quantified in all the samples. Missingness was evaluated in both data sets. Missingness maps were generated using the Amelia package in R. Maps show the peptides that were identified, blue areas represents quantified peptides and missing values are grey. These maps facilitate a general assessment of the magnitude of missingness in each sample. Figure 32A presents the complete data set generated in the experiment 1 and Figure 32B represents the data set generated in the experiment 2. Experiment 1 comprised 15706 peptides and 9211 (58.65%) were quantified in all the samples. On the other hand in experiment 2, 35357 peptides were identified and 16468 peptides were fully quantified (46.58%). Figure 32C and D present the distribution of missing values in each sample. Although experiment 2 identified 2.3 more peptides than experiment 1, the number of missing values were significantly higher, reducing the dataset by more than half. Interestingly, both experiments exhibits a similar pattern of missingness, with the sample labelled with tag 114 presenting the lower missingness in contrast to samples labelled with tag 115 and 118, which presented the highest number of missing values.

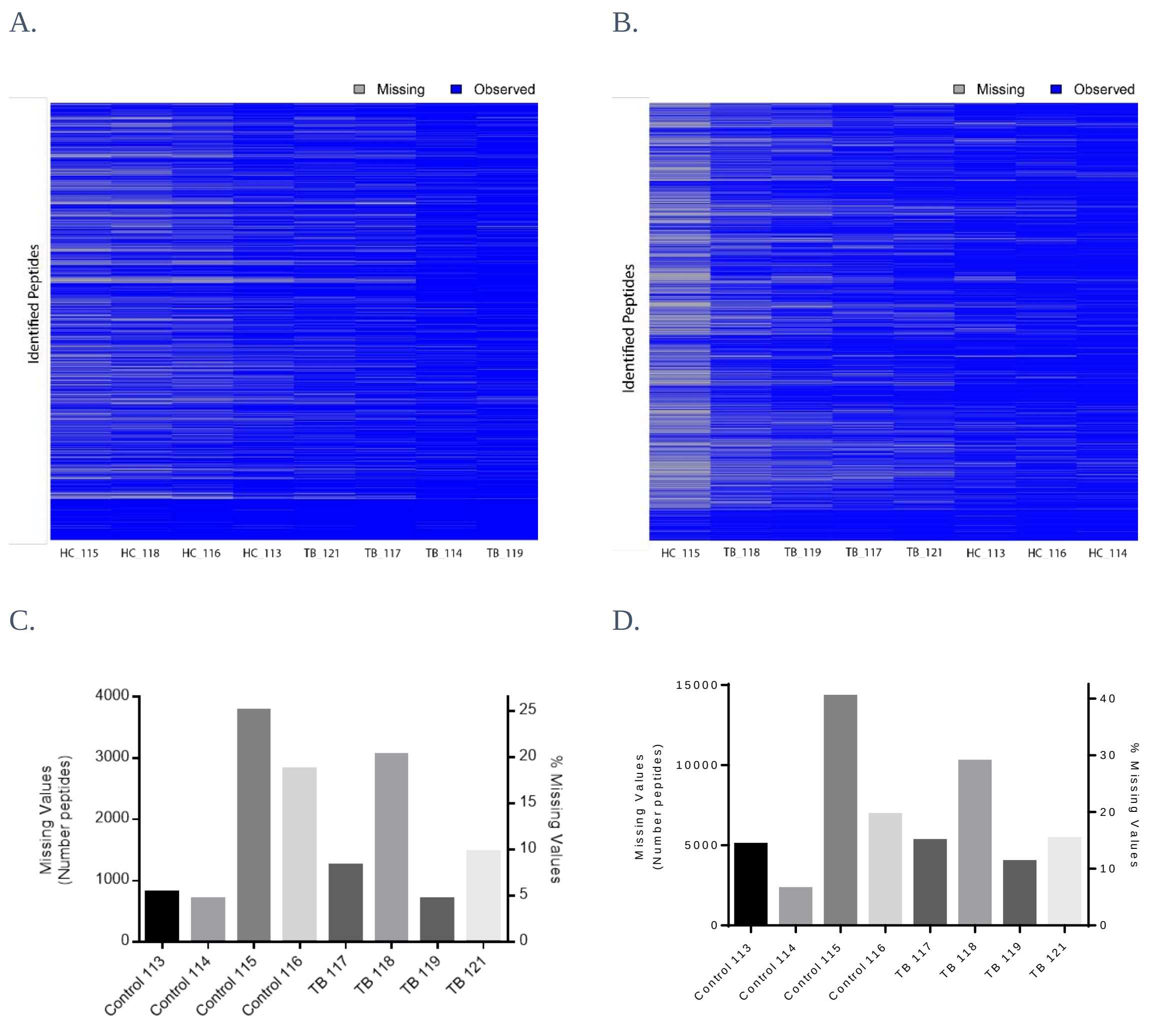


***Figure 32. Exploration of missingness***

***A – B.*** *Peptide coverage plots of raw peptides intensities showing missing values in each sample from experiment 1 and 2, respectively.* ***C – D.*** *Number and percentage of missing values in each channel from experiment 1 and 2, respectively.*

### 3.3.2 Quality Control and Normalisation at Peptide Level

A separate data set containing exclusively the fully quantified peptides was generated from each experiment; the distribution of the peptide intensities in each sample was evaluated using box and whiskers plots as shown in Figure 33A and B for experiment 1 and 2, respectively. Data was normalised using a median adjusted normalisation approach as described in section 2.6.1. Normalised peptide intensities distribution per channel is shown in Figure 33C and D for experiment 1 and 2, respectively. Principal component analysis was performed to evaluate clustering of data at peptide level. Importantly, most of variance in the data was explained with two components. Control and active tuberculosis groups were clearly discriminated as shown in Figure 33E and F for experiment 1 and 2, respectively.

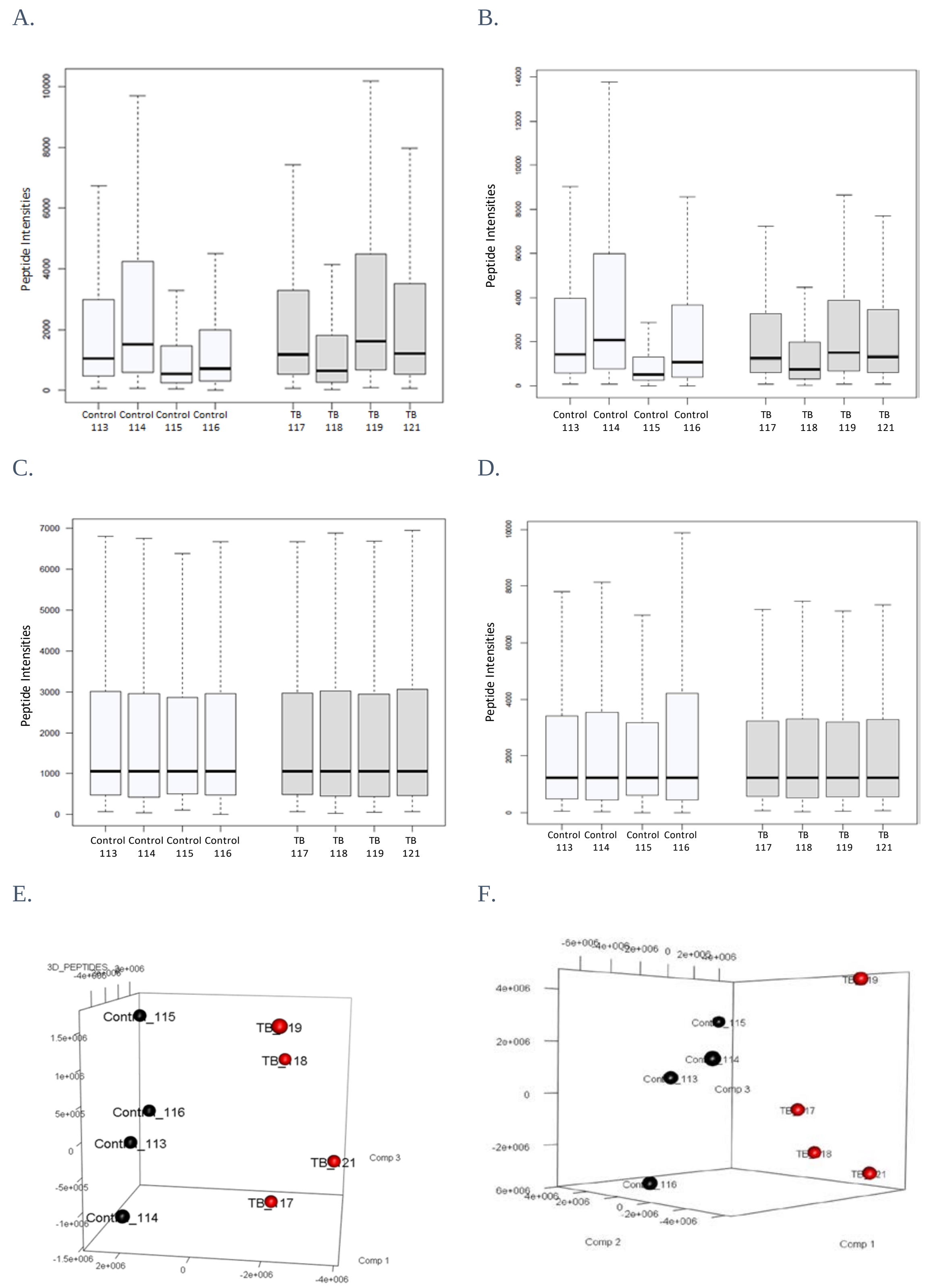


*Figure 33. Quality control analysis of raw data intensities*

***A- B.*** *Box and whiskers plots showing distribution of raw data intensities in each sample from experiment 1 and 2, respectively.* ***C – D.*** *Median-adjusted normalisation of raw intensities, experiment 1 and 2, respectively.* ***C – D.*** *Principal component analysis at peptide level, in black control individuals and red pulmonary tuberculosis patients for experiments 1 and 2, respectively.*

### 3.3.3 Partial Differential Plasma Proteome in Active Pulmonary Tuberculosis

The matrix generated using the median-adjusted normalised data was used to infer protein expression as described in section 2.6.1. In experiments 1 and 2, 557 and 500 proteins, respectively, were identified and fully quantified. 399 proteins were common between both experiments representing 60.6% of the total number of the quantified proteins. Pearson clustering was performed to evaluate grouping at the protein level between experiments presented in the Figure 34A. Overall, two main groups are observed discriminating control individuals from tuberculosis patients consistently in both experiments. The sample labelled with the tag 115 from experiment 1 appears as an outlier. Within the tuberculosis cluster, most of the samples group together according to the tag from both experiments. On the other hand, samples from the control cluster do not present clear discrimination between experiments or tags. Pearson correlation of common proteins between both experiments for each sample is presented in Figure 34B. R squared ranged from 0.5381 to 0.7270 for tag 115 and 118, respectively. Principal component analysis of common proteins clearly distinguish the tuberculosis group from the control group and indicates batch effects between experiments as shown in Figure 34C.

Significant modulation of protein expression in each experiment was assessed applying two different statistical strategies: one sample t-test and two sample t-test. Figure 34D presents the number of proteins significantly modulated calculated by both approaches in each experiment. Thirty-two proteins were common between experiments and confirmed by both tests.

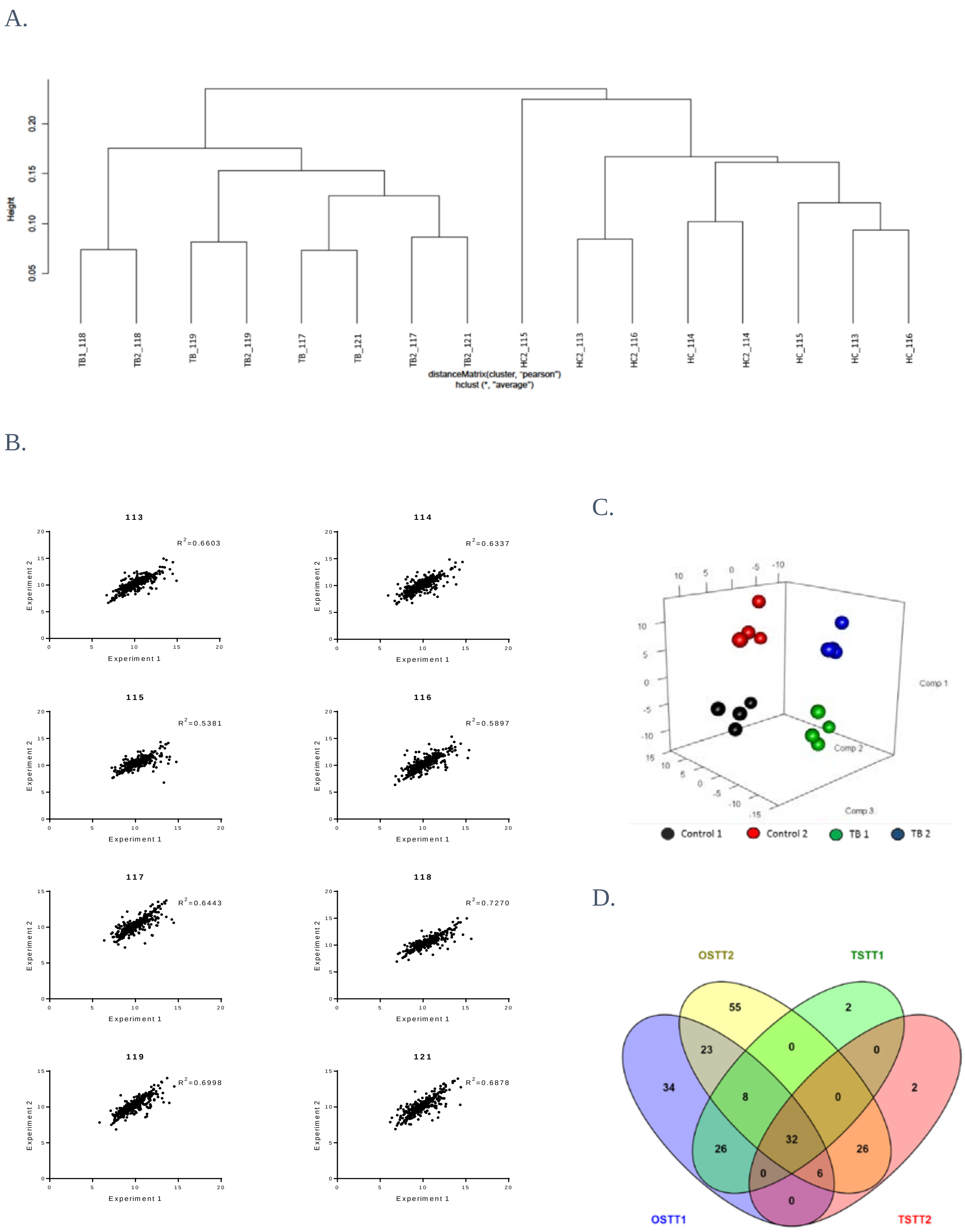


*Figure 34. Experiment comparison at protein level: clustering, correlation and statistical assessment*

***A.*** *Pearson clustering of samples at protein level across experiments 1 and 2.* ***B.*** *Correlation between experiments 1 and 2 across samples, $R^2$ calculated as Pearson correlation.* ***C.*** *Principal component analysis at protein level of control and tuberculosis groups in each experiment.* ***D.*** *Differential expression of proteins between control and tuberculosis groups in experiments 1 and 2 was evaluated using two different statistical approaches, one-sample t-test and two-sample t-test. Common proteins between both experiments and statistical test are presented in a Venn diagram. OSTT2 One sample t-test experiment 2, TSTT1 Two-sample t-test experiment 1, OSTT1 One sample t-test experiment 1, and TSTT2 Two-sample t-test experiment 2.*

Fold changes of these 32 proteins exhibit a high correlation between experiments ($R^2$ spanning from 0.9611 to 0.9820) as presented in Figure 35. The patterns of upregulation and downregulation are consistent among samples and experiments; 23 proteins were upregulated and 9 downregulated. Proteins and fold-changes are summarised in the Annexe 1.

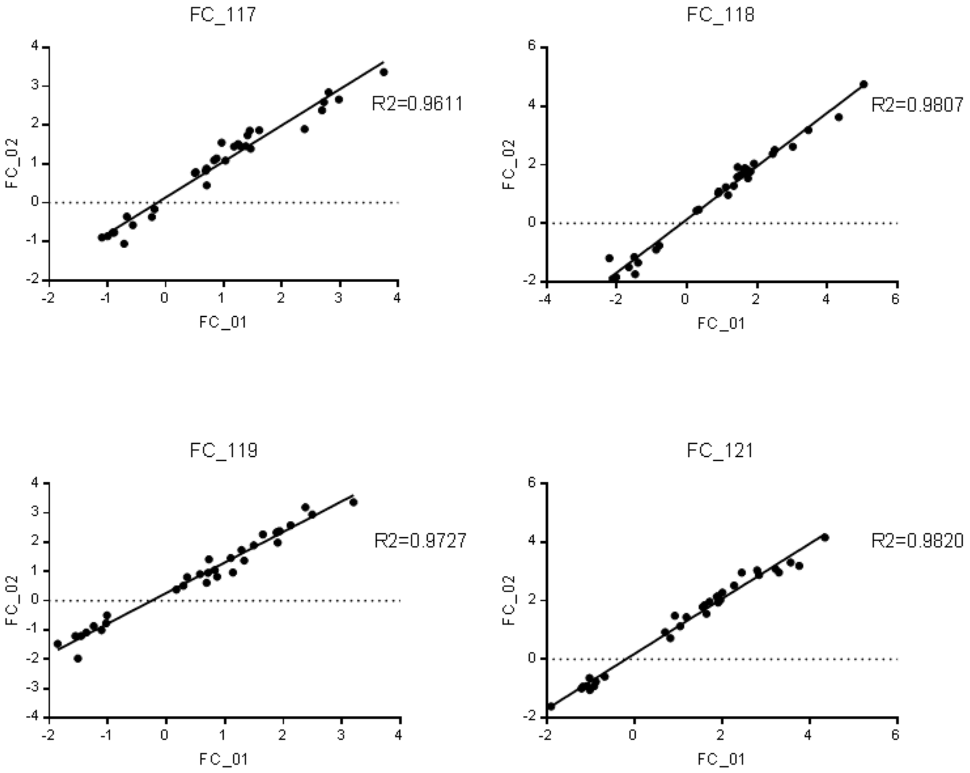


***Figure 35. Common proteins differently modulated between experiments 1 and 2***

*Thirty-two proteins were found common between experiments and confirmed by two independent statistical tests. Fold-changes correlation between experiments 1 and 2 across samples, $R^2$ calculated as Pearson correlation. FC_01: Fold-change experiment 1 and FC_02: Fold-change experiment 2.*

Although these preliminary experiments indicate that plasma proteome can be used for distinguishing tuberculosis patients form healthy donors and presents an acceptable reproducibility, the plasma proteome coverage is limited only to maximum 577 proteins, which is similar to previous reports. Various refinements to the method were necessary to improve its performance and increase the analytical power of this approach, including optimisation of the chromatographic parameters, orthogonality between offline and online separations and peptide recoveries from offline separations. These refinements are summarised in the following sections.

### 3.3.4 Size Exclusion Chromatography Optimisation

Three main parameters were evaluated for optimisation of SEC prefractionation of plasma: temperature of the column oven, number of columns and flow rate. The separation was performed using isocratic gradient of 6M guanidine HCl and 10% methanol and detection using UV-signal response at 280nm. Initially, temperature of the column oven was standardised and three different points were tested: 40°C, 42°C and 45°C. Figure 36A shows the separation of 120µL of human serum using 5 columns and a flow rate of 1.2mL/min. Separation conducted at 45°C showed sharper peaks suggesting better resolution without evident effect on retention times and separation profile. This temperature was selected for further analysis. Subsequently, the number of columns and flow rate was standardised.

Separation of 120µL human plasma was evaluated using 3, 4 and 5 columns. Figure 36B presents the chromatographic trace obtained with 3 columns 8.0mm I.D. x 300mm Shodex KW-804 serially connected operated at 45°C and 1.5mL/min. Figure 36C shows the traces obtained with 4 columns Shodex KW-804, 8.0mm I.D. x 300mm serially connected operated at 45°C and 1.0mL/min and 1.5mL/min. Under these conditions, the five segments are clearly separated and the fastest gradient expedites the separation 15 minutes with no detrimental effect on the separation resolution. Figure 36D presents the traces obtained using a 2x1x2 column configuration, this is: 2 columns Shodex KW-804, 8.0mm I.D. x 300mm, one column Shodex KW-802.5, 8.mm I.D. x 300mm and 2 columns Shodex KW-804, 8.0mm I.D. x 300mm serially connected, operated at 45°C and 1.5mL/min. The main differences between columns KW-804 and KW-802.5 are that the first one with particle size of 7 micron and maximum pore size of 1500Å exhibits a higher exclusion limit (1000000Da), whereas the column KW-802.5, with particle size of 5 micron and maximum pore size of 400Å, offers 4000 theoretical plates (TP) more than the KW-804. Figure 36D additionally shows duplicate runs for testing the reproducibility of the separation.

Considering that sharper peaks were obtained using 4 and 5 columns, these two conditions were further compared. Figure 36E compares the separation achieved with 4 and 5 columns (2x1x2 configuration), and Figure 36F presents the calibration curves obtained with 4 and 5 columns. BEH450 SEC (test mix Waters, Milford, USA) was used as a standard. This mix comprises 6 proteins: thyroglobulin dimer (1.40x10$^6$Da), thyroglobulin (6.69x10$^5$Da), IgG (1.50x10$^5$Da), bovine serum albumin (6.64x10$^4$Da), myoglobin (1.70 x10$^4$Da) and uracil (112Da).

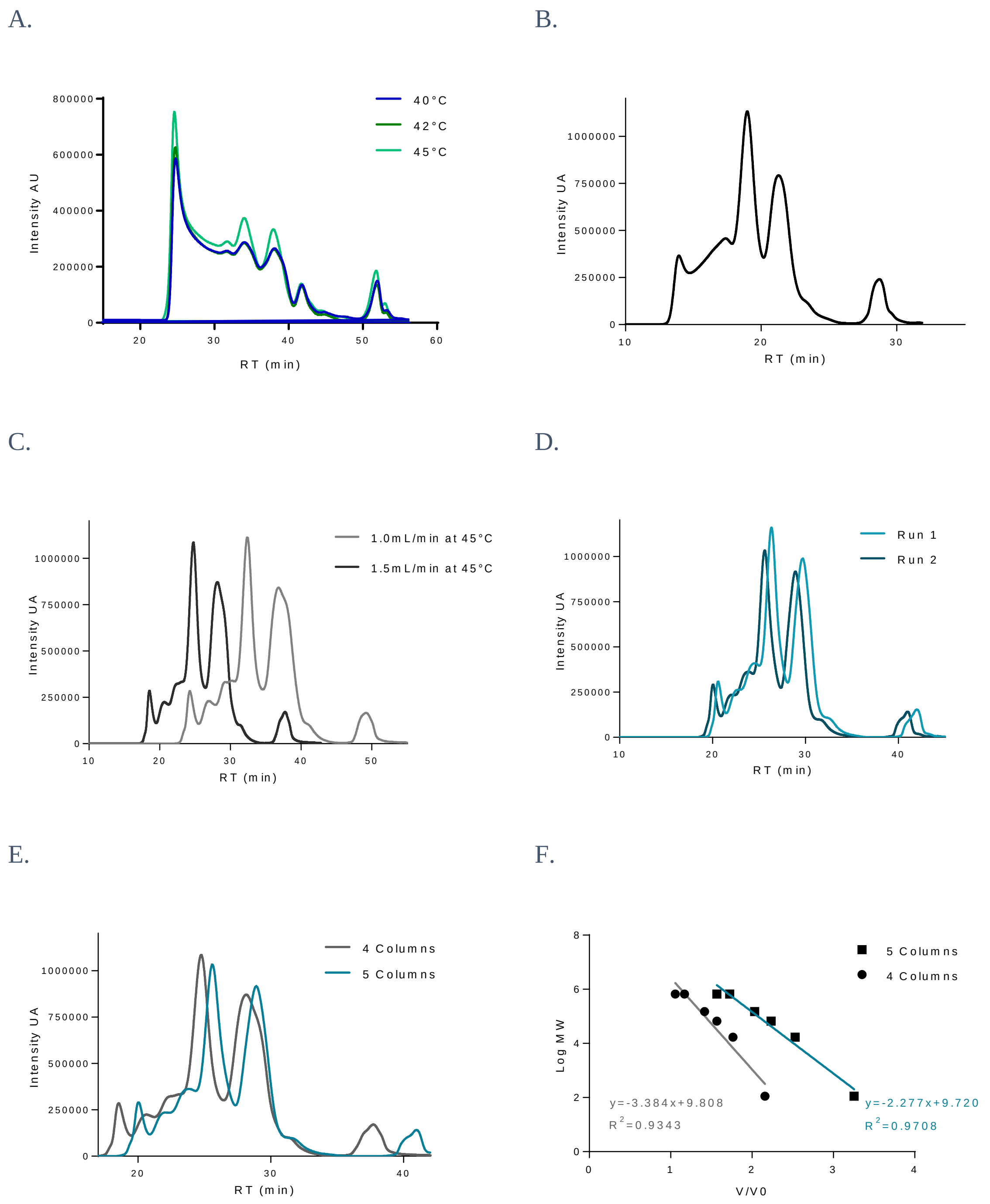


***Figure 36. Optimisation of chromatographic parameters for size exclusion fractionation (SEC)***

***A.*** *Temperature optimisation for SEC separation. Separation of human serum was conducted at three different temperatures, 5 columns serially connected were used and a flow rate of 1.2mL/min.* ***B.*** *Chromatographic trace of human plasma using 3 columns Shodex KW-804, 8.0mm I.D. x 300mm serially connected operated at 45°C and 1.5mL/min.* ***C.*** *Chromatographic traces of human plasma using 4 columns Shodex KW-804, 8.0mm I.D. x 300mm serially connected operated at 45°C.* ***D.*** *Technical replicate of human plasma SEC fractionation using 5 columns: 2 columns Shodex KW-804, 8.0mm I.D. x 300mm, one column Shodex KW-802.5, 8.mm I.D. x 300mm and 2 columns Shodex KW-804 serially connected, operated at 45°C and 1.5mL/min.* ***E.*** *Comparison between traces obtained using 4 and 5 columns (2x1x2 configuration) operated at 45°C and 1.5mL/min.* ***F.*** *SEC calibration curve using BEH450 and comparing separation with 4 and 5 columns, operated at 45°C and 1.5mL/min.*

The 5 column configuration did not increase time of separation substantially and sharper peaks were achieved, suggesting higher separation resolution. Additionally, the calibration curve indicates a better linearity using 5 columns. Taking these data together, parameters for SEC separation were selected: 5 columns (2x1x2 configuration) operated under an isocratic gradient at 1.5mL/min and 45°C.

### 3.3.5 Dialysis Purification and Protein Quantification

Following SEC separation, each segment is processed separately. The 8 samples that constitute one iTRAQ set were subjected to dialysis purification, the conditions reported previously by Al-Daghri, N., *et al.*(92), were kept but only ultrapure water was used as exchange solvent. Ultra-filtration based protein purification protocols were verified to be inferior to the dialysis exchange techniques in the original Garbis, S., *et al.* (2011) report and were therefore not further examined.

In terms of protein quantification, the method used for the preliminary experiments (infrared detection system) was replaced by Nanodrop. Precision and accuracy of measurements were investigated to compare both systems. For this purpose, a known amount of albumin was weighted and dissolved in 1mL of 0.5M TEAB and 0.05% SDS. Table 7 presents the evaluation of both systems. Considering that the Nanodrop system exhibits a significantly higher performance than the infrared system, it is selected as quantification method for following experiments.

***Table 7. Comparison IR quantification system and Nanodrop performance***

| *Parameter* | *Infrared detection system* | *Nanodrop system* |
|---|---|---|
| Reference concentration (mg/mL) | 3.40 | 4.80 |
| $\bar{x}$(mg/mL) | 4.27 | 4.25 |
| SD | 0.49 | 0.02 |
| %CV | 11.6 | 0.49 |
| %Error | 25.6 | 11.5 |

### 3.3.6 Solid Phase Extraction Cleaning Protocol and C4 Chromatography

Once protein from the 8 samples was quantified, trypsin digested and iTRAQ labelled, peptides were pooled and separated using RP-HPLC. From the pilot experiments discussed previously in this chapter, two features are required to be optimised in order to increase proteome coverage: orthogonality between offline and online separations and recovery of peptides masked by underivatised labelling reagents.

This set of experiments was performed using TMT labelling, an analogue of iTRAQ, following the workflow described in the Figure 31. The results obtained in this section are applied to iTRAQ labelling as will be demonstrated in the next chapters. TMT comprises 10 isotopic tags that were used to simultaneously optimise a suitable gradient for C4 based RP-HPLC separation and develop a method for cleaning labelled peptides using GracePure C18-Aq SPE cartridges (100mg, 50um particle size, 60Å pore size, GracePure, Hichrom, UK). Figure 37 depicts the workflow developed for cleaning, ensuring the minimal peptide losses. Ligands required to be activated using 100% acetonitrile and subsequently equilibrated using the same solvent for peptide reconstitution; 1% acetonitrile and 0.01% formic acid. Peptide fractions selected for cleaning are pooled and loaded into the cartridge, then samples and eluents need to be slowly dropwise eluted from the column (~ 4 drop/min). The eluents collected once the peptide pool is passed through the cartridge is loaded again twice in order to ensure the maximum interaction with the ligands. The next step involves washing the unbound material with two volumes of 1% acetonitrile and 0.01% formic acid. The elution of peptides is conducted stepwise, starting from 2% to 70% acetonitrile and 1% formic acid as shown in Figure 37, this constitutes an elution cycle. Complete elution of peptides requires two elution cycles.

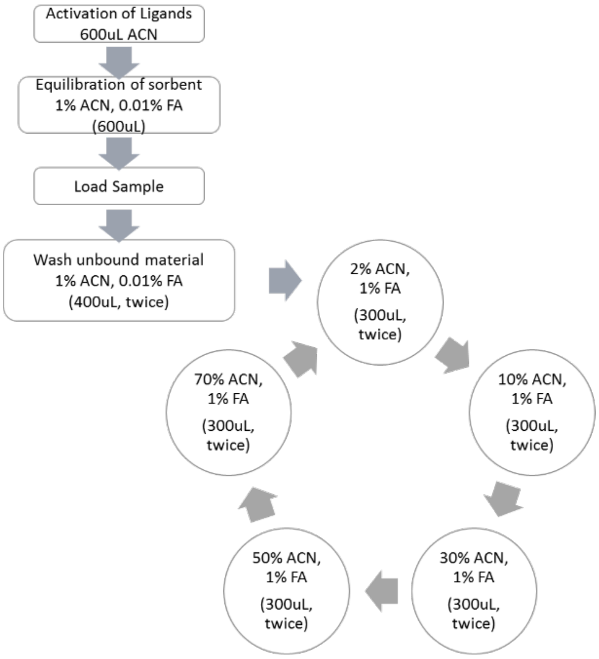
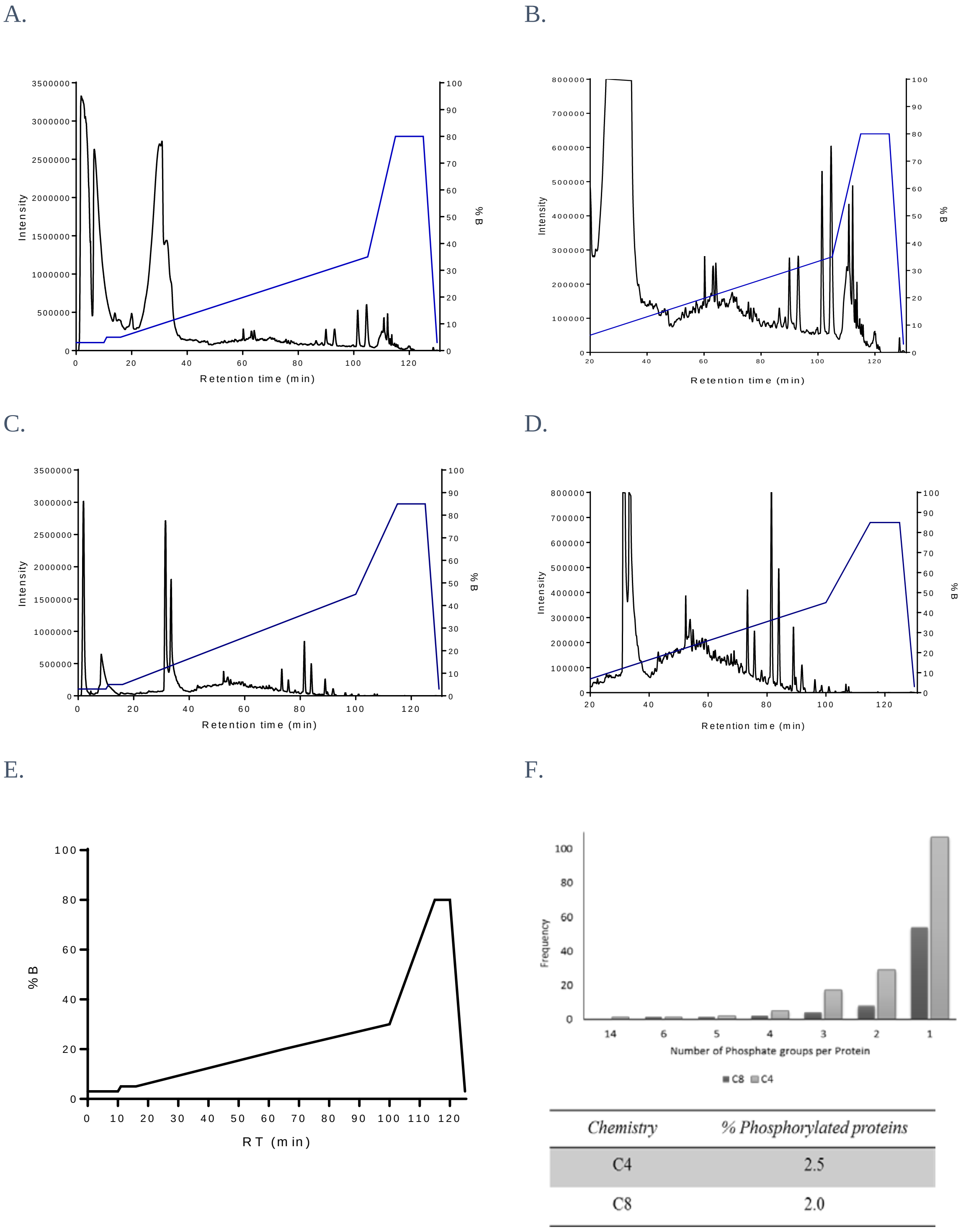


***Figure 37. Solid phase extraction protocol for iTRAQ/TMT labelled peptides***

*Protocol developed for peptide fractions subjected to iTRAQ/TMT labelling. It comprises five main steps: (1) Activation of ligands, (2) Equilibration of the stationary phase, (3) Loading sample, (4) Washing of unbound material, and (5) Stepwise elution of peptides.*

In order to evaluate peptide losses and the C4 gradient, labelled samples were split into two groups as shown in Figure 31. Considering that in this particular experiment a number of conditions are tested simultaneously (C4 vs. C8 chemistry, chromatographic gradients and SPE cleaning), it does not provide a definite comparison but it serves to indicate more optimal conditions for further experiments. In order to preliminarily evaluate the peptide loses resulting from the SPE protocol one set of 5 pooled samples were directly separated using a C4 column with a step gradient from 3% to 35% phase B in 105 minutes, followed by washing and equilibration steps. A set of 5 pooled samples was subjected to SPE cleaning and subsequently fractionated with a chromatographic gradient from 3% to 45% phase B in 100 minutes followed by washing and equilibration steps. Figure 38A presents the traces obtained for the first group of pooled peptides and B a details section of the chromatic window from 20 to 130 minutes of the chromatogram presented in Figure 38A. On the other hand, Figure 38C shows the separation of the second set of labels and D a detailed section from 20 to 130 minutes of the trace presented in Figure 38C. The pH of mobile phases was optimised to ensure alkaline pH within the working limit of the stationary phase of the C4 column. Mobile phase A is 0.08% ammonium hydroxide (pH~ 8). and mobile phase B is 99.92% acetonitrile and 0.08% ammonium hydroxide The step gradient from 3% to 35% phase B resulted in a better distribution of peaks along the chromatographic window and therefore a higher resolution of the separation. Figure 38E presents the optimised gradient for offline fractionation of labelled peptides using C4 chemistry.

The method reported by Al-Daghri, N., *et al.* (2014) was used for the fractionation of the labelled peptides using a C8 column, and this same approach was used to profile the preliminary tuberculosis proteome derived from segment 4 (SEC) as previously presented. A comparison between C4 and C8 was conducted to confirm suitability for proteome profiling. The C4 fractions collected from the first group of samples (no SPE cleaning) were pooled and subjected to SPE cleaning followed by C8 RP-HPLC separation. Fractions from both experiments C4 and C8 were subjected to nUPLC (C18) coupled to MS/MS analysis. Although C4 moiety is a reverse phase ligand, it is expected to interact in a greater extent with hydrophilic peptides than C8 chemistry and therefore to be more orthogonal to C18 (online separation). Certain posttranslational modifications such as phosphorylation increases hydrophilicity of proteins/peptides. The fraction of phosphorylated peptides was evaluated in both data sets as an indication of the ability of the column to capture such more hydrophilic peptides. Figure 38F presents a comparison between C4 and C8 chemistries. Although C4 captured only 0.5% more phosphorylated proteins and, significantly more proteins containing multiple phosphate groups were selectively retained by the C4 stationary phase. Proteins carrying up to 14 phosphate groups where identified using the C4 column in contrast to the C8 column, which allowed the identification of proteins carrying up to 6 phosphate groups.



| *Chemistry* | *% Phosphorylated proteins* |
|---|---|
| C4 | 2.5 |
| C8 | 2.0 |

***Figure 38. C4 Peptide fractionation***

*A. C4 peptide fractionation of pooled TMT labelled peptides with tags 126, 127C, 127N, 128C and 128N, conducted at 0.30mL/min and 30 $^{o}$ C, **B.** Chromatographic window of the previous trace showing a peptide rich region at relatively low intensity. **C.** C4 peptide fractionation of pooled TMT labelled peptides with tags 129C, 129N, 130C, 130N and 131 and subjected to SPE cleaning. Separation was conducted at 0.30mL/min and 30 $^{o}$ C **D.** Chromatographic window of the previous trace showing new captured peptides and a peptide rich region at relatively low intensity. **E.** Gradient of elution optimised for fractionation of iTRAQ/TMT labelled peptides. **F.** Enrichment of phosphorylated proteins using C4 chemistry compared to C8 separation.*

*Both data sets were compared in terms of number of proteins with variable number of phosphate groups and percentage of phosphorylated proteins.*

## 3.4 Discussion

iTRAQ data analysis imposes various processing challenges, and one common issue is data missingness. Considering the nature of this quantification strategy, the intensity of the observed peptide intensities depends upon diverse factors such as the abundance of the protein originating the peptide, the sensitivity of the instrument, variable amounts of loaded sample, ion suppression, and ionisation and fragmentation properties of peptides. Previously, it has been reported that this missingness is a non-random phenomenon in iTRAQ/TMT quantification (216). Missingness occurrence in the data sets analysed in this preliminary study suggest that there are significant batch effects between experiments (Figure 33E and F). Particularly, experiment 2 showed a lower performance in terms of quantification; although considerably more peptides were identified than in the experiment 1, only 46.58% of these peptides were completely quantified. Parameters such as trypsin digestion, labelling conditions and performance of the mass spectrometer must be kept highly controlled and standardised. Additionally, the number of missing values showed a clear pattern related to either sample identity, isobaric tag or both variables. This finding suggests that an adjusted experimental design should involve a block-randomised approach to distribute the samples and tag variability. The differential performance in terms of peptide intensities of these two preliminary experiments run as technical replicates highlights the importance of standardisation of the methods used and the conditions of the mass spectrometric devices.

In terms of proteome coverage improvement, this chapter presents optimisation and standardisation of the MudPIT method for plasma/serum proteomics first reported by Garbis, S., *et al*. (2011), and more recently optimised for iTRAQ quantification by Al-Daghri, N., *et al.* (2014) where exclusively segment 4 was explored (35, 92). Specifically, method development was focused on SEC separation parameters, offline to online orthogonality and iTRAQ labelled peptide fractions cleaning aimed to increase protein coverage and reproducibility.

With more than a half-a-century of history, SEC is a well-known chromatographic technique widely used for protein separation. However, its applicability has been relatively overlooked in the proteomics field, mainly due to low resolution and detrimental dilution of the protein fractions. Importantly, new SEC methods are under development such as ultrahigh pressure (UHP)-SEC for rapid and high-resolution separation of intact proteins for shotgun proteomics, which considerably accelerates the separation to few minutes and increases resolution by using small particle sizes (2µm) and organic/inorganic hybrid materials as BEH (ethylene bridged hybrid) (217).

On the other hand, method development of SEC separation aimed to plasma/serum proteomics has shown to significantly increase proteome coverage compared to depletion methods (35). However, a proper optimisation of SEC separation considering parameters as number of columns, temperature and flow rate has not been previously reported. The separation of proteins with SEC is based on their differences in hydrodynamic radius, which is analogous to their differences in MW. In other words, the larger the protein MW the larger the hydrodynamic radius that can be achieved under the appropriate chaotropic conditions of the sample solvent and matched mobile phase.

A very important element to the practice of SEC is the versatility of mobile phase conditions that dictates the resulting hydrodynamic radius of the proteins of interest along with other factures (i.e. viscosity). In this case, aqueous 6M guanidinium HCL and 10% methanol exhibits multimodal effects; it neutralises all protease activity, thus stopping from any additional protein degradation, and dissolves lipid micro-vesicular species while extracting their intact protein content (i.e. exosome-associated proteins). Additionally, physico-chemical properties of Guanidinium HCL effectively disrupt protein-protein or protein-ligand/co-factor interactions (albumin bound proteins); it serves as an excellent liquid-fixative thus capturing all in-situ clinically occurring events at the time of sampling. In a recent report, Guanidinium HCL was used for effectively studying exosome-derived proteins and its role in tumour metastasis (218). The efficient protein unfolding properties achieved at 6M GuaHCl also maximizes the hydrodynamic radius of all proteins, and therefore ensures their better SEC separation.

SEC is used in this MudPIT method as a preparative step, therefore only a crude separation of plasma samples into five segments according to molecular size is expected. Systematic evaluation of temperature (40°C, 42°C and 45°C), number of columns serially connected (3, 4 and 5) and flow rates (1.0mL/min, 1.2mL/min and 1.5mL/min) was conducted to determine the best conditions for plasma pre-fractionation. Linearised calibration curves were generated and goodness-of-fit of linear regression as $R^2$ evaluated to compare separation traces obtained with 4 and 5 columns (Figure 36F). The thyroglobulin dimer ($1.40x10^6$Da) standard is shown as an outlier in the curves of calibration, which is expected considering that the maximum exclusion limit of the column is $1.0x10^6$Da. Selected standardised chromatographic parameters for SEC were: 45°C, 1.5mL/min and 5 columns (2x1x2 configuration) for plasma prefractionation. Although these columns still exhibit sub-optimal sample carryover effects, not easily observable carryover was detected by the typical UV-based HPLC detection systems.

Introduction of iTRAQ to this MudPIT method for plasma proteomics was reported elsewhere (92). The manufacturer (AB Sciex, UK) recommends cation-exchange chromatography (for simple mixtures), high resolution cation-exchange chromatography (for complex mixtures) or ZipTip® for removal of reagents used for labelling such as buffer salts, SDS, high concentrations of organic solvents and underivatised iTRAQ/TMT. This is in part accomplished by the C8/C4 offline chromatography, which elutes these reagents at very early retention times. In the trace shown in the Figure 38A, these contaminants are eluted over the first 15 minutes and a second broad peak is eluted from the minute 22 to minute 38. Although contaminants are separated and eluted during the offline separation, some peptides with similar retention times might co-elute with them. Here, an easy and fast method for cleaning of iTRAQ/TMT labelled peptides was developed using solid phase extraction (SPE).

Peptide losses resulting from the SPE process were qualitatively assessed comparing peak intensities between the traces obtained with and without SPE cleaning. In both cases, the peptide peaks separated from the minute 40 to the minute 110 exhibited maximum intensities close to 400000AU at 215nm and the patterns of separation were similar. Notably, the broader peaks were considerably minimised and new peptide peaks were captured (Figure 38D). This cleaning protocol includes few steps of elution, low losses and the sorbent capacity allows loading of the peptide pool in one step, which is easier and faster than other methods.

The MudPIT approach relies completely on comprehensive multidimensional chromatographic techniques which significantly increases the peak capacity per time unit, compared to their unidimensional equivalents. Nevertheless, the maximum separation among analytes requires independent mechanisms of separation, which translates into a higher degree of orthogonality. Peptide separation imposes some challenges such as solubilisation, compatibility between dimensions, throughput, additive compatibility for mass spectrometry analysis and limitations to the range of liquid chromatographic techniques to couple. The preliminary plasma proteome profile was generated using C8 based RP-HPLC as offline dimension, however, a more hydrophilic chemistry is likely to benefit orthogonality considering that C18 chemistry is widely used in shotgun proteomics.

Data from mass spectrometric analysis of the C4 and C8 fractions were compared. Considering that the pooled peptides separated by C8 underwent two chromatographic separations which increases peptide losses, the results cannot be directly compared in terms of absolute number of proteins. However, the main purpose of using C4 instead of C8 is to increase the separation of analytes between the offline and online dimensions, therefore assessment of the proportion of hydrophilic peptides identified by each method can be used as an indicator of orthogonality. Moreover, certain posttranslational modifications such as phosphorylation increases the hydrophilicity of the proteins, hence the percentage of phosphorylated proteins identified by C4 and C8 was used to evaluate selective enrichment of hydrophilic peptides. As a consequence of phosphorylation, amino acids gain a double negative charge at physiological pH and, generally, this results in decreasing hydrophobicity as a result of two rearrangements in the protein surface in the phosphosites:  exposure of hydrophilic residues or/and burial of hydrophobic amino acids  (219). C4 chemistry enriched 0.5% more phosphorylated proteins than C8 and, interestingly, C4 was able to capture a wider variety of proteins carrying multiple phosphate groups up to 14 phosphate groups (Figure 38F).  Although, a modest increase of phosphopeptides was achieved with the C4 column, RP LC-MS approach generally exhibits poor retention behaviour of the more hydrophilic phosphopeptides. Moreover, no prior phosphopeptide chemical affinity enrichment was used (i.e. HILIC, ZrO2, TiO2, Fe- or Cu- IMAC, etc.) which usually is preferred for phosphoproteomic studies.

Additionally to the sample preparation, the bottleneck in systems biology research is processing and mining of large data sets and the central problem of accurately relating information derived from –omics experiments to biological processes. In other fields such as microarray analysis, relevant advances has been made detecting and removing experimental biases from experimental datasets using a wide range of methods (from simple scaling to non-parametric quantile normalisation). However, analysis methods of data from LC/MS-based proteomics tend to be relatively simple and are still under development. Raw data derived from proteomics experiments must be normalised to produce more accurate estimates of the underlying biological effects. Normalisation reduces the effect of outliers on the dataset and removes atypical signals resulting from experimental and instrumental biases. The median-adjusted normalisation approach applied in this preliminary study effectively centred the data as shown in Figure 33C and D. Global normalisation methods are widely used in iTRAQ proteomics, here the main objective is to realign the observed intensity distributions of the reporter ions from each quantification channel, such the median or mean of the distribution is equal across all the channels (220). Notably, control and tuberculosis groups are clearly discriminated at peptide (Figures 33E and F) and protein level (Figures 34A and C) in both experiments, even when batch effects are evident. These results indicate that this MudPIT approach is able to capture the differential proteome in active pulmonary tuberculosis using a partial analysis of the plasma proteome.

Protein expression was calculated from geometric averaging of normalised peptide intensities per protein and two separate statistical methods were applied to determine the significantly modulated proteins: One-sample t-test and two-sample t-test. In the first case, the fold changes were calculated and, then compared to 0 in order to test the null hypothesis. In the second case, control and tuberculosis groups were compared. The limitation of this first statistical approach is that t-test based statistics is not suitable for small sample sizes and assumes normal distribution. Although the protein expression was logarithmically transformed in this case, the distribution of data is not entirely normal. Assessment of size effect for discovery proteomic experiments benefit of more robust statistical approaches and correction for multiple testing. These different strategies will be explored in the following chapters.

The statistical approach based on t-test resulted in the identification of significantly modulated proteins reported in previous proteomic studies such as; serum amyloid A, transthyretin and C - reactive protein (114), apolipoprotein C-II and retinol binding protein 4 (36), S100-A9 (MRP14) (119). Importantly, the patterns of modulation were consistent with the previous reports. Additionally, new proteins were identified as significantly modulated which offers new opportunities for candidates.

The results of this first exploration of the active pulmonary tuberculosis proteome using a powerful multidimensional method are promising in terms of revealing novel biomarkers for early diagnosis of the infection. Importantly, this pilot study served to generate a “training” data set to explore the main features and structure of the generated data.  Additionally, the results pinpointed specific areas in the experimental design and methods that required further optimisation and standardisation aimed to increase proteome coverage, reproducibility and validity of the discovery phase.

In summary, chromatographic parameters for SEC plasma fractionation were standardised and defined for further application. In addition, a fast method for SPE based-cleaning for iTRAQ/TMT labelled peptides was developed. In parallel, a reverse phase chromatographic method for peptide fractionation based on C4 chemistry was established. A higher enrichment of hydrophilic proteins was achieved with C4 than C8, which suggests an improvement of the orthogonality between offline and online dimensions. These different modifications on the method previously reported (89) are aimed to increase coverage of the complex plasma proteome.  Having finalised my optimisation, I then proceeded to study the entire proteome (segments 1 – 4: Chapter 4) and then one segment in great detail to maximise statistical power (segment 4, Chapter 5).

# CHAPTER 4

## Comprehensive Plasma Protein Profile of Active Tuberculosis

### 4.1 Introduction

Previous chapters have presented preliminary evidence and method development demonstrating that the MudPIT strategy comprising SEC x RP-HPLC (C4/C8) x nUPLC (C18)-MS/MS can be used as a powerful tool to quantitatively profile the plasma proteome, a crucial milestone in the discovery of new biomarkers suitable for rapid tuberculosis diagnosis. Despite multiple serum proteomic studies in active pulmonary tuberculosis during the past 10 years, a universal biosignature remains elusive. As discussed in section 1.2.2, robust experimental designs encompassing confounding factors and proper validation are the most common pitfalls found in previous studies (103). Consequently, the biosignatures reported in these different studies exhibit poor correlation across proposed classifiers (section 1.3.3).

One of the main objectives of this research project is to profile the plasma proteome of active pulmonary tuberculosis applying a depletion-free optimised MudPIT approach. Taking together the data generated from the exploratory study previously discussed and the method optimisations presented in the Chapter 3, a more comprehensive experimental design is undertaken here. The discovery stage is performed uniquely with plasma samples from male individuals as tuberculosis immunopathology exhibits sexual dimorphism (14). Two different ethnicities (Peruvian and South African) are included in the experimental design. This allows exploration of alterations in the proteomic profiles driven by tuberculosis encompassing genetic background diversity and exposure to circulating *Mycobacterium* strains in different geographic locations (15, 117). Additionally, this approach will result in relevance of the protein biosignature to diverse ethnic contexts. In terms of experimental methods, additional quality control checkpoints are included in particular steps of the chromatographic fractionations. Additionally, a selection of four different statistical approaches is applied and the rationale used to select the subset of proteins chosen for further validation is presented.  This chapter presents the complete plasma profile of a set of seven samples, as first step of my research that aims to establish a universal plasma biosignature for active pulmonary tuberculosis suitable for early diagnosis.

## 4.2 Methods

### 4.2.1 Patient Cohort

Recruitment and ethics for the Peruvian and the South African participants are described in the section 2.2. As presented in the general description of the cohorts distribution, 7 participants were selected to profile the active pulmonary tuberculosis plasma proteome by analysing the SEC segments 1 to 4. Table 8 presents the clinical information of the individuals included in this experiments. Healthy control individuals included in this preliminary study presented a mean age ± SD of 29.3±6.43 (range 22-34 years) and BMI±SD of 23.17±1.58. In the case of active pulmonary TB patients, age ± SD was 32.3±6.65 (range 27-42 years) and BMI±SD was 21.07±2.35. For these clinical characteristics, there was no significant difference between groups ($p < 0.05$).

*Table 8. Clinical information of individuals participating in the proteome profiling study*

| *Variables* | *Healthy Controls* | *Pulmonary Tuberculosis* | *p Value* |
|---|---|---|---|
| n | 3 | 4 | |
| Gender | Male (100%) | Male (100%) | |
| Mean age ± SD (years) | 29.3±6.43 | 32.3±6.65 | 0.586[a] |
| Age range (years) | 22-34 | 27-42 | |
| Mean BMI ± SD | 23.17 ± 1.58 | 21.07 ± 2.35 | 0.143[a] |
| Smoking History | | | |
| • Non-smokers | 1 | 1 | 0.999[b] |
| • Current smokers | 2 | 2 | |
| • Ex-smokers | 0 | 1 | |
| Drug Treatment | | | |
| • None | 3 | 4 | |

[a] two-tailed *p*-value calculated by t-test

[b] two-tailed *p*-value calculated by Fischer’s exact test

### 4.2.2 Experimental Design

The plasma proteome of active pulmonary tuberculosis was profiled including samples of participants from South Africa and Peru. Additionally, a master pool was included for controlling variability across experiments. An aliquot of 20µL of all the samples available from South African control, Peruvian control, South African active tuberculosis and Peruvian active pulmonary, were pooled together to prepare the master pool and aliquoted to prevent freeze-thaw cycles. Figure 39 illustrates allocation of plasma samples within the 8-plex. Samples were randomised and the tag 113 was assigned to the master pool, which is maintained among 8-plexes.

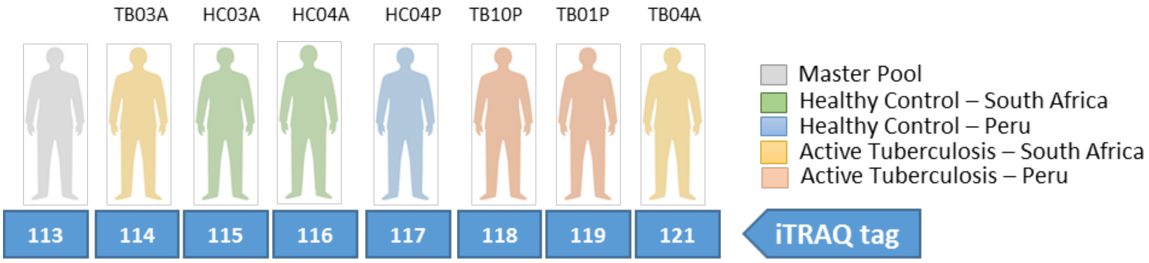


***Figure 39. Experimental design active pulmonary tuberculosis plasma proteome profile***

*Plasma samples allocation within the iTRAQ 8-plex. Control and tuberculosis samples from both ethnicities were randomised using the tool available in https://www.random.org/. The tag 113 was assigned to the master pool for variability control between -8-plex experiments.*

### 4.2.3 Sample Processing

Twenty-nine samples in total from both ethnicities available for this work, in addition to four aliquots of master pool, were individually subjected to SEC prefractionation under the optimised conditions discussed in section 3.4. These conditions includes: 5 columns configuration: 2 columns Shodex KW-804, 8.0mm I.D. x 300mm, one column Shodex KW-802.5, 8.mm I.D. x 300mm and 2 columns Shodex KW-804 serially connected, operated at 45°C and 1.5mL/min under isocratic elution with 6M guanidine hydrochloride and 10% ethanol. The five SEC segments were collected in a peak-dependent fashion detected at 280nm and then stored at -20°C until further analysis. The five SEC fractions were collected but only the first four were subjected to downstream analysis. As previously reported by Garbis, S. D., *et al.* (2011) the protein content in plasma/serum is completely fractionated in the first four segments and segment five contains mainly small molecules such as metabolites (35).

The first four segments collected from the prefractionation of the seven plasma samples selected for this study and a master pool aliquot were dialysed and quantified as described in section 3.3.5. 120µg of protein was reduced, alkylated and trypsin digested overnight (16h) as presented in section 2.5.5. iTRAQ labelling was conducted for 2 hours (section 2.5.6) and labelled peptides were dried in a speed vac at room temperature. Fractionation of labelled peptides was conducted using offline C4 – HPLC; peptides were analytically reconstituted and pooled together with 100µg of 3% phase mobile B (99.92% acetonitrile and 0.08% ammonium hydroxide) and 97% phase A (99.92% water and 0.08% ammonium hydroxide). Pooled peptides were then centrifuged at 16000xg for 10 minutes. The pellet was stored at -20°C and the supernatant was injected for separation. Offline fractions were collected in a peak-dependent fashion and detected at 215nm. The elution gradient is illustrated in the Figure 38E using a Kromasil C4 column (3.5µm, 2.1mm x 150mm) operated at 35°C and 0.3mL/min.

Offline fractions containing contaminants, the early and late fractions, were pooled together with the pellet obtained in the previous step, then peptides were cleaned using the SPE protocol provided in the Figure 37 and separated again by C4 HPLC. All the fractions collected were dried overnight in speed vac at room temperature and stored at -80°C.

Offline fractions were reconstituted in 31µL of loading solution (2% acetonitrile and 1% formic acid) and separated using an AcclaimPepMap RSLC, 75µm× 25cm, nanoViper, C18, 2µm particle column retrofitted to a PicoTip emitter (FS360-20-10-D-20-C7) and analysed by the high resolution nano-ESI-LTQ-Velos Pro Orbitrap-Elite mass spectrometer (Thermo Scientific). Specifications for the mass spectrometric analysis are presented in section 2.5.8, specifically for these experiments the scans were acquired at a resolution of 30000 for CID and 15000 for HCD.

### 4.2.4 Data Analysis

The spectrum files were processed using Proteome Discoverer 1.4 following the workflow and specifications detailed in section 2.6.1. Isotopic correction factors for iTRAQ reporter ions were applied. Raw intensities for all the PSM were extracted at 1% FDR, median-adjusted normalised and log2 transformed. Normalised intensities were averaged to calculate protein expression. Data was prepared for analysis according to two general strategies:

a. Independent analysis of each segment
b. Merged analysis from multiconsensus report generated from Proteome Discoverer 1.4

Figure 29B in section 2.6.1 summarises the main statistical pipelines selected for this work. Following the initial normalisation of data and protein relative expression calculation, four pipelines for the assessment of size effect between control and tuberculosis groups are applied:

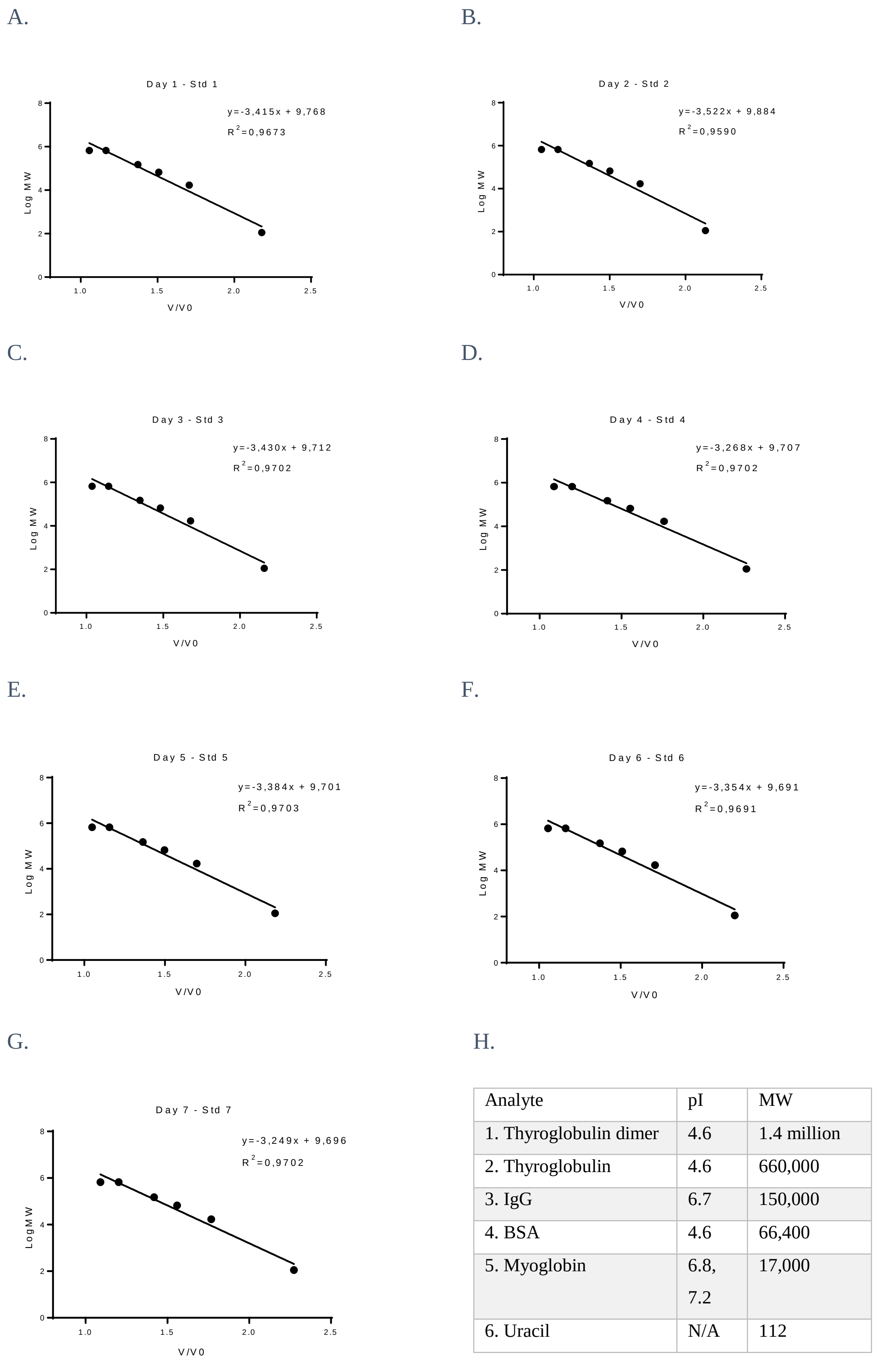
- (PL0) One-sample t-test: Ratios were calculated by $log_2 TB_x - \frac{1}{4}\sum_{i=1}^{4} log_2\, C_i$, where TB$_x$ is each tuberculosis sample and C$_i$ is healthy control sample and one-sample t-test applied considering as *null* hypothesis that the fold change is equal to zero.
- (PL1) Two- sample t-test: Protein expression of each sample are used to compare healthy control and tuberculosis groups where the *null* hypothesis implies that the mean of both groups is the same.
- (PL2) Permutation and LIMMA: Protein expression of each sample is used to calculate ratios, a permutation test (1000 cycles) is conducted to filter out ratios. Relative expression of proteins from filtered ratios where tested using linear modelling (221) (LIMMA built on R 3.3.3) to define significantly modulated proteins between groups.
- (PL3) LIMMA: Linear modelling using LIMMA in R environment was applied to assess significant size effects in protein expression between control and tuberculosis groups.

Box and whiskers plots, violin plots, heat maps and PCA analysis were generated in RStudio (version 3.3.1). Chromatographic traces, volcano and scatter plots were produced in GraphPad Prism 7. Gene ontology enrichment analysis was performed as detailed in section 2.6.2. Adobe Illustrator CS6 was used for final editing of figures.

## 4.3 Results

### 4.3.1 Prefractionation of Plasma Samples by Size Exclusion Chromatography

The complete available cohort of plasma samples collected from Peruvian and South African participants (29 samples) and four aliquots of the master pool were individually SEC pre-fractionated using the standardised conditions presented in the Chapter 4. Each chromatographic separation was conducted over 45 minutes and distributed in such manner as to require a minimum number of days to complete the separations. This precaution was taken in order to reduce day-to-day variability. Additionally, a pool of human sera was run as first analysis each day to ensure optimal technical conditions and the BEH450 SEC standard was separated daily to evaluate system performance and reproducibility. Figure 40A-G presents the calibration curve obtained from each day of analysis, $R^2$ values ranged from 0.9590 to 0.9703.



| Analyte | pI | MW |
|---|---|---|
| 1. Thyroglobulin dimer | 4.6 | 1.4 million |
| 2. Thyroglobulin | 4.6 | 660,000 |
| 3. IgG | 6.7 | 150,000 |
| 4. BSA | 4.6 | 66,400 |
| 5. Myoglobin | 6.8,<br>7.2 | 17,000 |
| 6. Uracil | N/A | 112 |

***Figure 40. Daily SEC calibration curves***

***A-G.*** *The BEH450 test mix containing six standard proteins with molecular weights ranging from 1.4x$10^6$Da to 112Da was separated on each day of analysis to evaluate the separation performance and calibration curves* *are presented (Normalised elution volume to void volume V/$V_0$vs. molecular weight, MW). Completion of prefractionation of the samples was achieved within 7 days. Equation of linear regression fit and R squared for measure of linearity are shown. H. Composition of standard mix by Waters http://www.waters.com/webassets/cms/support/docs/720003385en.pdf*

Figure 41A-E presents the chromatographic traces obtained from 33 plasma samples including master pool aliquots, from which the 7 samples for complete proteomic analysis as presented in Table 8 were selected. The patterns of separation particularly for the protein fraction, segments 1 to 4, revealed a consistent pattern of fractionation across groups with slight sample variation. Therefore, the five segments were independently collected in a peak-dependent fashion in order to maintain as much consistency as possible between segments from the different samples. Notably, the peak containing mostly metabolites (segment 5) exhibited important differences in terms of intensity and shape between Peruvian and South African specimens. Each collection is initiated in the exact inflection point between segments previously defined in section 2.5.2. Only one sample of the pulmonary tuberculosis group (TB03) from South Africa presents a shift in the separation time, however the pattern is completely consistent with the other samples.

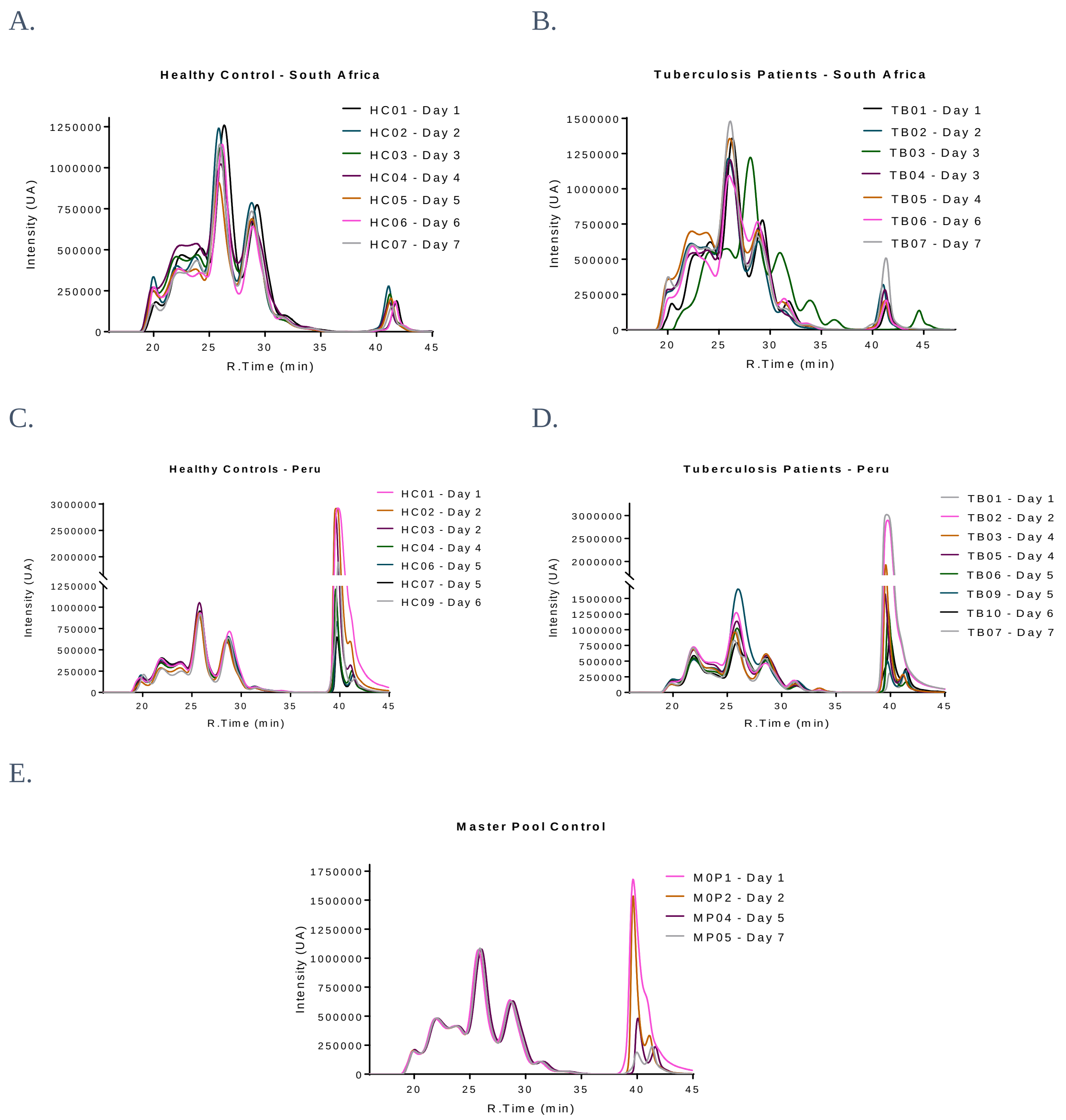


***Figure 41. SEC prefractionation of plasma samples***

*Isocratic chromatographic traces of plasma samples are presented. Intensity was evaluated at 280nm over 45 minutes. Separation was conducted at 1.5mL/min and 45°C.* ***A.*** *Overlapping of SEC chromatographic traces of healthy control plasma samples from South Africa.* ***B.*** *Overlapping of SEC chromatographic traces of active pulmonary tuberculosis plasma samples from South Africa.* ***C.*** *Overlapping of SEC chromatographic traces of healthy control plasma samples from Peru.* ***D.*** *Overlapping of SEC chromatographic traces of active pulmonary tuberculosis plasma samples from Peru.* ***E.*** *Overlapping of SEC chromatographic traces of master pool samples separation.*

Despite sample processing over 7 days, the collection time points presented low variation among samples within a same group as shown in Figure 42A-C, suggesting good technical reproducibility. All samples within the South African group, excepting TB03A where elution was delayed by about three minutes compared to the elution times in the group, presented a low standard deviation per segment ranging from 0.4113 to 0.8154 in segment 1 and 5, respectively. In the case of the Peruvian group, the standard deviation ranged from 0.1714 to 0.2866 in segment 1 and 5 and, finally in the case of the master pool the standard deviation ranged from 0.0763 to 0.3309 in segments 2 and 5.

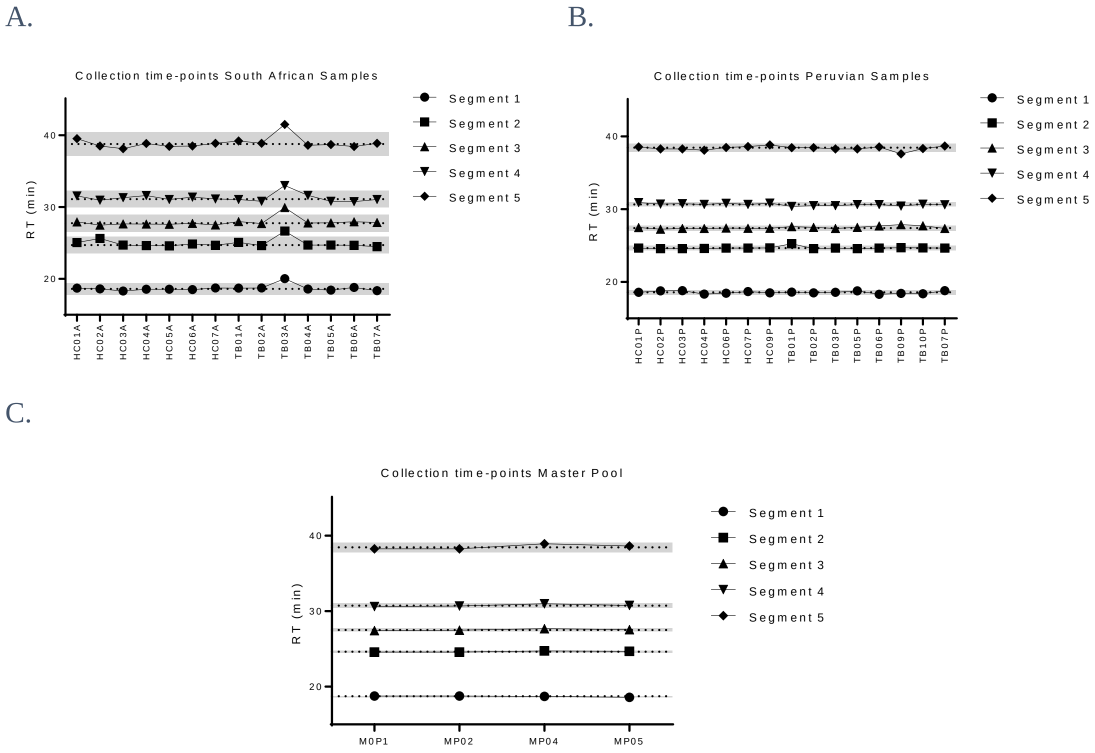
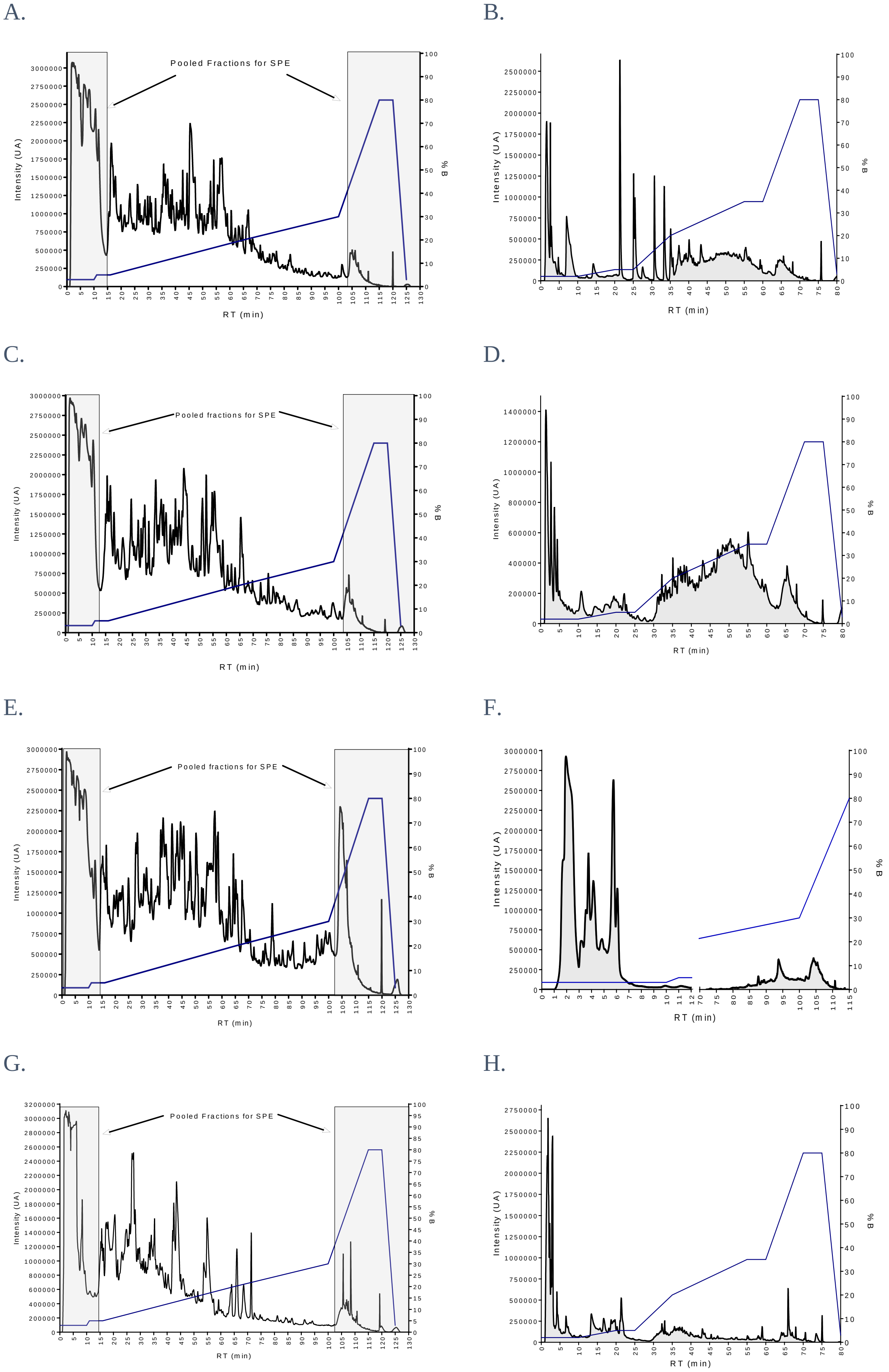


***Figure 42. Quality assessment of collection time-points in SEC separations***

*Collection time-points of each segment are plotted for each sample per group **A.** South African samples **B.** Peruvian samples and **C.** Master pool samples. (RT: retention times). Grey bands indicate 2SD.*

### 4.3.2 Peptide Fractionation by C4 nUPLC

From the block-randomised experimental design presented in Figure 39, the set of seven samples described in Table 8 and one master pool aliquot were chosen to conduct a complete profiling of the plasma proteome. Figure 43A shows the C4 chromatographic traces obtained from the fractionation of the pooled iTRAQ labelled peptides of segment 1. Highlighted peaks in grey indicate the fractions that were selected for SPE cleaning and additional C4 separation presented in Figure 43B. A total of one hundred and nine fractions were obtained for online nUPLC-MS/MS analysis in segment 1. Figure 43C presents C4 trace for segment 2 and Figure 43D SPE cleaned fractions. Ninety-one fractions were obtained in total for subsequent MS analysis. Figure 43E presents the C4 trace for segment 3 and Figure 43F SPE cleaned fractions. Ninety fractions were collected in total for MS analysis. Figure 40G presents C4 trace for segment 4 and Figure 43H SPE cleaned fractions. One hundred and two fractions were collected for subsequent MS analysis. Importantly SPE cleaned fractions from segment 3 were run using the same gradient applied for the initial fractionation, however the central region of the chromatogram appeared barely occupied, therefore a new gradient was adjusted to expedite the analysis. This gradient was used for SPE cleaned fractions C4 separations for segments 1, 2 and 4.



## *Figure 43. C4 HPLC chromatographic traces of iTRAQ labelled peptides*

*C4 chromatographic traces of pooled iTRAQ labelled peptides were performed at 0.3min/mL and 30°C. Intensity was evaluated at 215nm. Blue line indicates gradient of elution.* ***A.*** *Chromatogram of segment 1 separation. Highlighted areas were pooled together and subjected to SPE cleaning.* ***B.*** *Chromatogram of SPE* *cleaned fractions from segment 1 separation.* ***C.*** *Chromatogram of segment 2 separation. Highlighted areas were pooled together and subjected to SPE cleaning.* ***D.*** *Chromatogram of SPE cleaned fractions from segment 2 separation.* ***E.*** *Chromatogram of segment 3 separation. Highlighted areas were pooled together and subjected to SPE cleaning.* ***F.*** *Chromatogram of SPE cleaned fractions from segment 3 separation.* ***G.*** *Chromatogram of segment 4 separation. Highlighted areas were pooled together and subjected to SPE cleaning.* ***H.*** *Chromatogram of SPE cleaned fractions from segment 4 separation.*

### 4.3.3 Partial Profiling of Tuberculosis Plasma Proteome

#### 4.3.3.1 Analysis of independent segments: Exploration of the plasma proteome

Plasma proteome from segments 1 and 4 were MS/MS profiled, and each segment constitutes an independent experiment. Raw peptide intensities from each experiment were extracted from Proteome Discoverer 1.4 (Thermo Scientific) at 1% FDR and 5% FDR for comparison. The trade-off between the percentage of co-isolation excluding peptides from quantification and the number of fully quantified proteins in the preliminary experiments was considered to determine the cut-off for this parameter as presented in Figure 44. 50% co-isolation excluding peptides from quantification was selected as optimal cut-off and applied for extraction of all the data presented in this work.

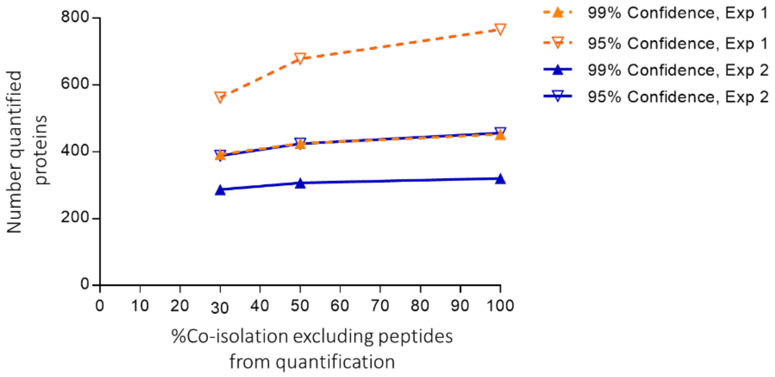


***Figure 44. Percentage of co-isolation excluding peptides from quantification***

*Number of identified proteins in the preliminary experiments 1 and 2 were compared using three different cut-off values: 30%, 50% and 100% co-isolation excluding peptides from quantification for proteins defined peptides extracted at 1% and 5%FDR confidence.*

Considering the total number of identified peptides in each dataset the percentage of iTRAQ labelled peptides was determined for each sample as shown in the Figure 45. Labelling efficiency was consistently over 70% among the four different experiments (segment 1 to 4); therefore, all the samples were kept for downstream analysis. The dispersion of the labelling efficiency ranged 2.70%CV in segment 1 to 9.09%CV in segment 4.

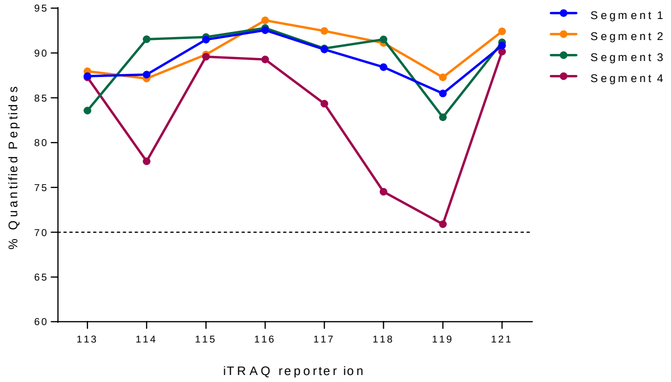
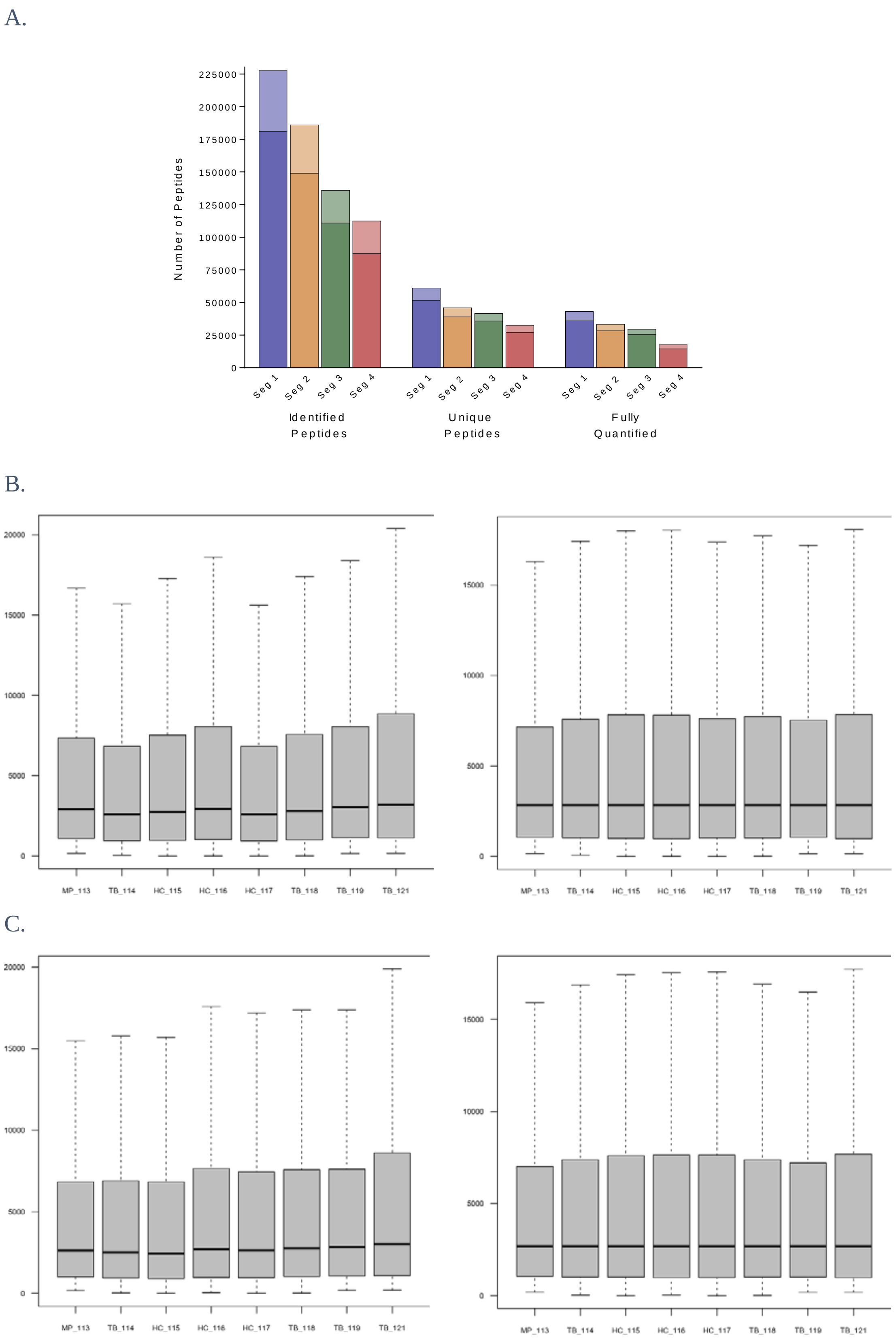


***Figure 45. iTRAQ labelling efficiency segments 1 to 4***

*As an indirect measurement of the labelling efficiency, the percentage of labelled peptides was calculated considering the number of quantified peptides labelled with a specific iTRAQ reporter in relation to the total number of identified peptides in the whole dataset. All the samples presented a labelling efficiency above 70%.*

Figure 46A presents the number of identified, unique and fully quantified peptides in each segment. In the context of this work, unique peptides refers to the peptides that are traceable to only a specific protein or protein group and fully quantified peptides indicate the number of peptides that were quantified in all 8 samples. Metrics for 1%FDR and 5%FDR are presented in the Figure 46A. In terms of number of peptides at 1%FDR; 181086 peptides were identified from which, 51622 were unique peptides used for quantification and 36657 peptides were quantified in all samples, in segment 1. In segment 2, 149073 peptides were identified, from which 39158 were unique peptides and 28428 peptides were quantified in all samples. In segment 3, 110938 peptides were identified, from which 35908 were unique peptides and 25558 peptides were quantified in all samples. In segment 4, 87582 peptides were identified, from which 26994 were unique peptides and 14612 peptides were quantified in all samples. Overall, in comparison to the peptide metrics extracted at 1%FDR, a 21% increase of peptide number is observed when data extracted at 5%FDR. Peptides resulting from the most strict FDR will be used for further statistical analysis.

Considering the technical variability inherent to shotgun proteomic experiments, the centre and total spread of the data distribution among different experiments was reasonably comparable. For instance, at 1% FDR the master pool from segment 1, median was 3793 and the IQR was 7000.25. On the other hand, at the same confidence in segment 3, median was 2671 and the IQR was 6931.40. Box and whisker boxplots for segments 1 to 4 are presented in the Figure 46B to E, respectively.


A.
Number of Peptides
225000
200000
175000
150000
125000
100000
75000
50000
25000
0
Seg 1
Seg 2
Seg 3
Seg 4
Identified Peptides
Unique Peptides
Fully Quantified
B.
20000
15000
10000
5000
0
MP_113
TB_114
HC_115
HC_116
HC_117
TB_118
TB_119
TB_121
C.

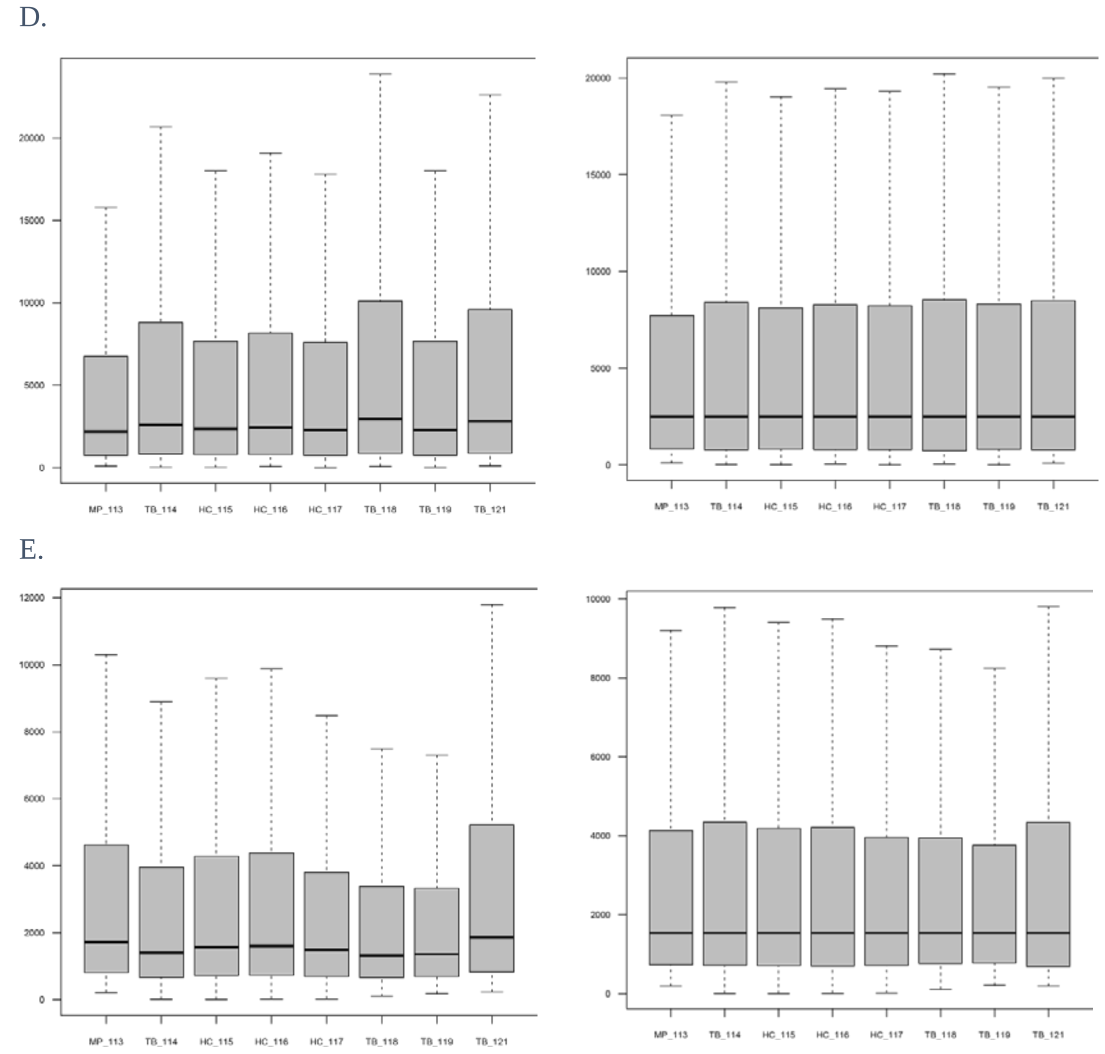


***Figure 46. Data exploration at peptide level of plasma proteome***

***A.** Total number of identified, unique and fully quantified peptides per segment. Dark colour represents the number of peptides extracted at 1%FDR and light colour the additional peptides obtained when extracted at 5%FDR. B – E. Box and whiskers plots of raw and normalised peptide intensities at 1%FDR. Median-adjusted normalisation was conducted. **B.** Segment 1; **C.** Segment 2; **D.** Segment 3; and **E.** Segment 4. Raw and normalised from left to right.*

Peptides extracted from Proteome Discover 1.4 (Thermo Fisher) at 1%FDR and 5%FDR were used to calculate the relative protein expression values as described in section 2.6.1. Proteins with missing values were included for the first inspection of the data. Figure 47A presents the distribution of molecular weights resulting from the processing of each segment. The median (IQR) for segment 1 was 78.87KDa (48.68 – 120.7KDa), for segment 2 was 83.25KDa (52.45 – 130.4KDa), for segment 3 was 55.74KDa (36.43 – 96.41KDa) and for segment 4 was 49.15KDa (23.74 – 95.55KDa). As presented in Figure 47B at 1% FDR, 725 proteins were identified in segment 1 from which 616 were quantified in all the samples. In segment 2, 678 proteins were identified and 572 fully quantified. In segment 3, 869 proteins were identified, 763 fully quantified; in segment 4, 717 proteins were identified, and 547 fully quantified. In segment 1 at 5% FDR 1778 proteins were identified and 1310 fully quantified. In segment 2, 1872 proteins were identified and 1357 fully quantified. In segment

3, 1441 proteins were identified and 1140 fully quantified. And in segment 4, 1555 proteins were identified and 912 fully quantified.

Considering that SEC is a preparative technique, there is an overlap of the different segments during the separation; therefore, it is expected to profile a subpopulation of shared proteins across segments. The degree at which the quantified proteins were shared across segments is indicated in the Figure 47C. Overall, the plasma proteome profiled in this work resulted in 4174 identified from which 3196 were fully quantified at 5%. Two hundred and thirty one proteins were shared among the four segments. Six hundred and fifty eight proteins were exclusively profiled in segment 1, 669 proteins in segment 2, 569 proteins in segment 3 and 424 in segment 4.

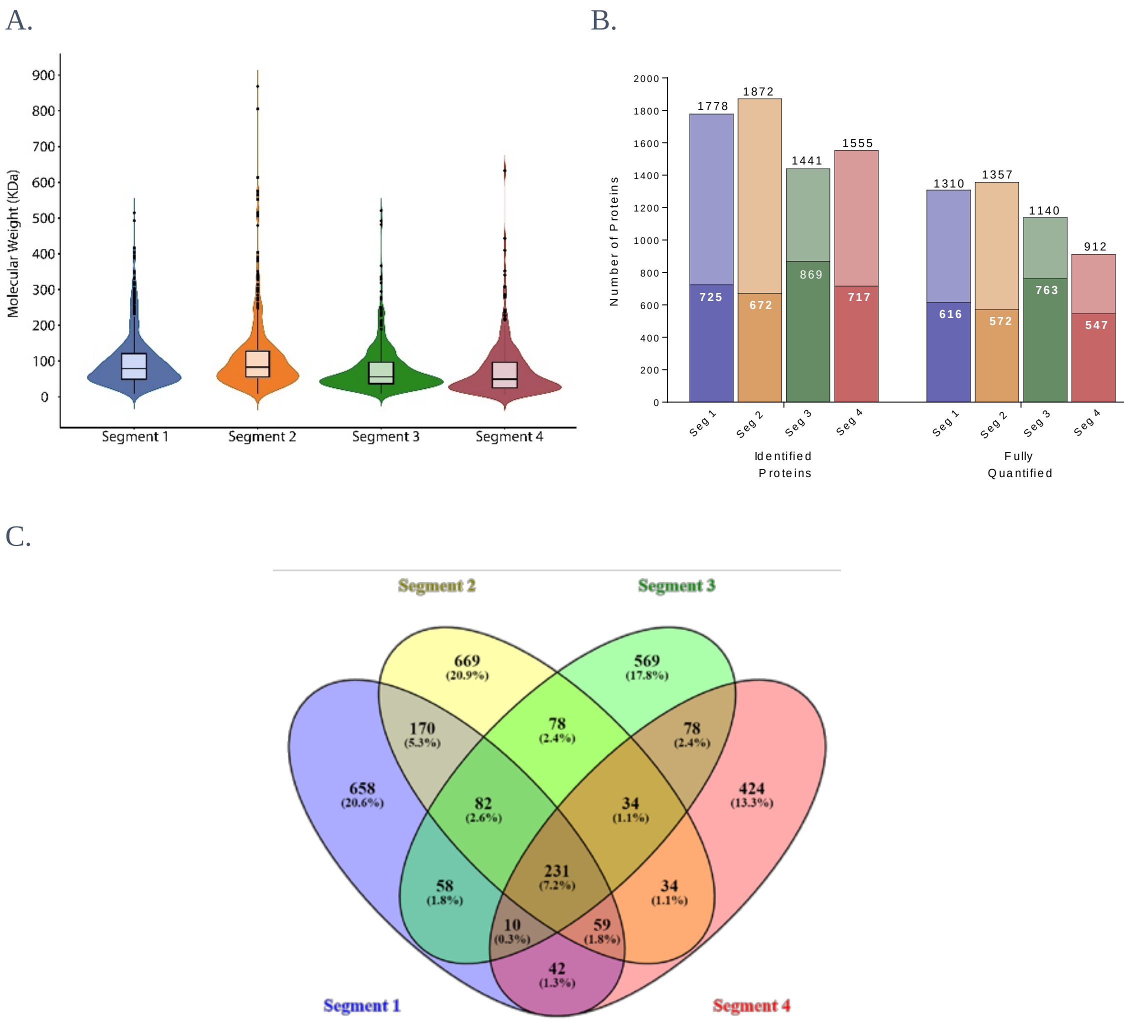


*Figure 47. Data exploration at protein level of plasma proteome*

***A.** Violin plots representing the distribution and density of the molecular weights of the proteins quantified in each segment at 1%FDR. Boxplots inside the violin plots indicates the median and interquartile range. Filled area represents the probability density plot. Molecular weights retrieved from UniProt. **B.** Number of identified and fully quantified proteins in segment 1 to 4. Dark colour indicates the number of proteins extracted at 1%FDR and light colour the additional proteins obtained when extracted at 5%FDR. **C.** Venn diagram of proteins fully quantified at 1%FDR in segments 1 to 4.*

#### 4.3.3.2 Multiconsensus report: an integrated evaluation of the plasma proteome

Taking in consideration that there is a considerable number of shared proteins among SEC segments and thus, peptides derived from a particular protein may be found across different segments, due to the intrinsic lateral diffusion of the protein analytes during their SEC separation process, an alternative approach can be applied in order to improve on statistics and power of analysis. The following section describes the analysis of the data extracted as a multiconsensus report. In this case, the data from the four experiments is merged and Proteome Discoverer 1.4 generates a single report.

Integrated analysis of proteomic data resulted in 5022 identified proteins from which 3577 were quantified in all the samples at 5%. On the other hand, 1876 proteins were identified and 1435 proteins quantified at 1%FDR as presented in Figure 48A. In comparison to independent analysis, the integrated evaluation of the data increased the coverage of the proteome in about 20%. Figure 48B presents the distribution of proteins depending on the number of peptides profiled per protein, 1095 proteins were profiled with 2 or more peptides at 1%FDR. The plasma proteome profile quantified in this work at high confidence exhibited a dynamic range of 11 eleven orders of magnitude from pg/mL to mg/mL as shown in Figure 48C. This ranged from proteins classically found in plasma such as albumin and immunoglobulins to signalling proteins such as CCL5 and MEGF8 (Multiple epidermal growth factor-like domains protein 8) and included enzymes and leakage products from tissues such as NAGPA (N-acetylglucosamine-1-phosphodiester alpha-N-acetylglucosaminidase) and TIMP2 (metalloproteinase inhibitor 2).  The illustrated linear dynamic range only approximately represents the true linear dynamic that was achieved given that many proteins relatively quantified have an unknown native concentration level.

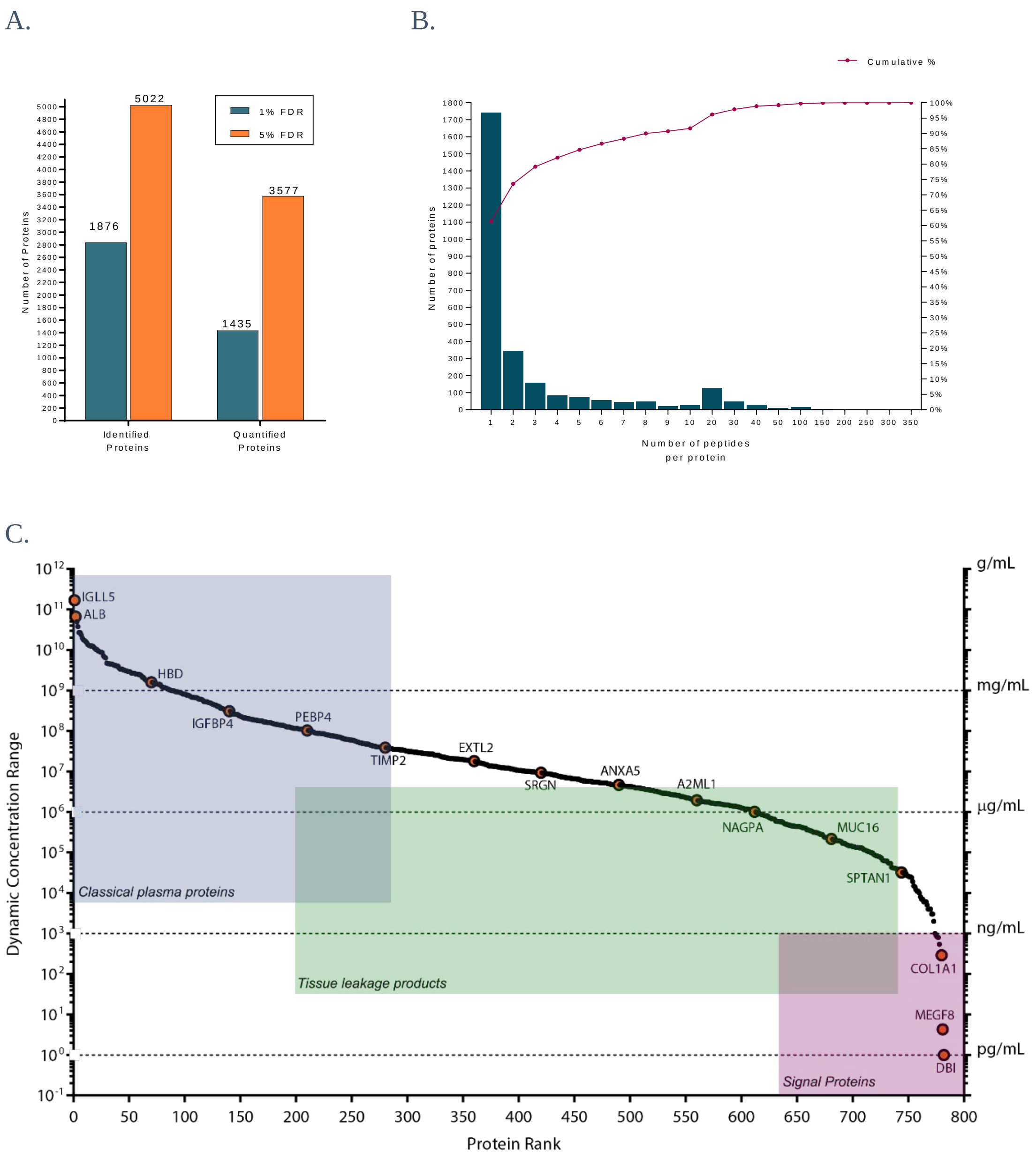


*Figure 48. Multiconsensus report: coverage and dynamic range of profiled proteins*

*Integrated analysis of segments 1 to 4 resulting from multiconsensus report. A. Total number of proteins identified and quantified at 1%FDR and 5%FDR. B. Distribution of the number of peptides profiled by protein at 1%FDR and C. Dynamic range of concentration for proteins fully quantified at 1%FDR. Highlighted in orange some proteins as illustration of the profiled proteins along the dynamic range. Proteins with reported concentration in plasma or serum are only included. Protein circulating abundances were retrieved from Homo sapiens plasma integrated database available from the protein abundance database PaxDb$^{4.1}$ (https://pax-db.org/species/9606) or from the plasma proteome database (http://www.plasmaproteomedatabase.org).*

### 4.3.3.3 Statistical assessment of proteomic data

In the section 4.2.4 the four statistical pipelines selected for this work are described: (P0) One-sample t-test, (PL1) Two-sample t-test, (PL2) Permutation and LIMMA and (PL3) LIMMA. A comparison based on the number of regulated proteins was conducted to select the most suitable pipeline for future analysis. Figure 49 demonstrates the number of regulated proteins resulting from each statistical pipeline. PL0 resulted in 360 significantly regulated proteins, PL1in 185 proteins, PL2in 114 proteins and PL3 in 240 proteins. For this first inspection, segments were independently analysed and only nominal *p* values were considered. Thirty six proteins were common to all pipelines.

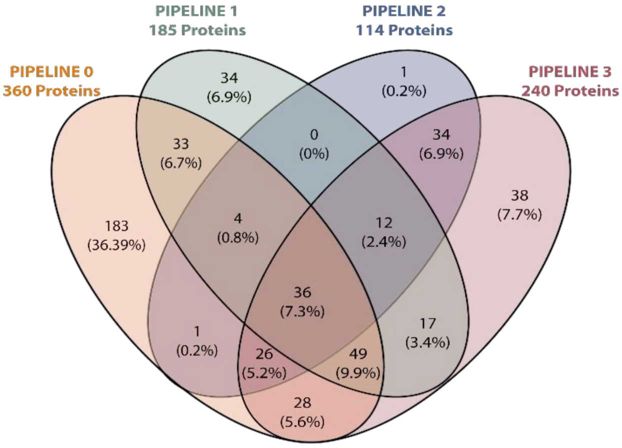


***Figure 49. Comparison of four statistical pipelines for assessment of significant protein regulation***

*Four statistical pipelines were compared for evaluation of significant size effect between control and tuberculosis group. (PL0) One-sample t-test, (PL1) Two-sample t-test, (PL2) Permutation and LIMMA and, (PL3) LIMMA. Only nominal p values included. Significant modulation considered when $p < 0.05$*

Missing values are frequently considered a common nuisance derived from relative quantification in proteomics and therefore partially quantified proteins are usually excluded from downstream analysis. However missing data could provide biologically relevant information, specially related to an on/off phenomena. In this work, additionally to the statistical testing for significant changes in protein abundance resulting from tuberculosis infection, the data was manually evaluated looking for patterns of missingness associated to the groups. Annexe 2 presents a summary of the proteins significantly regulated per segment including those with a pattern of missingness based on sample groups.

PL2 resulted in the strictest statistical approach and the fold-changes from shared proteins significantly regulated across various segments were compared to evaluate the variability of the relative quantification in the independent experiments. Relative quantification of 20 shared proteins among independent experiments presented reasonably reproducible values as shown in Figure 50.

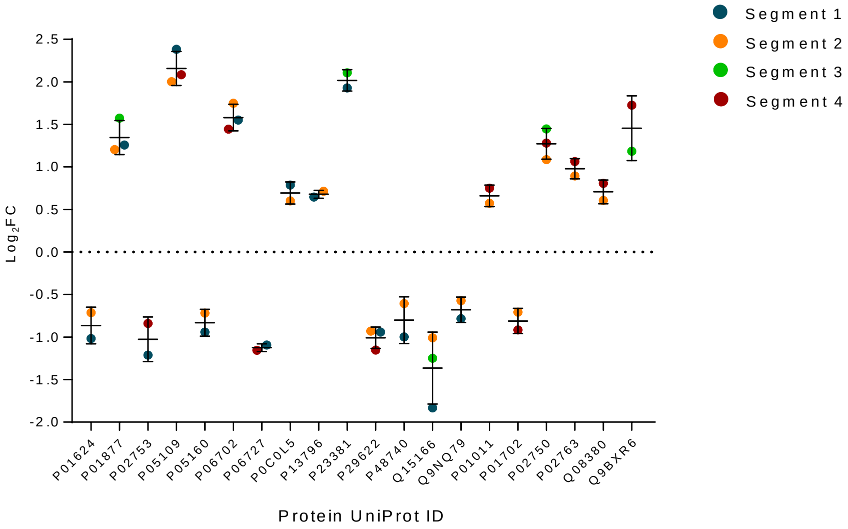


***Figure 50. Variability relative quantification of 20 shared proteins across independent experiments***

*Relative quantification of the 20 shared proteins significantly regulated across the four segments. Permutation and linear modelling (PL2) was used to assess the significance of the size effects. Only nominal p values considered ($p < 0.05$)*

Additionally relative quantification of proteins extracted at high confidence from the multiconsensus report were used to evaluate the modulation of proteins in plasma resulting from active tuberculosis infection. Pipelines 2 and 3 were applied on this dataset. As shown in Figure 51A, 32 proteins were significantly modulated from the PL2 and 157 from PL3. Twenty eight proteins were common between these two approaches. Considering that most of proteins from PL2 were included in the list derived from PL3, this later list of proteins was used for additional analysis. The multidimensional scaling or MDS plot presented in Figure 51B indicates that most of the variance of the data is explained by the group variable: healthy controls and tuberculosis patients. These results confirm the ability of this proteomic signature to distinguish active tuberculosis from healthy status. Protein expression pattern and clustering of both samples and proteins is shown in a heatmap in Figure 51C and further visualisation of the significantly regulated proteins is presented in the volcano plot in Figure 51D. Forty four proteins were significantly upregulated and 75 significantly downregulated.

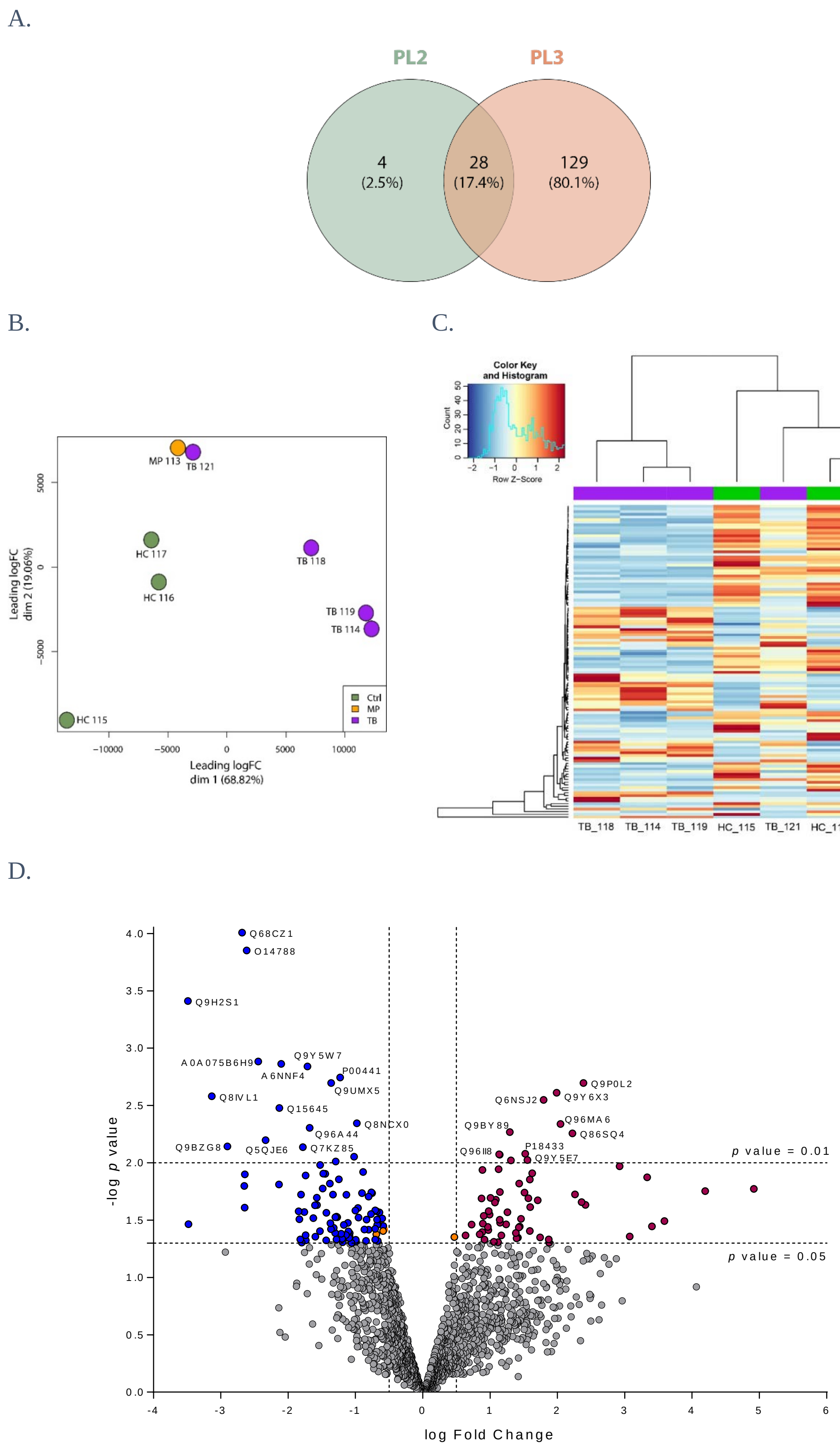


***Figure 51. Differential plasma proteome driven by active tuberculosis***

*Significantly modulated proteins were determined conducting pipeline 2 (Permutation and LIMMA) and pipeline 4 (Only LIMMA). A. Comparison of the number of significantly regulated proteins between the two statistical approaches including partially quantified data. PL2: pipeline 2 and PL3: pipeline 4. B.*

*Multidimensional scaling plot generated using the significant and fully quantified proteins derived from PL3 (119 proteins). Orange indicates master pool, purple active tuberculosis patients and green healthy donors. C. Heatmap with significant and fully quantified proteins derived from PL3. Heatmap was generated using the function heatmap.2 in R. Blue indicates significantly downregulated proteins and red significantly upregulated proteins. Dendograms showing hierarchical clustering of both protein expression (rows) and samples (columns). Purple indicates active tuberculosis samples and green healthy donors. D. Volcano plot derived from the analysis of data through PL3. Red indicates significantly upregulated protein, blue significantly downregulated proteins, grey no significant proteins and orange the four additional proteins resulting from PL2. Proteins with p value <0.01 annotated.*

#### 4.3.3.5 Functional analysis of the plasma proteome in active tuberculosis

Gene ontology analysis of the proteome generated indicates that over 43% of the proteins were annotated to extracellular space/region and structures expected for proteins from plasma. Over 128 proteins were annotated as exosome derived (14%). The remaining proportion of proteins was distributed in a wide range of cellular compartments including cytoplasm, lysosomes and plasma membrane as presented in Figure 52A.

Gene ontology network analysis enriched for biological process terms (Figure 52B) shows enrichment for acute inflammatory response, blood coagulation regulation of protein secretion, macromolecular complex remodelling, fibrinolysis, hydrogen peroxide catabolism and protein activation cascade including defence response to bacterium and antimicrobial humoral response. All of these terms are relevant to plasma proteome and response to tuberculosis infection.

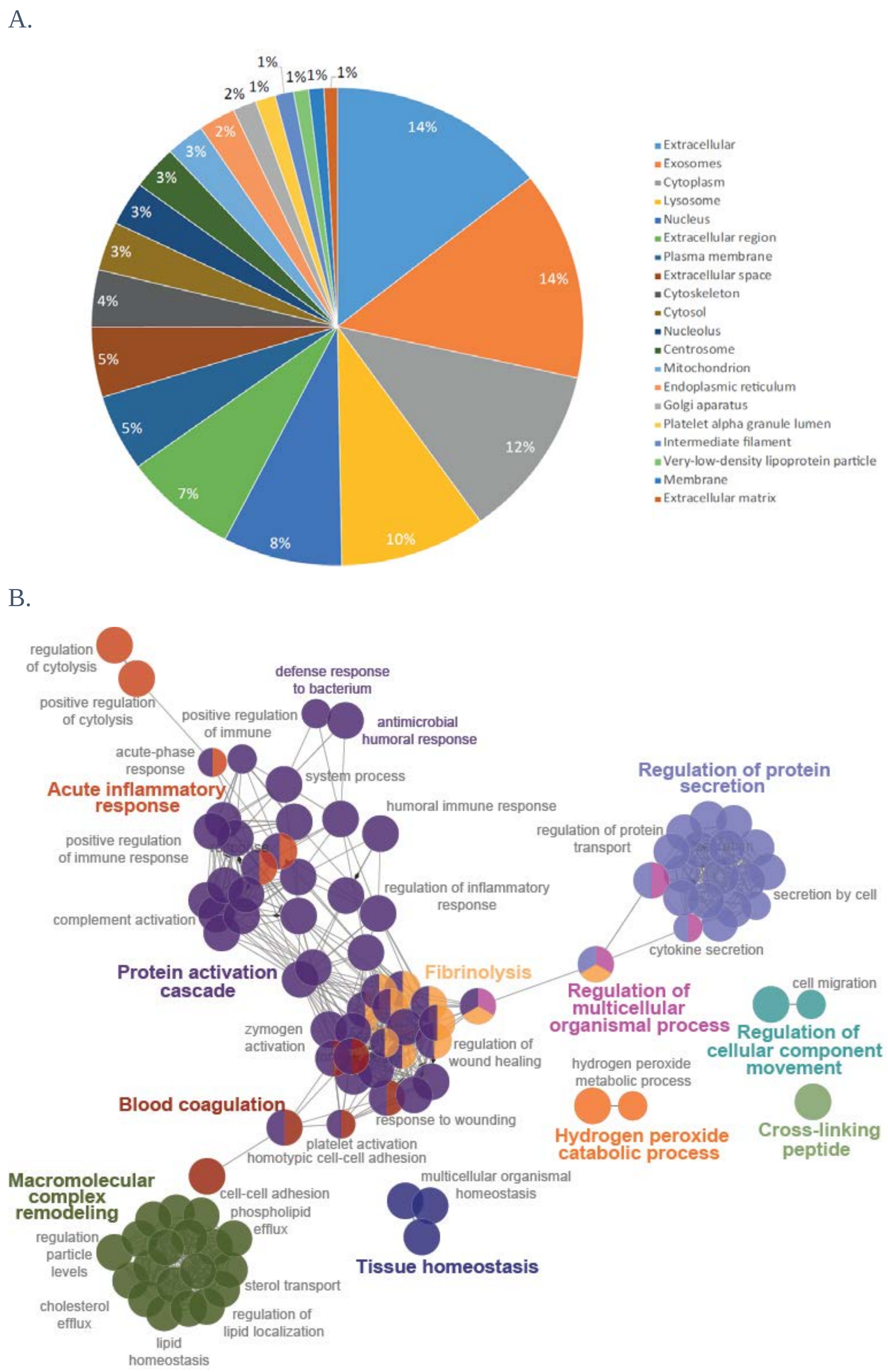


***Figure 52. Gene ontology enrichment of the plasma proteome in active tuberculosis***

*Gene ontology enrichment of the plasma proteome, A. Cellular compartment enrichment generated with FunRich 3.1.3 (222) B. Biological processes enrichment GO network generated with the app ClueGo in Cytoscape based on the database EBI-QuickGO-GOA (11.09.2017). Only annotations derived from experimental evidence were allowed and significant pathways with p value < 0.001 represented. For clarity purposes only some of the most relevant GO terms are presented. Size of nodes represent enrichment significance and edges kappa-statistic relations.*

An alternative network analysis based on the correlation of protein expression was used to find patterns of expression using Biolayout Express 3D. Two main clusters were produced with a Pearson correlation coefficient <0.85; the first one with proteins consistently downregulated and the second with proteins consistently upregulated in the tuberculosis group. Sample labelled as TB_121 again showed a different pattern to the tuberculosis group. Figure 53 shows the networks and the gene ontology enrichment associated to each one.

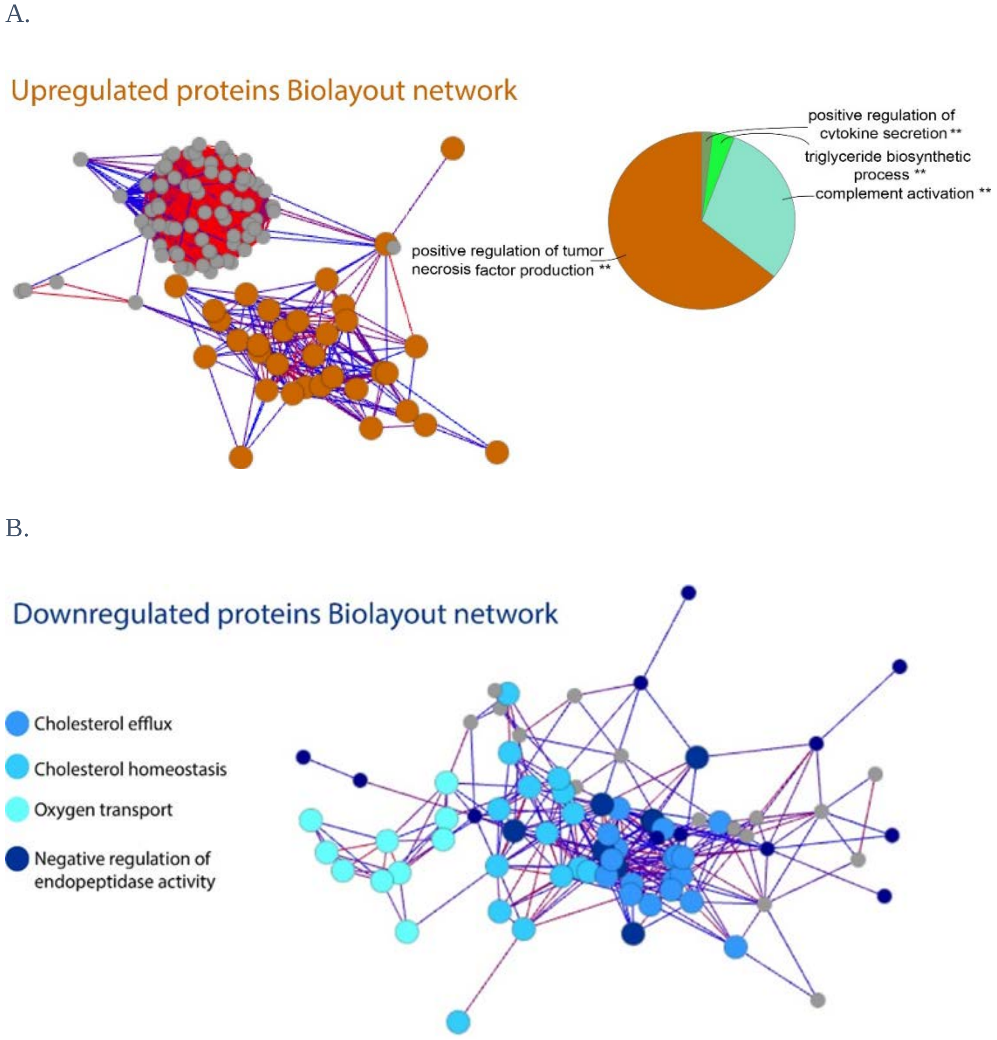

A.
Upregulated proteins Biolayout network
positive regulation of cytokine secretion **
triglyceride biosynthetic process **
complement activation **
positive regulation of tumor necrosis factor production **
B.
Downregulated proteins Biolayout network
● Cholesterol efflux
● Cholesterol homeostasis
● Oxygen transport
● Negative regulation of endopeptidase activity

***Figure 53. Correlation network analysis of plasma proteome associated to pulmonary tuberculosis***

*Co-expression networks generated using BioLayout Express 3D(215). Network of proteins with expression profiles correlated with a Pearson coefficient R>0.85. Nodes represent proteins and edges degree of co-expression. A. In orange a cluster of proteins upregulated in the tuberculosis group and B. In blue, clusters of proteins downregulated in the tuberculosis group. Additionally, main GO biological process term associated to each node.*

## 4.2 Discussion

The current state of biomarker discovery for diagnostics of tuberculosis has relied upon depletion methods, which has delivered limited coverage proteomes and partial biosignatures (36, 117, 119, 150, 151). This work presents the characterization of a comprehensive plasma proteomic profile for tuberculosis using an optimised MudPIT strategy that has the potential to achieve unbiased deep coverage profiling.

This particular MudPIT approach is based upon hyper-fractionation of the plasma samples using preparative and analytical chromatographic separations at both protein and peptide level. Figures 41 and 42 show high reproducibility of the chromatographic traces for both sample and master pool fractionations resulting from SEC generated over seven days. Consequently, variability of the traces across groups is mainly explained by the expected biological differences. Unexpectedly, the Peruvian group exhibited a particularly variable peak in segment 5 in comparison to the South African group (Figure 41). This peak contains mainly primary and secondary metabolites, mRNAs and other oligonucleotide species, along with small organic molecules; therefore, this fragment was excluded since the main scope for this project is limited to the protein content (35). Different additives used in the tubes during sample collection, although this is only a suggestion and will require further investigation, might explain the differences between Peruvian and South African samples in terms of segment five. The seven plasma samples selected for the discovery experiments were matched, such that groups did not exhibit significant differences of age, BMI or smoking status (Table 7), and therefore protein modulation can be mainly associated to the tuberculosis infection status.

Fractionation of iTRAQ-labelled peptide pools resulted in abundant traces for each segment (1-4) as presented in Figure 43. Additionally, early and late fractions containing underivatised labelling reagents, and other contaminants were subjected to SPE cleaning, using the protocol presented in the Figure 37. The second separation of the cleaned peptide fractions showed recovery of a significant number of peptides. Initially, the same chromatographic gradient was used for separation of both pooled peptides and SPE cleaned fractions as shown in Figure 43F. Considering that the SPE cleaned fractions represent a section of highly hydrophilic and hydrophobic peptides, the gradient was further optimised to increase the separation efficiency, and expedite separation. This gradient was used for segments 1, 2 and 4 (Figures 43B, D and H). The traces obtained with the new gradient were abundant and the separation was reduced by 50 minutes, representing an important methodological improvement.

Segments 1 to 4 were MS profiled and both raw peptide intensities and iTRAQ protein ratios were extracted form Protein Discoverer 1.4 (Thermo Scientific) using a FDR threshold of 1% and 5%, which implies high and moderate confidence at peptide level identification, respectively. Peptides were efficiently labelled in all cases (Figure 45) with missingness percentages below 30%. Labelling notably improved under optimised conditions when compared to the pilot study presented in previous chapter. Data missingness is not at random phenomenon in iTRAQ proteomics, it has been suggested that the probability that a protein is missing is related to its abundance, and therefore this feature of the data could further contribute to gain biological information (216). The analysis presented in this work explored proteins quantified in all the samples and missingness patterns as a complementary approach.

Extensive datasets from proteome profiling were generated from segment 1 to 4 at peptide level. Considering the most stringent conditions at 1% FDR, the dataset comprising only unique peptides ranged from 51622 to 26994 peptides for segment 1 to 4, respectively. These data sets were used to calculate protein expression. The total number of unique and fully quantified peptides, presented in Figure 46A, suggest an extensive proteome coverage supported as well by the total number of inferred proteins. As part of the data processing, median normalisation is performed in order to correct the differences in the total observed protein abundance across samples. Assuming that expression from most of the proteins in a biological system are not regulated, the median peptide intensities in each sample should be the same. If that is not the case, it may point out experimental bias such as quantification errors, which must be corrected. Medians derived from each dataset exhibited an acceptable variation, ranging from 6.95%CV in the segment 2 dataset to 12.07%CV in the segment 4 dataset, which in overall presented the lowest intensities of all datasets. The raw data was effectively normalised and adjusted as presented in Figure 46B-E.

Initial SEC prefractionation of the plasma is based on the hydrodynamic radius of the analytes that is most closely related to the molecular weight when the separation is conducted under denaturing conditions. The premise of this first step is the reduction of the complexity of the matrix by fractionating based on molecular size. The violin plots in Figure 47A suggest a tendency to smaller molecular sizes from segment 1 to 4 as expected. Relevant for plasma proteomics, abundant proteins such as albumin and immunoglobulins are concentrated in different fractions according to their sizes. For instance, albumin exhibits a molecular size of 69.37KDa and it is concentrated in segment 3, confirmed by the retention time of the BEH450 standard mix, which contains BSA. This feature of the method allows an unbiased and in-depth profiling of the plasma proteome, since the reduction of the complexity of the matrix is achieved avoiding the depletion of the biological sample.

Considering that SEC is a preparative technique, protein separation into segments is completed only partially. Lateral diffusion during the chromatographic separation results in a shared subpopulation of proteins across segments. Nevertheless, each segment represents a sub-proteome and these can be mined independently depending on the scope of the experiment. In this chapter, a comprehensive analysis of the plasma was intended; therefore, a single list of proteins was derived according to the distribution of proteins presented in Figure 47C. This list comprises 5022 identified proteins from which 3577 were fully quantified, representing the most comprehensive plasma proteome to date in the tuberculosis field. In contrast previous proteomic studies based on depletion have yield limited plasma proteomes. For instance, in one of the most recent plasma profiles for tuberculosis Chen, C. *et al.*, (2018) profiled 716 proteins using iTRAQ proteomics (153).

Additionally, a particular feature included in Proteome Discoverer 1.4 allows generating a single report that includes multiple reports from independent experiments. Different reports are treated as biological replicates and merged in a multiconsensus report. Additionally, more unique peptides per protein group can be captured that exhibit the same trend in differential expression across multiple segments. Consequently, this approach increases peptide and protein statistical power since evidence found of a particular protein from different segments is merged in one single protein report. This strategy increased in 17% the identified proteins and 10% the quantified proteins.

Optimisation and development of this MudPIT approach led to a notable coverage increase of the plasma proteome compared to similar studies, and plasma proteins spanning a minimum of 11 orders of magnitude were confidently quantified (Figure 48). As previously stated, the actual dynamic range achieved is most likely substantially larger as many of the relatively quantified proteins have an unknown native plasma concentration level. By contrast, Xu, D., *et al*. (2015) only reported 434 plasma proteins quantified using iTRAQ (95% confidence) (119) and Wang, C., *et al.*, (2016) only quantified 160 serum proteins using iTRAQ (95% confidence) (223). Given that these were mostly generic and of high native abundance levels, their linear dynamic range was around 6-orders of magnitude or less. Deeper coverages increases the opportunities to discover novel biomarkers with clinical relevance (133).

Isobaric labelling offers multiple advantages such as multiplexing of samples, which reduces the technical variability during MS analysis which benefits high-throughput quantification (186). The possibility of multiplexing samples within one experiment eliminates the need to compare multiple LC-MS/MS runs, therefore reducing the total analytical time and minimising run-to-run variation. An additional benefit is that iTRAQ/TMT exhibits a wide dynamic range, it can be utilised to profile high and low-abundance proteins as demonstrated in this work (224). However, assessing proteomic data for significant changes of protein abundance is a central task that faces multiple challenges, for example, small sample sizes result in large uncertainty of the variability estimates which impairs t-test statistics. There is not a clear consensus in the field regarding the best approach to determine significant size effects from data generated by iTRAQ/TMT proteomics. Usually the best approach is selected from the evaluation of the data structure, sources of variability and experimental design features.

FDR estimation at the peptide level has a significant impact on the number of proteins identified. Since this analysis is mainly orientated to the selection of candidates for further validation, only high confidence identifications were selected for downstream analysis. In this work, four different reported approaches for analysis were tested. (*PL0*) One-sample t-test, (*PL1*) Two-sample t-test, (*PL2*) Permutation and LIMMA and (*PL3*) LIMMA. According to the total number of significantly modulated proteins pipelines were characterised from less to most strict; *PL0 – PL3 – PL1 – PL2*.

t-Test based statistics presents reduced power when only small sample sizes are available and is limited to the assumption of distribution normality (225). This is usually the case for proteomic discovery experiments. Particularly, one-sample t-test (*PL0*) was initially considered since the main output from Proteome Discovery at protein level is restricted to ratiometric expression. However, this approach results in a large number of false positives since the null hypothesis is tested at the fold-change level and the variability within groups is ignored. Calculation of protein expression based on normalised and log2-transformed peptide intensities allows the group comparison including individual sample variation. Two-sample t-test is a more adequate test than one-sample t-test for processing of these datasets. However, assumptions of normality and equality of variance limit its power for proteomic application. Recently, linear models such the LIMMA (Linear Models for MicroArray data) package component of Bioconductor, an R-based open-source software project, have expanded to proteomics from large-scale gene expression data. LIMMA has proved be particularly powerful with small sample numbers by using the full dataset to shrink observed sample variances towards an estimate allowing for variance distribution (225, 226). This empirical Bayes approach results in a more realistic distribution of biological variances compared to other methods (226). Permutation was introduced prior to LIMMA in *PL2*, to filter-out less stable ratios, reducing the number of false positives. This approach has proved to reduce false positives in microarray data processing for microarray data (227). In *PL3* permutation was omitted to reduce stringency of the analysis. Multiple testing correction was applied (FDR) to all the pipelines, however nominal *p* values were mostly considered since this correction reduced dramatically the number of significantly regulated proteins. Additionally, the top hits were subjected to validation in independent cohorts.

Since most of the proteins significantly modulated calculated from *PL2* are included in the list derived from *PL3* (Figure 51A), this last list was used for downstream analysis. On hundred and fifty-seven proteins were determined as significantly regulated from *PL3* and the 117 fully quantified proteins from that group were used to discriminate tuberculosis patients from healthy controls as presented in MDS plot (Figure 51B). Sample 121 is separated from the tuberculosis groups and closer to the master pool, which correlates with the clinical data available from this patient. This individual presented with a normal CRP and only minor inflammation on lung radiography, indicating less advanced disease than the other patients. This therefore shows the sensitivity of the proteomic signature to stratify patients within the tuberculosis group. However, this observed biological heterogeneity highlights the necessity to increase the number of biological replicates in order to study the nature of variation and increase the confidence on the biological conclusions drawn from this study.

Clustering of samples based on the relative expression levels of these 117 proteins is dictated mainly by the cases and controls grouping. The heatmap in Figure 51C presents the pattern of expression and consistently with the MDS plot sample 121 is clustered with the control group. Consistent patterns of up- and down-regulation resulting from the tuberculosis infection are distinguishable, including sets of proteins highly upregulated and downregulated, as well as moderately regulated proteins. Proteins with the most consistent regulation and stronger statistics were selected for further validation. Validation of potential candidates will be discussed in Chapter 5. The list of 117 proteins is presented in the Annexe 3.

Gene ontology enrichment analysis of quantified proteins at cellular component level was consistent with those predicted for plasma samples as presented in Figure 52A. Network analysis of biological process GO enrichment at high confidence ($p < 0.001$) indicates activation of the immune system at innate and adaptive levels with nodes such as complement activation, protein activation cascade and acute-phase response (Figure 52B). Related to these biological processes, other nodes were represented such as defence response to bacterium and antimicrobial humoral response, fibrinolysis, cytokine secretion and tissue homeostasis. These results are consistent with the acute response to tuberculosis infection and clinical manifestation of the active disease.

A complementary strategy to explore the proteomic data generated in this work was employed. Calculation of *p* value presents increasingly recognised shortcomings (228-230) particularly when testing hundreds of classifiers. The complexity of the biology systems encoded in large datasets requires alternative approaches to complement statistics based on *p* values and "*null* hypothesis testing". Biolayout Express (3D) (215) was used to evaluate the proteins fully quantified at high confidence from the multiconsensus report. Network analysis clusters proteins based on coefficient of correlation, which results in patterns of expression that provides a biological relevant insight of the data. Analysis based on patterns of co-expression were consistent with tuberculosis immunopathogenesis as shown in Figure 52C. Proteins downregulated were mainly associated to cholesterol homeostasis and negative regulation of endopeptidase activity. These biological processes are consistent with the altered systemic state induced by tuberculosis pathogenesis resulting in clinical manifestations such as cachexia (231). Upregulated proteins were related with specific processes such as positive regulation of TNF production, complement activation and secretion of cytokines.

The small sample size constitutes an important limitation of this profiled proteome, and consequently the biological insights that can be derived with confidence from this analysis are limited by a high false discovery rate. However, this proteome of active pulmonary tuberculosis represents the most comprehensive plasma proteome described so far for this global health threat. In this chapter the capabilities of this MudPIT approach to generate extensive and unbiased proteomes has been demonstrated. Additionally, refined statistical and bioinformatic approaches were developed in order to mine the large datasets produced from this approach. Increased statistical power is required to harness the potential of this approach to reveal new biomarkers and generate novel knowledge about tuberculosis immunopathology. The next chapter will present increased biological replicates to increase statistical power and validation on independent cohorts of selected candidates.

# CHAPTER 5

# Comprehensive plasma proteomic profiling reveals novel diagnostic biomarkers for active tuberculosis

## 5.1 Introduction

Chapter 4 presented an optimised method for an extensive profiling of the entire plasma proteome based on hyper-fraction, which circumvents the depletion of the matrix. This methodological approach offers a more comprehensive representation of the human plasma proteome ranging from classical plasma proteins at high circulating levels to signalling proteins expressed at pg/mL concentrations. The in-depth proteomic profiling, as described, constitutes a low-throughput and manually intensive process requiring extensive instrument time. This also translates to increased analysis costs. These factors limited the number of samples examined in the discovery phase, and thus curbing the statistical power.

Considering the potential of this method to reveal new biomarkers suitable for point-of-care tests and its importance to improve our current diagnostic tools urgently required for tackling the transmission of tuberculosis, this chapter presents a deeper profile of one of the segments. Segment 4 was selected for a detailed proteomic analysis by increasing the sample size, analysing 10 control and 11 TB patients. This particular segment is enriched for small proteins ranging from 5KDa to over 600KDa. Most of profiled proteins were distributed around 50KDa, which is below the immunoglobulins and albumin molecular weights (Figure 47A). This suggests that the concentration of most of these highly abundant proteins is significantly reduced favouring the profiling of a more diverse subproteome. Independent statistical analysis additionally indicated that segment 4 profiling led to the largest number of modulated proteins compared to the other segments (Annexe 2). Additionally, since pulmonary tuberculosis involves destructive immunopathology of the lung, some degradation products might be captured as well in this segment.

After the in depth analysis of segment 4, this chapter describes the rationale behind the candidate prioritisation, validation of selected proteins using immunoassays in two independent cohorts. These readouts validate the proteomic findings. In addition, I undertook further bioinformatic analysis of this more robust proteomic dataset for biological interpretation.

## 5.2 Methods

### 5.2.1 Patient Cohort

Recruitment and ethics for the Peruvian and the South African participants are described in the section 2.2. Additionally to the data previously produced, two additional iTRAQ sets were processed to increase the sample size for segment 4 and generate a detailed analysis of this particular segment. As presented in the general description of iTRAQ experiments in Table 4 seven additional samples from heathy donors were included in the study and 8 samples for the tuberculosis group. Table 9 presents the clinical data for the overall cohort of samples used to analyse segment 4, including the sample set analysed in the previous chapter. Healthy control individuals included in this study group presented a mean age ± SD of 28.3±4.30 (range 22-35 years) and BMI±SD of 23.50 ± 1.29. In the case of active pulmonary TB patients, age ± SD was 31.7±7.86 (range 21-44 years) and BMI±SD was 20.7 ± 2.22. The participant individuals were age matched with no significant difference between groups ($p$ < 0.05). However, nutritional status in patients with active tuberculosis is often compromised. Pulmonary tuberculosis may lead to reduction in appetite, nutrient/micronutrient malabsorption, and altered metabolism leading to wasting (232). The patients included in this cohort of samples reported significantly lower BMI to the control group. Samples were collected prior treatment initiation and there was not significant differences in the smoking status between groups.

*Table 9. Clinical information of individuals participating in the proteome profiling study*

| *Variables* | *Healthy Controls* | *Pulmonary Tuberculosis* | *p Value* |
|---|---|---|---|
| n | 10 | 11 | |
| Gender | Male (100%) | Male (100%) | |
| Mean age ± SD (years) | 28.3±4.30 | 31.7±7.86 | 0.229$^{(a)}$ |
| Age range (years) | 22-35 | 21-44 | |
| Mean BMI ± SD | 23.50 ± 1.29 | 20.7 ± 2.22 | 0.002$^{(a)}$ |
| Smoking History | | | |
| • Non-smokers | 6 | 2 | 0.07$^{(b)}$ |
| • Current smokers | 3 | 3 | |
| • Ex-smokers | 1 | 6 | |
| Drug Treatment | | | |
| • None | 10 | 11 | |

$^{(a)}$ two-tailed $p$-value calculated by t-test

$^{(b)}$ two-tailed $p$-value calculated by Fischer's exact test

### 5.2.2 Experimental Design

The plasma proteomic profile associated to pulmonary tuberculosis infection was evaluated in depth using the segment 4 generated by the SEC fractionation of plasma samples. As described in the previous chapter, this work included participant samples from South Africa and Peru and a master pool used for controlling variability across experiments. Figure 54 illustrates allocation of plasma samples for the three iTRAQ 8-plex experiments by block-randomisation. Samples were randomised and the tag 113 was assigned to the master pool. Set A refers to the group of samples processed as part of the whole plasma profiling described in the previous chapter. Sets B and C were run one year after that initial analysis, but using the same protocol. Only some modifications at the MS analysis were applied (details below).

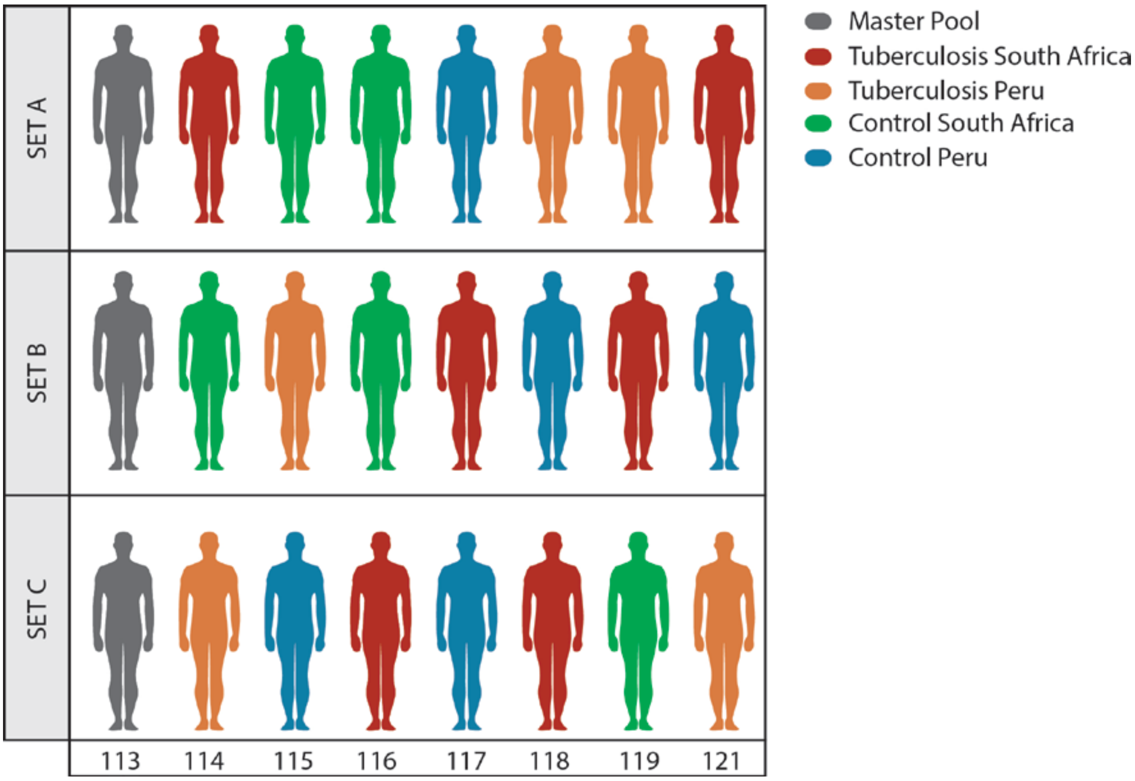


***Figure 54. Experimental design for depth profiling of plasma segment 4 in pulmonary tuberculosis context***

*Plasma samples allocation for three iTRAQ 8-plex. Control and tuberculosis samples from Peruvian and South African ethnicities were randomised using the tool available in https://www.random.org/. The tag 113 was assigned to the master pool for variability control inter-8-plex experiments*

### 5.2.3 Sample processing and data analysis

SEC segments were produced as described in section 4.2.3 and the chromatographic traces are shown in Figure 41. Segments were generated and stored at -20°C until further processing. All the procedures were performed as described in section 4.2.3. Briefly, segment 4 from the selected samples were dialysed and quantified. 120µg of protein was trypsin digested and iTRAQ labelled according to the design presented in the Figure 54. Peptides were analytically reconstituted, pooled together and fractionated using offline C4 – HPLC. Gracepure SPE C18-AQ cartridges were used to clean selected fractions. Offline fractions were further separated using an AcclaimPepMap RSLC, 75µm× 25cm, nanoViper, C18, 2µm particle column retrofitted to a PicoTip emitter (FS360-20-10-D-20-C7) and analysed by the high resolution nano-ESI-LTQ-Velos Pro Orbitrap-Elite mass spectrometer (Thermo Scientific). Specifications for the mass spectrometric analysis are presented in section 2.5.8, particularly for these experiments the scans were acquired at a higher resolution (60000 for CID) than the previous experiments (30000 in Chapters 3 and 4) and 15000 for HCD.

Resolution was increased since Data analysis was conducted as described in section 4.2.4 by pipelines *PL2* and *PL3*.

Further bioinformatic mining of the data was conducted to interrogate biologically relevant information captured by the more comprehensive plasma proteomic approach conducted. Initially, the three datasets generated for segment 4 were merged and subsequently, a new dataset containing the common proteins was corrected for batch effect. Two approaches were tested for batch effect correction, normalisation of the relative protein expression levels to the master pool and ComBat. ComBat is a software tool originally developed to adjust batch effects in microarray data using empirical bayes methods (233) but has been widely extended to multiple applications. Variance distribution of the data was inspected using PCA and the method providing the best effect on the data was selected for downstream analysis.

Thus, weighted correlation network analysis or WGCNA (234) was applied to define protein co-expression trends across control and pulmonary tuberculosis groups. WGCNA 1.63 was run in RStudio available from CRAN (https://cran.r-project.org/web/packages/WGCNA/index.html). Networks of highly interconnected proteins were constructed using a soft-thresholding power = 0.9 and modules were identified using a minimum module size of 15. Module significance was calculated as a measurement of the correlation between biological traits, such as disease or group, ethnicity and smoking status and the protein expression profiles. Various visualisation tools available from the package were used to identify modules strongly correlated to biologically relevant covariates.

### 5.2.4 Validation of prioritised candidates

Serum and/or plasma samples collected from two different cross-sectional studies were included in this work for validation of the proteins that were prioritised as potential candidates to pulmonary tuberculosis biomarkers. The first cohort comprised 196 plasma samples from healthy volunteers, patients with respiratory symptoms requiring clinical assessment or patients recently diagnosed with pulmonary tuberculosis recruited at the Ubuntu HIV/TB clinic and GF Jooste Hospital, Cape Town, South Africa. These samples are part of a larger study described previously by Walker, N. *et al.*, (2017) (211, 235)  and included patients co-infected with HIV. This study was approved by the University of Cape Town Human Research Ethics Committee (REF 516/2011). Sputum acid-fast bacilli culture, Xpert MTB/RIF and chest radiographic evidence confirmed pulmonary tuberculosis diagnosis. Table 10 presents the general demographics for the cohort used in this work for validation, which included individuals with negative (115 samples) and positive (81 samples) co-infection with HIV.

*Table 10. Validation cohort from South Africa*

| Variables | Healthy Control | Pulmonary Tuberculosis | Respiratory Symptomatic |
|---|---|---|---|
| n (total) | 70 | 90 | 36 |
| *HIV uninfected coinfection* | | | |
| n | 54 | 38 | 23 |
| Gender | Female (22%) | Female (34%) | Female (22%) |
| Mean age ± SD (years) | 28 ± 7.8 | 37 ± 11.2 | 29 ± 9.1 |
| Age range (years) | 16-51 | 24-73 | 20-50 |
| Mean BMI ± SD | 22.9 ± 6.42 | 21.8 ± 4.09 | 21.9 ± 8.69 |
| BMI range | 17.46-52.45 | 17.63-33.20 | 18.76-55.65 |
| *HIV coinfection* | | | |
| n | 16 | 52 | 13 |
| Gender | Female (68%) | Female (46%) | Female (53%) |
| Mean age ± SD (years) | 32.5 ± 4.6 | 32 ± 7.1 | 32 ± 9.3 |
| Age range (years) | 26-42 | 23-56 | 19-58 |
| Mean BMI ± SD | 26.9 ± 6.14 | 23 ± 3.77 | 25.8 ± 5.51 |
| BMI range | 17.9-38.39 | 17.44-37.08 | 17.14-35.11 |

A second independent cohort was included for validation of candidates in this work comprising 124 samples from a cross-sectional study conducted in the United Kingdom (https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/mimic/). This study was sponsored by the University Hospital Southampton NHS Foundation Trust and approved by the National Research Ethics Service Committee South Central - Southampton A (ref 13 SC 0043). Table 11 presents the demographics for this validation cohort. Sputum acid-fast bacilli culture, Xpert MTB/RIF and chest radiographic evidence confirmed pulmonary tuberculosis diagnosis and QuantiFERON positivity confirmed tuberculosis latent infection. Both validation cohorts included patients attending the clinics with symptoms suggestive of pulmonary tuberculosis finally diagnosed with a different respiratory infection. This group is referred as respiratory symptomatic.

*Table 11. Validation cohort from UK (MIMIC Study)*

| Variables | Healthy Control | Latent Tuberculosis | Respiratory Symptomatic | Pulmonary Tuberculosis | Extrapulmonary Tuberculosis |
|---|---|---|---|---|---|
| *Negative HIV coinfection* | | | | | |
| n | 30 | 30 | 26 | 32 | 6 |
| Gender | Female (67%) | Female (73%) | Female (42%) | Female (56%) | Female (50%) |
| Mean age ± SD (years) | 43 ± 16 | 39 ± 18 | 57 ± 12 | 40 ± 15 | 34 ± 8 |
| Age range (years) | 20 - 69 | 19 - 80 | 30 - 78 | 20 - 72 | 25 - 45 |
| Mean BMI* ± SD | 27.3 ± 6.9 | 26.52 ± 5.1 | 27.56 ± 6.2 | 25.72 ± 4.8 | 20.63 ± 2.1 |
| BMI range* | 15 - 46 | 19 - 40 | 21.6 - 39 | 20 - 35.8 | 17.8 - 23 |

* Only partial data available

Details for quantification of these proteins on the validation cohort by ELISA or Luminex are described in Section 2.7.

Performance of the validated candidates was in first instance assessed by calculating receiver operating curves (ROC) for individual proteins and combined proteins in each validation cohort. The statistical package IBM SPSS Statistics 25 was used for this purpose. ROC analysis was conducted by setting pulmonary tuberculosis as a positive test and binary logistic regression probabilities were calculated when analysis of combined markers was performed. Coordinates of the curves was exported to estimate potential cut-off values.

## 5.3 Results

### 5.3.1 Peptide Fractionation of Segment Four by C4 HPLC

The chromatographic traces resulting from the C4 fractionation of SEC segment 4 derived from the samples selected for sets B and C and SPE cleaned peptide fractions are shown in Figure 55. Fractionation of samples from set B resulted in 98 offline fractions and from set C in 121 fractions (Figure 55A and C, respectively). SPE cleaning of the samples recovered additional peptides as presented in the traces in Figure 55B and D.

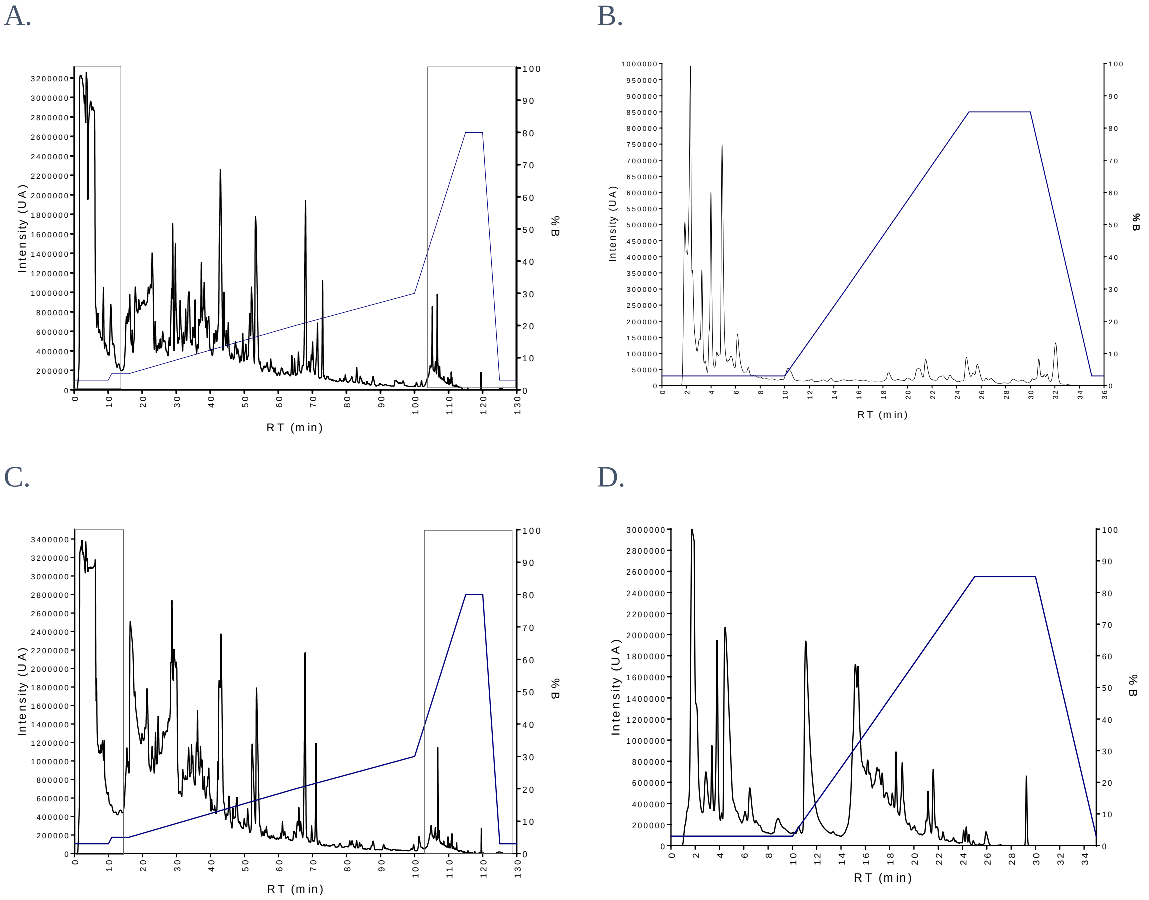


***Figure 55. C4 HPLC chromatographic traces of iTRAQ labelled peptides – Segment 4***

*C4 chromatographic traces of pooled iTRAQ labelled peptides were performed at 0.3min/mL and 35°C. Intensity was evaluated at 215nm. Blue line indicates elution gradient.* ***A.*** *Chromatogram of segment 4 set B.*

*B. Chromatogram of SPE cleaned fractions from segment 4 set B.. C. Chromatogram of segment 4 set C. D. Chromatogram of SPE cleaned fractions from segment 4 set C.*

### 5.3.2 In-Depth Analysis of Plasma Subproteome of Pulmonary Tuberculosis Infection

A general description of the data generated from segment 4 in sets A, B and C are presented in Figure 56A and indicate that all the samples were successfully labelled with a percentage of quantified peptides ranging from 71% to 93.58%. Figure 56B depicts the number of peptides identified in each one of the sets, indicating unique and fully quantified peptides. This group of unique and fully quantified peptides (Set A = 14152 peptides, Set B = 29868 peptides and Set C = 29855, 1%FDR) was used for the determination of the protein expression. Figure 56 C presents the total number of identified and fully quantified proteins. At 1%FDR, 733 proteins were identified and 527 fully quantified in set A, 1228 proteins were identified and 1062 fully quantified in set B and 1248 proteins were identified and 1009 fully quantified in set C.

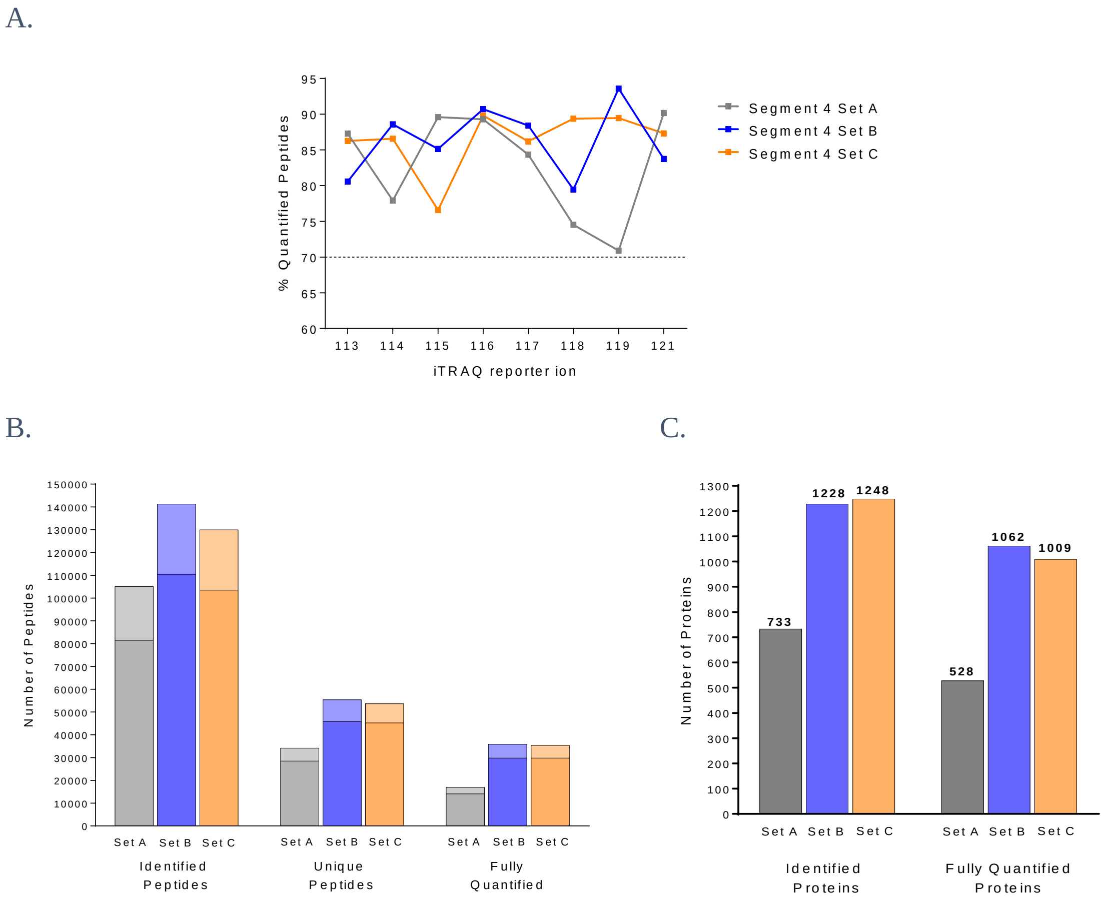


***Figure 56. General metrics for the plasma profile resulting from segment 4***

*General description of the data generated for set A to C from segment 4. A. Percentage of labelled peptides. B. Metrics for number of identified, unique and fully quantified peptides. Dark colour 1%FDR and in lighter colour additional peptides profiled at 5%FDR. C. Number of identified and quantified proteins in each set at 1%FDR.*

The three datasets generated from each iTRAQ set were inspected to evaluate batch effects and distribution of the data. The variance of the data is mainly explained by the batch (dimension 1), however the group effect is distinguishable when considering dimensions 2 and 3 (~15% variance), as illustrated in the MDS plots in Figure 57A. Additionally, Figure 57B presents a heatmap generated with the common proteins among the three sets (426 proteins)

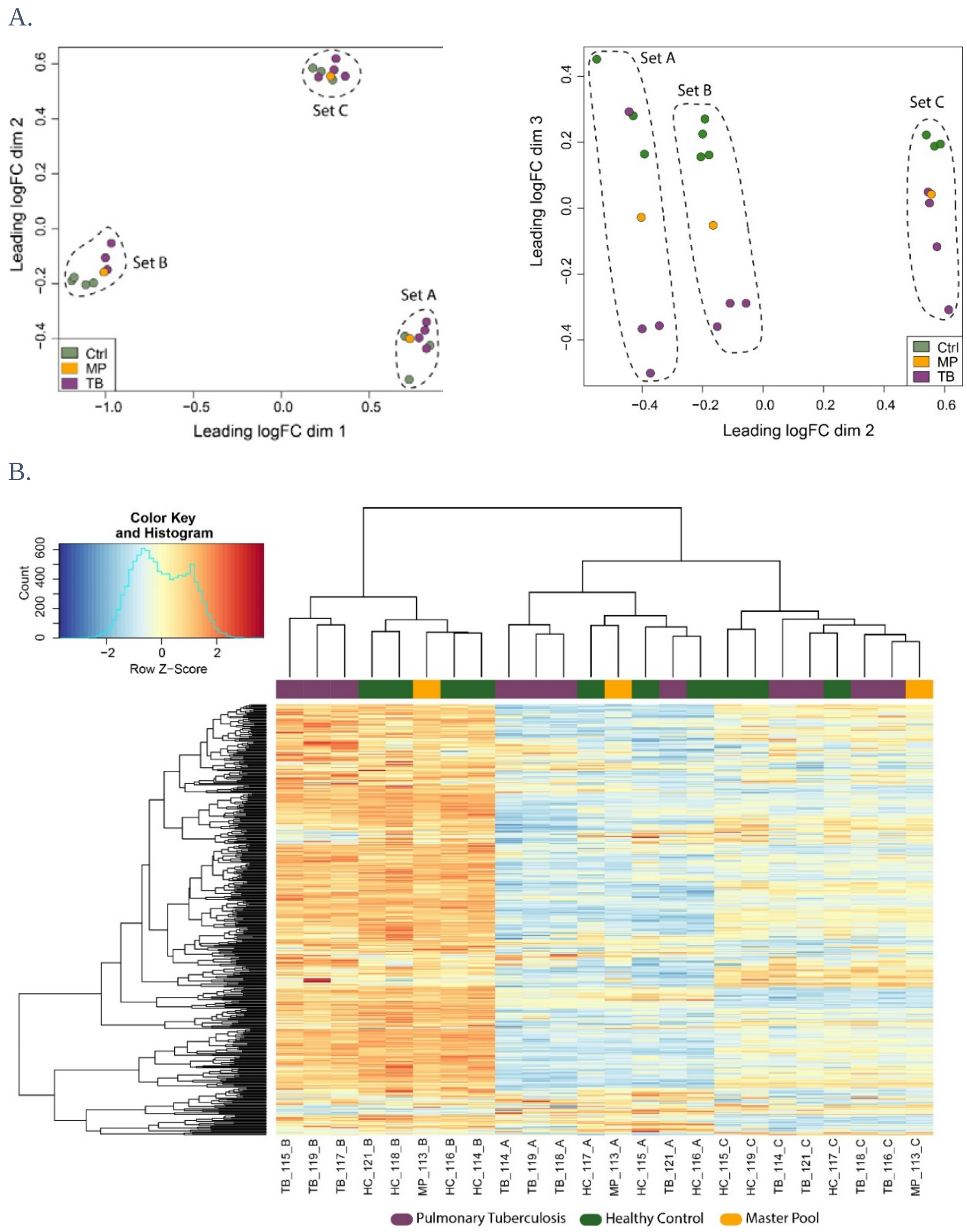


***Figure 57. Data distribution and clustering from sets A to C***

*Distribution of the data generated from sets A, B and C. A. MDS plots and variance distribution. In purple tuberculosis, green controls and orange master pool. Histogram for the variance distribution is presented as well. B. Heatmap with fully quantified proteins derived from PL3. Heatmap was generated using the function heatmap.2 in R. Dendograms showing hierarchical clustering of both protein expression (rows) and samples (columns). Green indicates healthy controls individuals, purple tuberculosis patients and orange master pool.*

### 5.3.3 Selection of Candidates for Validation

The primary goal of my PhD was to identify novel diagnostic markers for tuberculosis, and therefore my first aim in analysing the dataset was to identify candidates with diagnostic potential.  A small set of proteins were selected for validation on larger independent cohorts. Validation is a challenging step of the biomarker pipeline that implies translation of platform from mass spectrometric based-discovery to antibody-based quantification (ELISA/Luminex) and a significant investment of resources. Therefore, the selection of biomarkers was based on the best statistical evidence. The dataset of fully quantified proteins extracted at 1%FDR resulting from the analysis of segment 4 sets A to C were merged. Four hundred and twenty six proteins were common among the three sets and the statistical assessment of groups was based on this set of proteins (Figure 58A).

Pipelines *PL2* and *PL3* were applied for calculation of *p* values (significance considered when $p$ <0.05) and alternatively, analysis based on the correlation patterns of expression (Pearson correlation score R=0.75) using Biolayout 3D Express was utilised as well. *PL2* and *PL3* were conducted including experiment as a covariant, and only proteins whose regulation is explained by the group, controls vs. tuberculosis patients were included. The Venn diagram in Figure 58B indicates that all the proteins considered significantly modulated resulting from *PL2* are included in the group of significantly regulated proteins determined by *PL3*, additionally 36 out of 38 proteins defined by Biolayout were included to the list of proteins from *PL3*. Figure 58C presents a volcano plot with the proteins significantly upregulated (red) and downregulated (blue) determined by *PL3*. Proteins with *p* values FDR adjusted are indicated with a thicker outline. Fifty-eight proteins were significantly upregulated and 116 downregulated. Alternatively, analysis based on the co-expression patterns (R=0.75) led to two main clusters defined by status group. Cluster 1 contains 23 proteins that were significantly upregulated and Cluster 3 includes 14 proteins downregulated in the tuberculosis group as shown in Figure 58D.

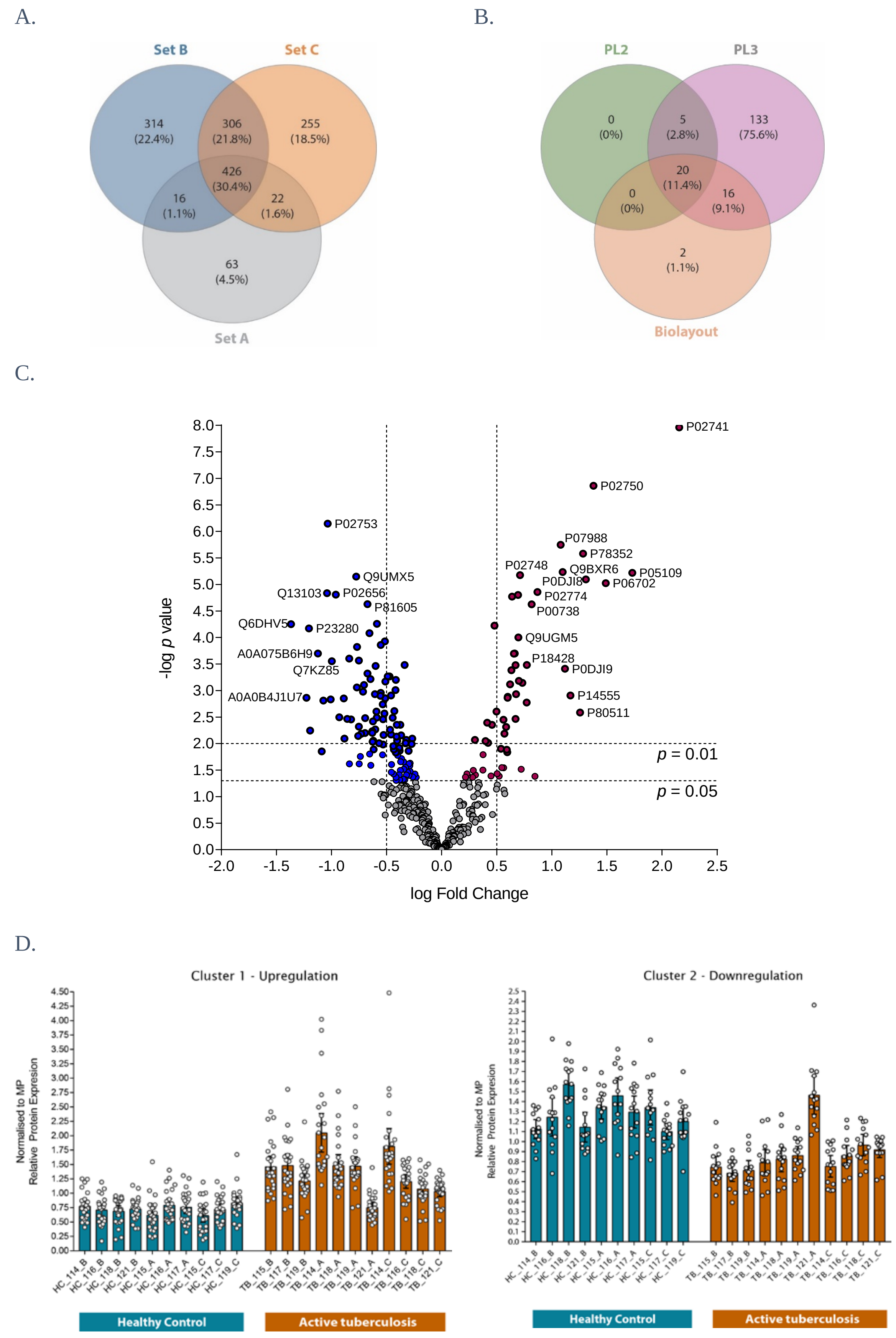


***Figure 58. Differential expressed proteins common among sets A to C from segment 4***

*A. Comparison of fully quantified proteins across sets A, B and C. B. Comparison pipelines of analysis PL2, PL3 and Biolayout (R=0.75). C. Volcano plot of common proteins fully quantified and analysed by PL3. In red* *proteins significantly upregulated, in blue significantly downregulated. Dots with thicker outline indicates proteins with p values FDR adjusted. The proteins exhibiting the largest fold changes were annotated using UniProt ID. D. Main clusters resulting from Biolayout 3D Express, 58 proteins were significantly upregulated and 116 downregulated. Bars with median and 95% CI.*

Annexe 4 presents the list of proteins defined as significantly regulated determined by *PL2* and Annexe 5 by *PL3*. Noticeably, all the 25 proteins defined by *PL2* were FDR corrected for multiple comparison and 127 out of 174 proteins were FDR corrected in the list defined by the *PL3* pipeline. Table 12 lists the 36 proteins common among the *PL2*, *PL3* and Biolayout pipelines.

***Table 12. Common proteins to PL2, PL3 and Biolayout significantly modulated by tuberculosis infection in segment 4***

| Protein Accession | Protein Name | logFC | *p* Value | Adj. *p* Value |
|---|---|---|---|---|
| **P02741** | C-reactive protein (CRP) | 2.155535 | 1.10E-08 | 4.67E-06 |
| **P05109** | S100 calcium binding protein A8 (S100A8) | 1.730152 | 6.02E-06 | 0.000336 |
| P06702 | S100 calcium binding protein A9 (S100A9) | 1.489748 | 9.47E-06 | 0.000367 |
| **P02750** | Leucine rich alpha-2-glycoprotein 1 (LRG1) | 1.377648 | 1.38E-07 | 2.93E-05 |
| **P0DJI8** | Serum amyloid A1 (SAA1) | 1.308297 | 8.03E-06 | 0.000342 |
| P78352 | Discs large MAGUK scaffold protein 4 (DLG4) | 1.282465 | 2.62E-06 | 0.000223 |
| P14555 | Phospholipase A2 group IIA (PLA2G2A) | 1.168357 | 0.001234 | 0.009192 |
| P0DJI9 | Serum amyloid A2 (SAA2) | 1.118803 | 0.000388 | 0.00435 |
| **Q9BXR6** | Complement factor H related 5 (CFHR5) | 1.09874 | 5.83E-06 | 0.000336 |
| **P07988** | Surfactant protein B (SFTPB) | 1.080633 | 1.79E-06 | 0.000191 |
| P00738 | Haptoglobin (HP) | 0.816506 | 2.36E-05 | 0.000559 |
| **P18428** | Lipopolysaccharide binding protein (LBP) | 0.773199 | 0.00033 | 0.003916 |
| P02763 | Orosomucoid 1 (ORM1) | 0.693382 | 1.58E-05 | 0.000448 |
| P50281 | Matrix metallopeptidase 14 (MMP14) | 0.671028 | 0.003418 | 0.018631 |
| P15907 | ST6 beta-galactoside alpha-2,6-sialyltransferase 1 (ST6GAL1) | 0.670154 | 0.000331 | 0.003916 |
| P01011 | Serpin family A member 3 (SERPINA3) | 0.639662 | 1.69E-05 | 0.00045 |
| P25311 | Alpha-2-glycoprotein 1, zinc-binding (AZGP1) | 0.632819 | 0.000413 | 0.004507 |
| P28062 | Proteasome subunit beta 8 (PSMB8) | 0.600077 | 0.001312 | 0.009471 |
| Q9Y275 | Tumour necrosis factor superfamily member 13b (TNFSF13B) | 0.598276 | 0.001366 | 0.009525 |
| Q10588 | Bone marrow stromal cell antigen 1 (BST1) | 0.497329 | 0.002479 | 0.015174 |
| P02671 | Fibrinogen alpha chain (FGA) | 0.479315 | 5.97E-05 | 0.001211 |
| P01009 | Serpin family A member 1 (SERPINA1) | 0.413837 | 0.004043 | 0.020751 |
| P19652 | Orosomucoid 2 (ORM2) | 0.398525 | 0.008966 | 0.035365 |
| P05090 | Apolipoprotein D (APOD) | -0.28712 | 0.02343 | 0.073392 |
| P05452 | C-type lectin domain family 3 member B(CLEC3B) | -0.41938 | 0.000981 | 0.008195 |
| P24592 | Insulin like growth factor binding protein 6 (IGFBP6) | -0.47325 | 0.000543 | 0.005505 |
| P02766 | Transthyretin (TTR) | -0.51538 | 0.000118 | 0.00201 |
| P16035 | TIMP metallopeptidase inhibitor 2(TIMP2) | -0.55365 | 0.001104 | 0.008872 |
| P02652 | Apolipoprotein A2 (APOA2) | -0.58763 | 5.53E-05 | 0.001192 |
| Q16819 | Meprin A subunit alpha (MEP1A) | -0.59883 | 0.000343 | 0.003945 |

|  |  |  |  |  |
|---|---|---|---|---|
| P02655 | Apolipoprotein C2 (APOC2) | -0.60503 | 0.00117 | 0.009089 |
| P35443 | Thrombospondin 4 (THBS4) | -0.6732 | 0.000475 | 0.005063 |
| P02654 | Apolipoprotein C1 (APOC1) | -0.69471 | 0.003317 | 0.018591 |
| P06727 | Apolipoprotein A4 (APOA4) | -0.83936 | 0.00025 | 0.003431 |
| P02656 | Apolipoprotein C3 (APOC3) | -0.96146 | 1.56E-05 | 0.000448 |
| P02753 | Retinol binding protein 4 (RBP4) | -1.03432 | 7.14E-07 | 0.000101 |

Seven proteins of the top 15 proteins listed in Table 12 were selected for validation in two independent cohorts. C-reactive protein (CRP) and Serum amyloid A1 (SAA1) were included as positive control in the validation group since these proteins are well recognised major acute-phase proteins and are expected to be increased in the individuals with pulmonary tuberculosis. S100 calcium binding protein A8 (S100A8) and Lipopolysaccharide binding protein (LBP) have been described in other proteomic profiles of tuberculosis; therefore, the expression of these proteins on the specific cohorts defined for the validation of this work is valuable information for the design of a multi-marker panel.  Identification of four proteins already reported in the literature suggested our approach was robust. Novel proteins such as Complement factor H related 5 (CFHR5), Leucine rich alpha-2-glycoprotein 1 (LRG1) and Surfactant protein B (SFTPB) were additionally selected for validation. Proteins closely associated to the selected proteins were excluded for further validation (S100 calcium binding protein A9 (S100A9) and Serum amyloid A2 (SAA2)) since biological independency is recognised to benefit performance of multi-marker panels.

The multimarker panel designed took place before the detailed proteomic data analysis for segment four, therefore additionally to the 7 proteins selected for validation from this highly curated group of proteins, a subset of 8 proteins from segments 1, 2 and 3 were included in the validation process. Although the statistical power for these segments was low and therefore the likelihood of validation reduced, exploration of proteins with the most consistent patterns of expression might lead to possible biomarkers. We accepted that there may be relatively greater failure rate at the validation stage due to reduced statistical power of the initial proteomic analysis. Table 13 presents the list of proteins selected for validation from segments 1, 2 and 3.  Since this particular subset of proteins was defined prior to the additional proteomic experiments described in this chapter, one protein from segment 4 based on the analysis of only set A was included in this list.

*Table 13. Proteins selected for validation from segments 1, 2 and 3.*

| SEC Segment | Protein Accession | Protein Name | logFC | *p* Value | Adj. *p* Value |
|---|---|---|---|---|---|
| Segment 1 | Q68CZ1 | Protein fantom (RPGRIP1L) | -2.6862852 | 6.42E-05 | 0.0081542 |
| Segment 2 | Q08830 | Fibrinogen-like protein 1 (FGL1) | 1.0763064 | 4.41E-02 | 0.1095433 |
| Segment 2 | Q9NZM1 | Myoferlin (MYOF) | -1.56217 | 0.02545 | 0.109543 |
| Segment 3 | P49747 | Cartilage oligomeric matrix protein (COMP) | 2.2372189 | 4.50E-02 | 0.3941814 |
| Segment 3 | Q12905 | Interleukin enhancer-binding factor 2 (ILF2) | 3.8421888 | 1.06E-02 | 0.3941814 |
| Segment 3 | Q9H2S1 | Small conductance calcium-activated potassium channel protein 2 (KCNN2) | -3.5749196 | 9.24E-04 | 0.1607798 |
| Segment 3 | O14788 | Tumour necrosis factor ligand superfamily member 11 (TNFSF11) | -2.6732522 | 1.52E-03 | 0.1607798 |
| Segment 4 | O94822 | E3 ubiquitin-protein ligase listerin (LTN1) | 3.7360754 | 2.11E-02 | 0.0793238 |

### 5.3.4 Validation of Selected Candidates in Independent Cohorts

Two independent cohorts recruited in South Africa and United Kingdom where included for the validation of the 15 proteins selected, namely: RPGRIP1L, FGL1, COMP, ILF2, MYOF, KCNN2, TNFSF11, LTN1, CRP, S100A8, LRG1, SAA1, CFHR5, SFTPB, and LBP. Demographic description of the two cohorts is presented in Tables 10 and 11.

ELISA or Luminex were used to measure the concentration levels of the selected proteins in plasma and serum depending on the availability of assays. ELISA measurements comprised candidates for which there are commercially available kits, such as: RPGRIP1L, FGL1, COMP, ILF2, KCNN2, LTN1, LRG1 and SFTPB. Three luminex multimarker arrays were custom-made by Protavio; a 4-plex array including LBP, COMP, TNFSF11 and CFHR5, one 2-plex including S100A8 and MYOF and two single-plexes for SAA1 and CRP. The validation of candidates is a multistep process that included sequential steps. Figure 59 depicts the three main conducted steps for validation. Firstly, each one of the immunoassays was optimised. Specifically for the ELISA kits, the dilution factor for plasma and serum was determined for each protein and the incubation time for the primary antibodies. In terms of the Luminex arrays the optimisation of the assays involved determination of optimised dilution factor for serum and plasma samples, composition of the sample diluent buffer to reduce the cross-reactivity of heterophilic antibodies in plasma/serum, doublet discriminator (DD) gating and sensitivity (High RP1 target).

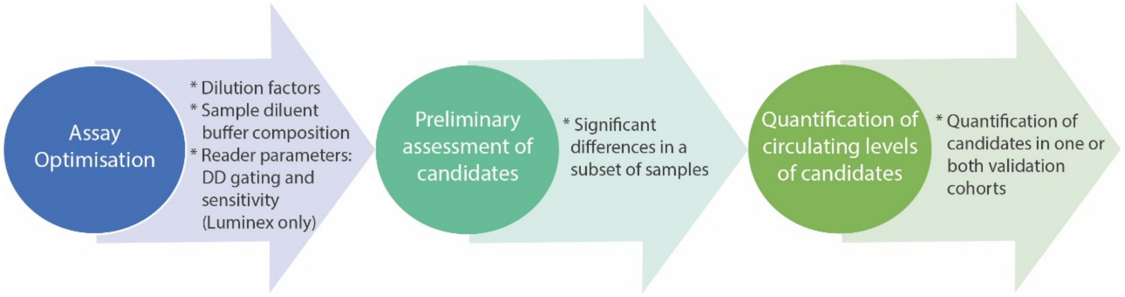

Assay Optimisation
* Dilution factors
* Sample diluent buffer composition
* Reader parameters: DD gating and sensitivity (Luminex only)
Preliminary assessment of candidates
* Significant differences in a subset of samples
Quantification of circulating levels of candidates
* Quantification of candidates in one or both validation cohorts

***Figure 59. General workflow for antibody-based validation of selected candidates***

*Validation of 15 selected proteins was conducted as a stepwise process. Initially each one of ELISA or Luminex assays were optimised. Under optimised conditions a subset of samples were tested to evaluate significant differences. Candidates that showed differences close to significance or significant when comparing control vs. tuberculosis patients were validated on one or both complete validation cohorts.*

Optimised dilution factors for each one of the analytes are presented in Table 14. In the case of the Luminex assays, the sample diluent buffer composition that provided the best reduction of matrix effects of serum/plasma for the 2-plex was 30% Pierce™ Protein-Free T20 (PBS) blocking buffer (ThermoFisher Scientific, UK), 0.5% polyvinyl alcohol and 0.8% polyvinylpyrrolidone. For the 4-plex the buffer previously described was supplemented with 0.5% Super ChemiBlock™ Heterophile Blocking Agent (KC) (Merck, UK). The plates were read using 3000 to 20000 range for the Doublet Discriminator gating on the Luminex. Candidates were validated in one or both cohorts depending on sample volume availability.

***Table 14. Dilution factors for validation of selected proteins***

| Protein | Validation platform | Dilution factor |
|---|---|---|
| RPGRIP1L | ELISA | No dilution |
| FGL1 | ELISA | 1:10 |
| COMP | ELISA | 1:10 |
| COMP | 4-plex Luminex | 1:625 |
| ILF2 | ELISA | 1:5 |
| KCNN2 | ELISA | 1:10 |
| LTN1 | ELISA | 1:5 |
| LRG1 | ELISA | 1:50 |
| SFTPB | ELISA | 1:2/1:5 |
| LBP | 4-plex Luminex | 1:625 |
| TNFSF11 | 4-plex Luminex | 1:625 |
| CFHR5 | 4-plex Luminex | 1:625 |
| S100A8 | 2-plex Luminex | 1:2 |
| MYOF | 2-plex Luminex | 1:2 |
| SAA1 | 2-plex Luminex | 1:100 |
| CRP | 2-plex Luminex | 1:500 |

The following section will present the results of validation of the selected analytes. Figure 60 shows the proteins with no significant differences under optimised conditions in a validation subset of samples using ELISA kits. This subset comprised plasma samples from the South African cohort.

Figure 60A shows RPGRIP1L, Figure 60B FGL1, Figure 60C COMP, Figure 60D KCNN2 and Figure 60E SFTPB.

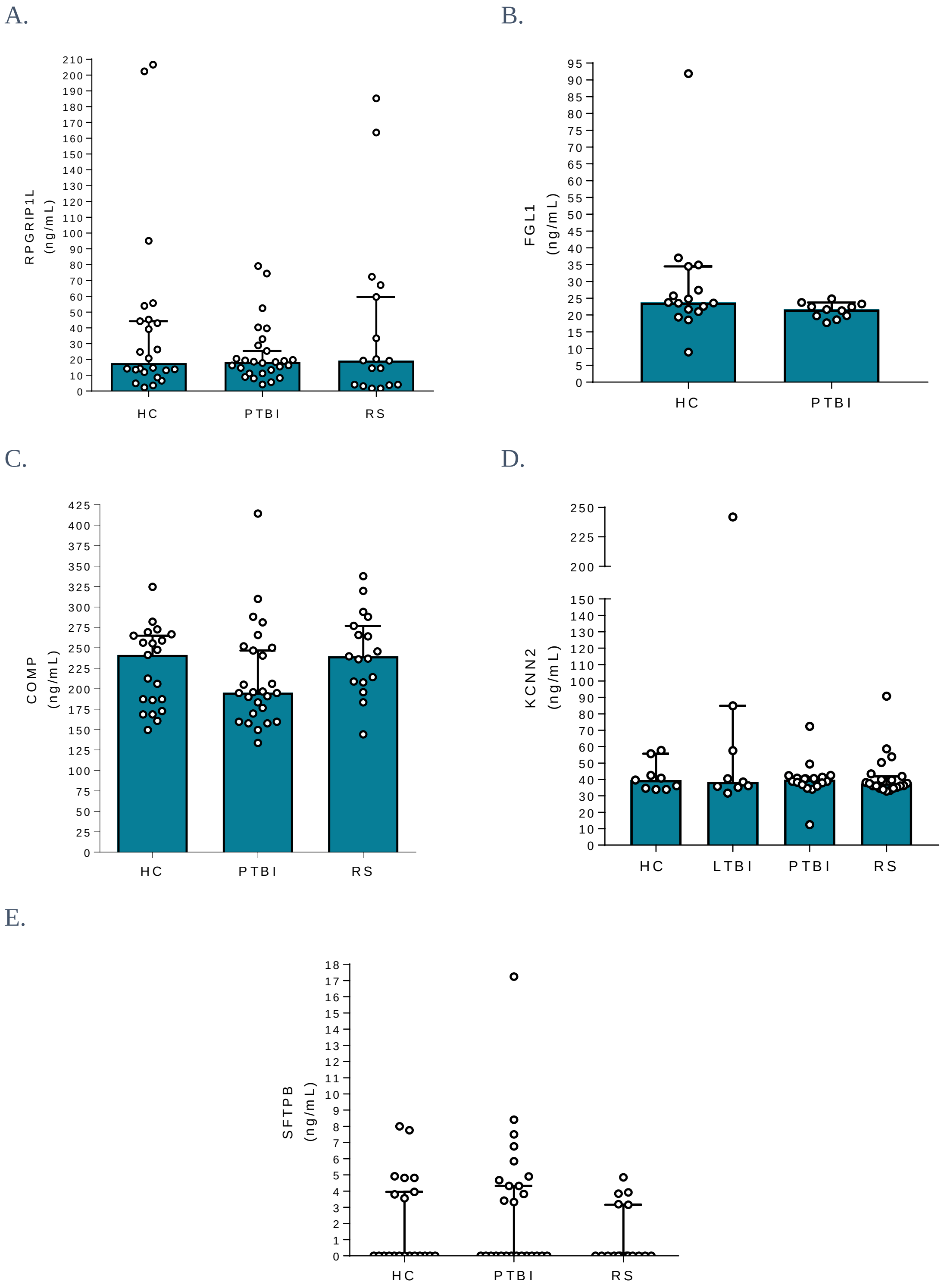


*Figure 60. Assessment of candidates by ELISA on validation subset: Proteins with no significant differences*

*Five candidates showed no significant differences between controls and pulmonary tuberculosis patients when measured by ELISA on a subset of plasma samples from the South African validation cohort. A. RPGRIP1L. B. FGL1. C. COMP. D. KCNN2 and E. SFTPB. Bars with medians and 95%CI, p value considered significant when p <0.05 and calculated from Mann-Whitney test for two groups comparison and Kruskal-Wallis test and multiple Dunn's multiple comparison test for three or more groups. HC: Healthy controls, PTBI: Pulmonary* *tuberculosis infection and RS: Respiratory symptomatic patients. RPGRIP1L: Protein fantom, FGL1: Fibrinogen-like protein 1, COMP: Cartilage oligomeric matrix protein, KCNN2: Small conductance calcium-activated potassium channel protein 2 and SFTPB: Surfactant protein B.*

Figure 61 presents the proteins that passed the preliminary test on the validation subset and were successfully validated on the complete South African or MIMIC cohort by ELISA testing under optimised conditions. Figure 61A shows validation of ILF2 on the MIMIC cohort. Interleukin enhancer-binding factor 2 protein was significantly upregulated on latent and active tuberculosis and respiratory symptomatic patients. Figures 61B and C present validation of LTN on a subset of samples from the South African cohort and the complete MIMIC cohort, respectively. The E3 ubiquitin-protein ligase listerin was significantly upregulated exclusively on pulmonary tuberculosis patients. Figure 61D shows validation of LRG1 on the MIMIC cohort. Leucine rich alpha-2-glycoprotein 1 was significantly upregulated in pulmonary tuberculosis and respiratory symptomatic patients.

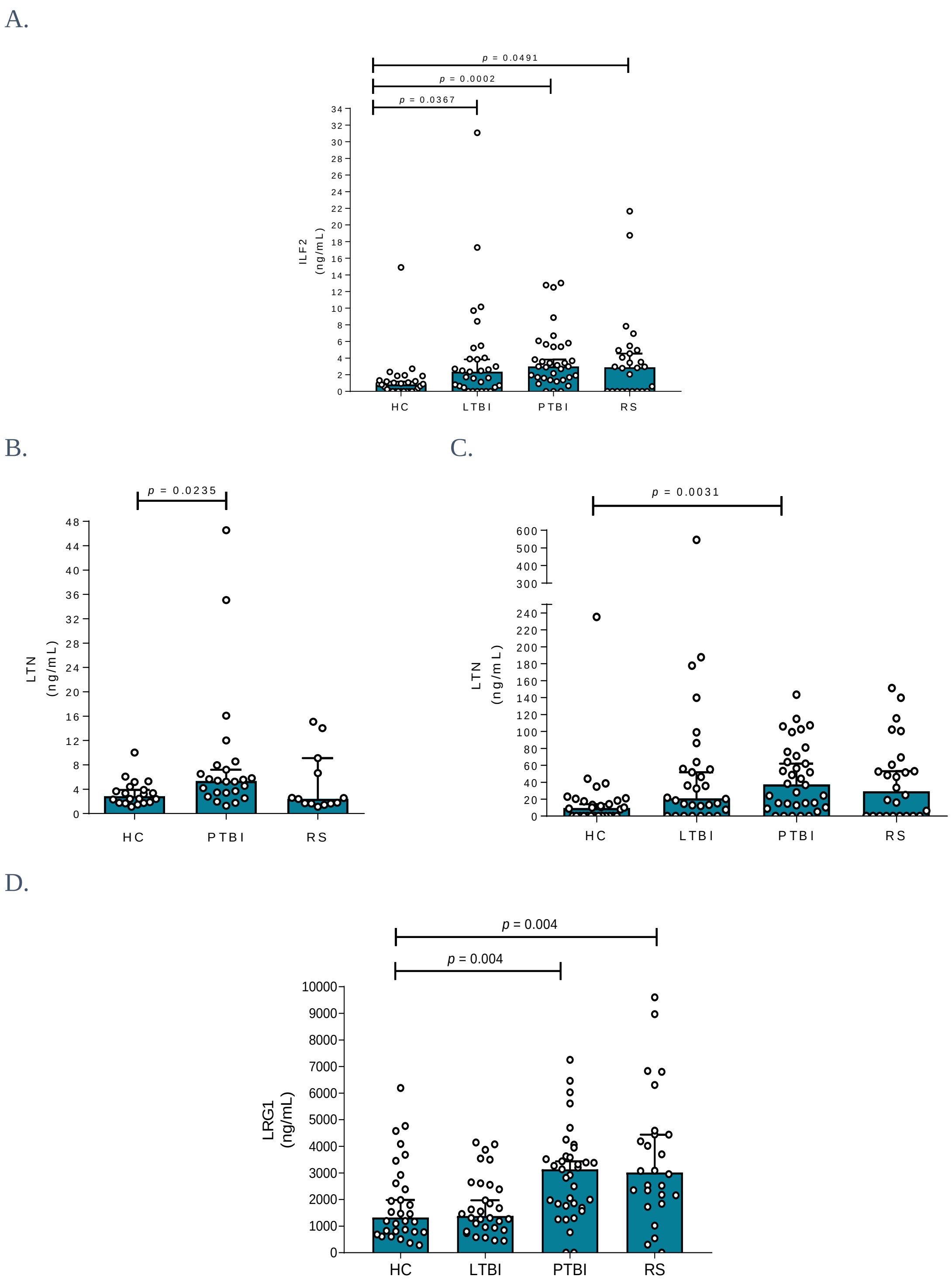


***Figure 61. Proteins validated by ELISA on MIMIC and South African cohort***

*Three candidates were significantly upregulated during pulmonary tuberculosis infection measured by ELISA. A. Validation of ILF2 on the MIMIC validation cohort. B. Validation of LTN on a subset of samples negative for HIV coinfection from the South African cohort. HC (n=20), PTBI (n=24) and RS (n=13). C. Validation of LTN on the MIMIC validation cohort. D. LRG1 validation on the MIMIC validation cohort. Bars with medians and 95%CI, p value considered significant when p <0.05 and calculated from Kruskal-Wallis test and multiple Dunn's multiple comparison test for three or more groups. HC: Healthy controls, PTBI: Pulmonary tuberculosis infection and RS: Respiratory symptomatic patients. IFL2: Interleukin enhancer-binding factor 2, LTN: E3 ubiquitin-protein ligase listerin and LRG1: Leucine rich alpha-2-glycoprotein 1.*

Figure 62 presents the results of validation by Luminex multiplex arrays on the mimic validation cohort and Figure 63 on the South Africa cohort. Figure 62A shows validation of CFHR5 on MIMIC and 63A on the South African cohort. CFHR5 was significantly upregulated in the pulmonary tuberculosis patients on both cohorts. This protein was upregulated as well in the extrapulmonary tuberculosis and respiratory symptomatic patients on the MIMIC cohort. CFHR5 was upregulated in patients with pulmonary tuberculosis that were HIV co-infected. Figure 62B shows validation of LBP on the MIMIC cohort and 632B on the South African cohort. Similarly to CFHR5, the LBP was significantly upregulated in the pulmonary tuberculosis patients on both cohorts. LBP was upregulated as well in the extrapulmonary tuberculosis and respiratory symptomatic patients on the MIMIC cohort but not on the South African cohort. Additionally, LBP was upregulated in patients with pulmonary tuberculosis who were HIV co-infected on the South African cohort. Figure 62C presents validation of TNFSF11 on MIMIC and 63C on the South African cohort. The TNFSF11 (RANKL) was not significantly modulated on the MIMIC nor South African cohort. Figure 62D shows the validation of COMP on MIMIC and 63D on the South African cohort. The COMP was not significantly regulated on any cohort in the tuberculosis context. However, data on Figure 63D suggests that COMP is upregulated in patients HIV infected but not co-infected with *Mtb*.

CRP and SAA were included in the panel as positive controls for the validation process and as expected, these two acute-phase proteins were upregulated on both cohorts as depicted in Figures 62E, 63E for SAA1 from MIMIC and South Africa, respectively. Figure 62F for CRP on MIMIC. C-reactive protein evaluation on the South African cohort was part of the blood test profile at enrolment (data not shown). SAA-1 was significantly upregulated in the extrapulmonary tuberculosis and respiratory symptomatic patients on the MIMIC cohort. On the contrary, SAA1 is only upregulated in tuberculosis patients HIV negative and positive on the South African cohort.

MYOF and S100A8 were tested together in 2-plex array. However, most of concentration levels on the tested samples were below detection and/or quantification levels, which impaired the group comparison. Therefore data from these 2-plex is not presented.

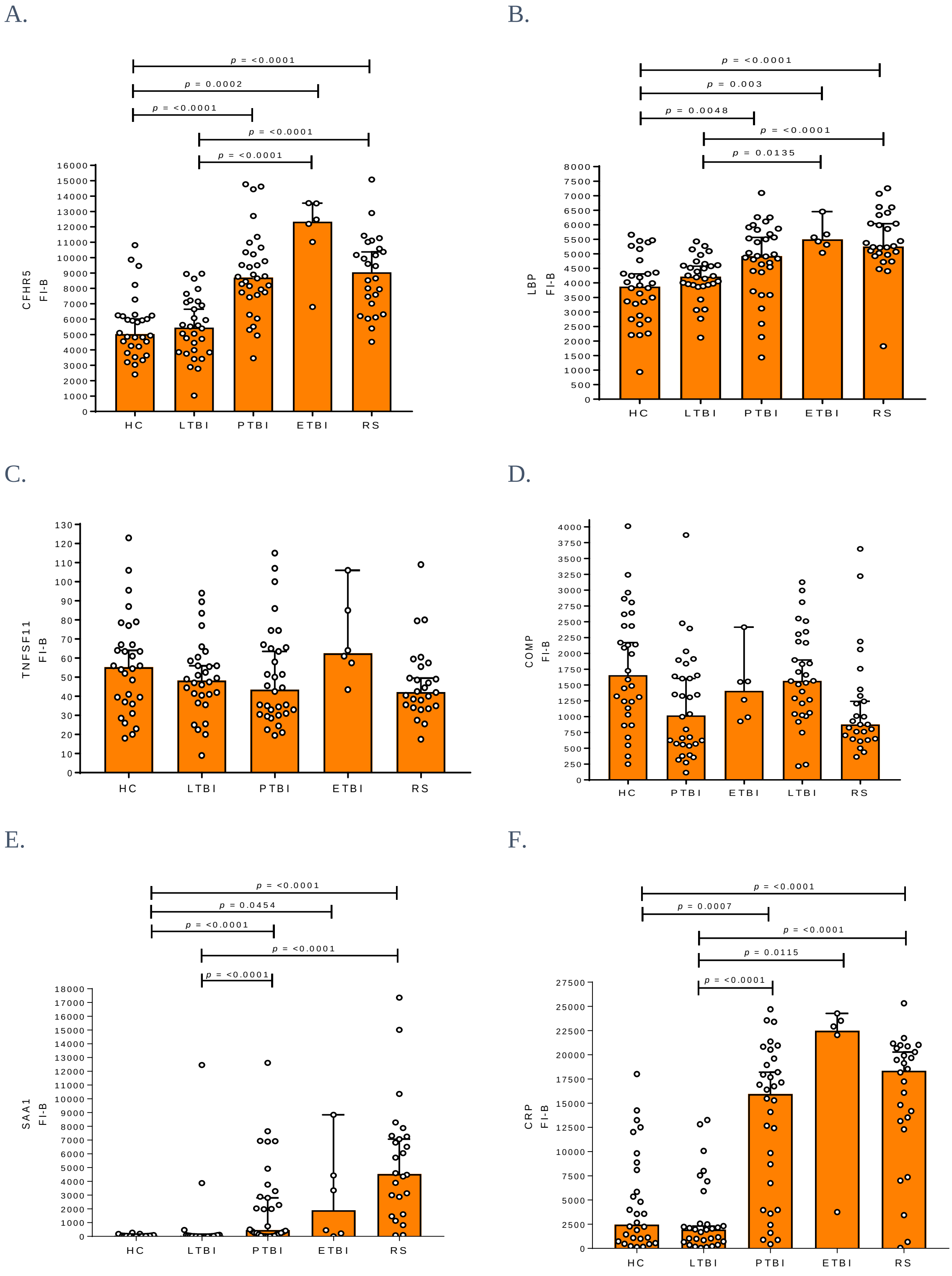


***Figure 62. Proteins evaluated on MIMIC validation cohort by Luminex***

*Six protein candidates were evaluated by Luminex arrays on both validation cohorts. A set of four different arrays were designed and optimised for this purpose: one 4-plex array including LBP, COMP, TNFSF11 and CFHR5, one 2-plex comprising S100A8 and MYOF and two single-plex for SAA1 and CRP. Fluorescence intensities are presented on the y-axis. A. Validation of CFHR5. B. Validation of LBP. C. Determination of TNFSF11. D. Evaluation of COMP. E. Validation of SAA1 and F. Validation of CRP. Bars with medians and 95%CI, p value considered significant when p <0.05 and calculated from Kruskal-Wallis test and multiple Dunn's multiple comparison test. HC: Healthy controls, PTBI: Pulmonary tuberculosis infection and RS:*

*Respiratory symptomatic patients. CFHR5: Complement factor H related 5, LBP: Lipopolysaccharide binding protein, TNFSF11: Tumour necrosis factor ligand superfamily member 11, COMP: Cartilage oligomeric matrix protein, SAA1: Serum amyloid A1 and CRP: C-reactive protein.*

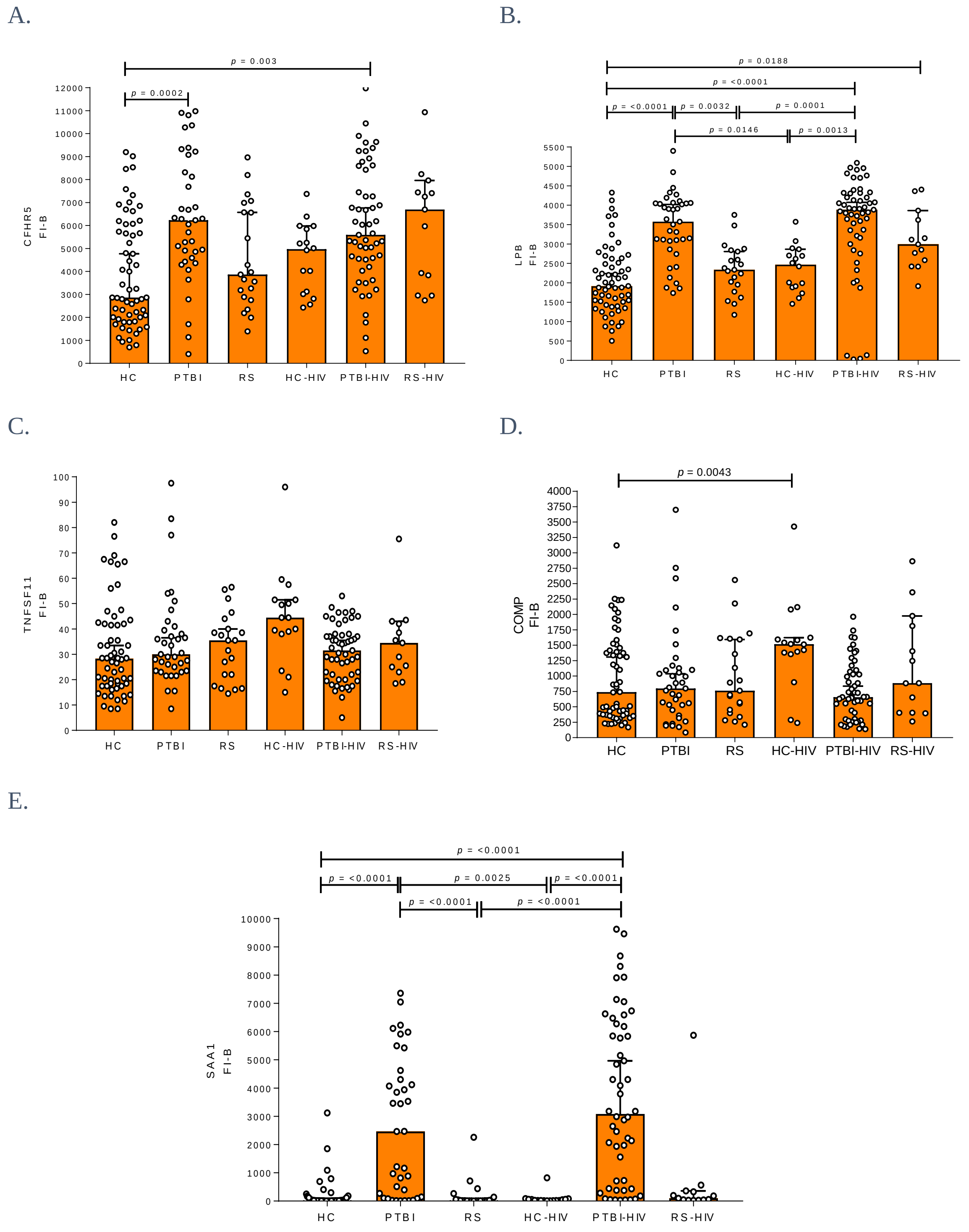


***Figure 63. Proteins evaluated on South African validation cohort by Luminex***

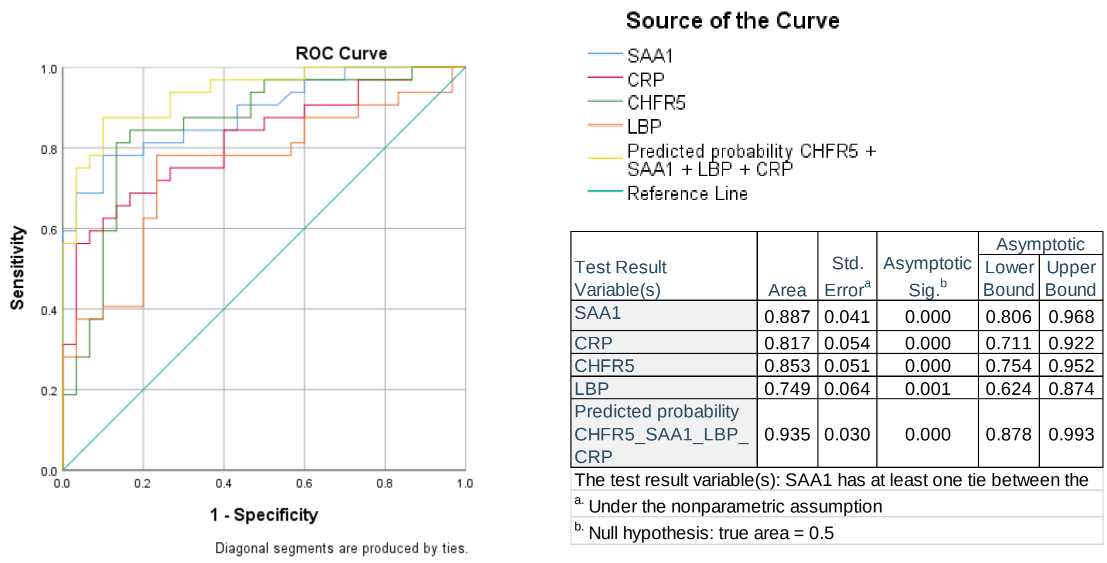
*Six protein candidates were evaluated by Luminex arrays on both validation cohorts. A set of four different arrays were designed and optimised for this purpose: one 4-plex array including LBP, COMP, TNFSF11 and CFHR5, one 2-plex comprising S100A8 and MYOF and two single-plex for SAA1 and CRP. Fluorescence intensities are presented on the y-axis. A. Validation of CFHR5. B. Validation of LBP. C. Determination of TNFSF11. D. Evaluation of COMP and E. Validation of SAA1. Bars with medians and 95%CI, p value* *considered significant when p <0.05 and calculated from Kruskal-Wallis test and multiple Dunn's multiple comparison test. HC: Healthy controls, PTBI: Pulmonary tuberculosis infection and RS: Respiratory symptomatic patients. CFHR5: Complement factor H related 5, LBP: Lipopolysaccharide binding protein, TNFSF11: Tumour necrosis factor ligand superfamily member 11, COMP: Cartilage oligomeric matrix protein and SAA1: Serum amyloid A1.*

Therefore, taking the ELISA and luminex data together, of the 15 proteins that were selected, 7 were statistically significantly elevated (57%). Three proteins where the assay could not detect the analyte in plasma are excluded. When proteins identified in segment 4 by analysis of 21 samples are considered, from 7 selected, 5 were validated (71.4%), showing that the more comprehensive early analysis increases significantly the predictive power, even considering the two proteins (SFTPB and S100A8) where the assay could not detect the proteins in plasma.

A preliminary assessment of the performance of the validated proteins as classifiers of tuberculosis disease compared to the controls was conducted using receiver operator characteristic (ROC) curves and the area under the curve (AUC) as indicator of such performance (93). Fluorescence intensities or concentration levels of significantly regulated proteins were utilised to generate individual ROC curves for each analyte. Additionally, binomial logistic regression was performed to predict the probability of tuberculosis disease (dichotomous dependent variable) based on a set of predictors, in this case the combination of the validated proteins. These probabilities were used to calculate the combined ROC curves.

Figure 64 presents individual and combined ROC curves with statistics for the AUC using significantly regulated proteins as predictors. Figure 64A shows ROC curves and AUC analysis based on fluorescence intensities of SAA1, CRP, CFHR5 and LBP in the MIMIC cohort. The best AUC was achieved with combination of the four proteins (AUC=0.935, 95% confidence interval: 0.878-0.993, $p$ < 0.001). Figure 64B shows ROC curves and AUC analysis based on the concentration (ng/mL) of LTN, LRG1 and ILF2 in the MIMIC cohort. The best AUC was achieved with ILF2 alone (AUC=0.826, 95% confidence interval: 0.690-0.961, $p$ < 0.001). Figure 64C shows ROC curves and AUC analysis based on fluorescence intensities of SAA1, CFHR5 and LBP in the South African cohort. The best AUC was achieved with combination of the three proteins (AUC=0.886, 95% confidence interval: 0.819-0.952, $p$ < 0.001).

A.



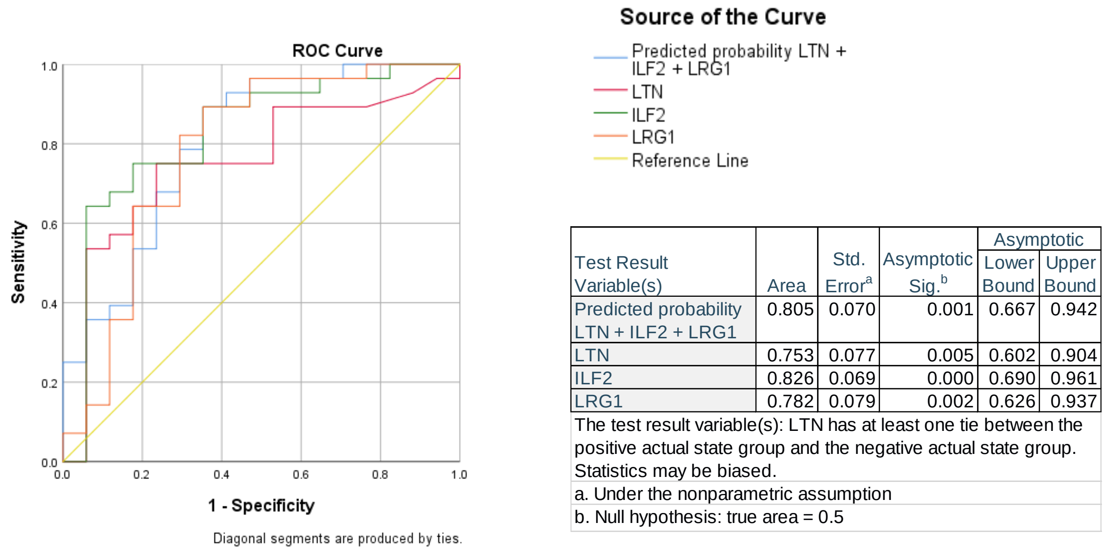
| Test Result Variable(s) | Area | Std. Error[a] | Asymptotic Sig.[b] | Asymptotic Lower Bound | Upper Bound |
|---|---|---|---|---|---|
| SAA1 | 0.887 | 0.041 | 0.000 | 0.806 | 0.968 |
| CRP | 0.817 | 0.054 | 0.000 | 0.711 | 0.922 |
| CHFR5 | 0.853 | 0.051 | 0.000 | 0.754 | 0.952 |
| LBP | 0.749 | 0.064 | 0.001 | 0.624 | 0.874 |
| Predicted probability CHFR5_SAA1_LBP_CRP | 0.935 | 0.030 | 0.000 | 0.878 | 0.993 |

The test result variable(s): SAA1 has at least one tie between the

[a] Under the nonparametric assumption

[b] Null hypothesis: true area = 0.5

B.



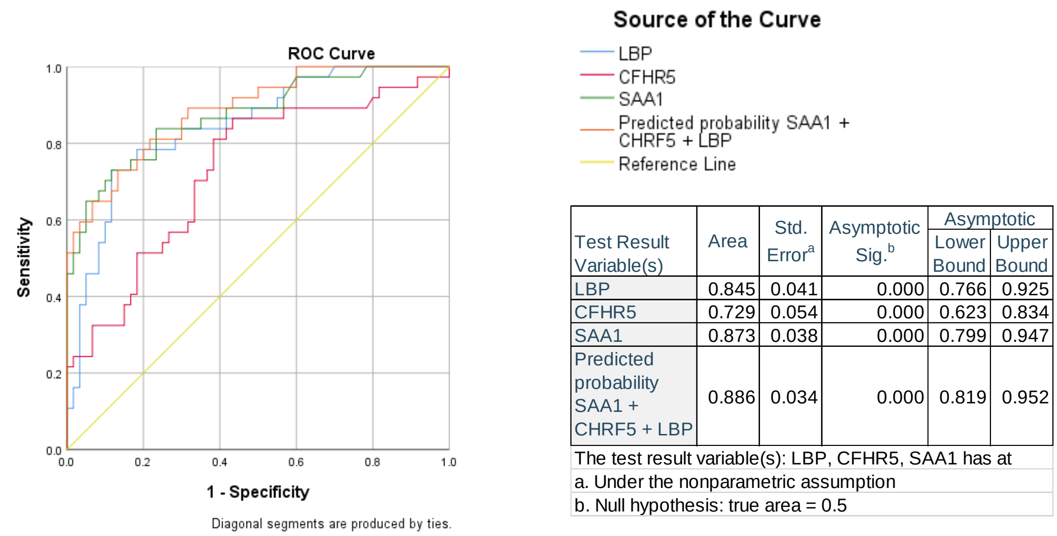
| Test Result Variable(s) | Area | Std. Error[a] | Asymptotic Sig.[b] | Asymptotic Lower Bound | Upper Bound |
|---|---|---|---|---|---|
| Predicted probability LTN + ILF2 + LRG1 | 0.805 | 0.070 | 0.001 | 0.667 | 0.942 |
| LTN | 0.753 | 0.077 | 0.005 | 0.602 | 0.904 |
| ILF2 | 0.826 | 0.069 | 0.000 | 0.690 | 0.961 |
| LRG1 | 0.782 | 0.079 | 0.002 | 0.626 | 0.937 |

The test result variable(s): LTN has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

C.



| Test Result Variable(s) | Area | Std. Error[a] | Asymptotic Sig.[b] | Asymptotic Lower Bound | Upper Bound |
|---|---|---|---|---|---|
| LBP | 0.845 | 0.041 | 0.000 | 0.766 | 0.925 |
| CFHR5 | 0.729 | 0.054 | 0.000 | 0.623 | 0.834 |
| SAA1 | 0.873 | 0.038 | 0.000 | 0.799 | 0.947 |
| Predicted probability SAA1 + CHRF5 + LBP | 0.886 | 0.034 | 0.000 | 0.819 | 0.952 |

The test result variable(s): LBP, CFHR5, SAA1 has at

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

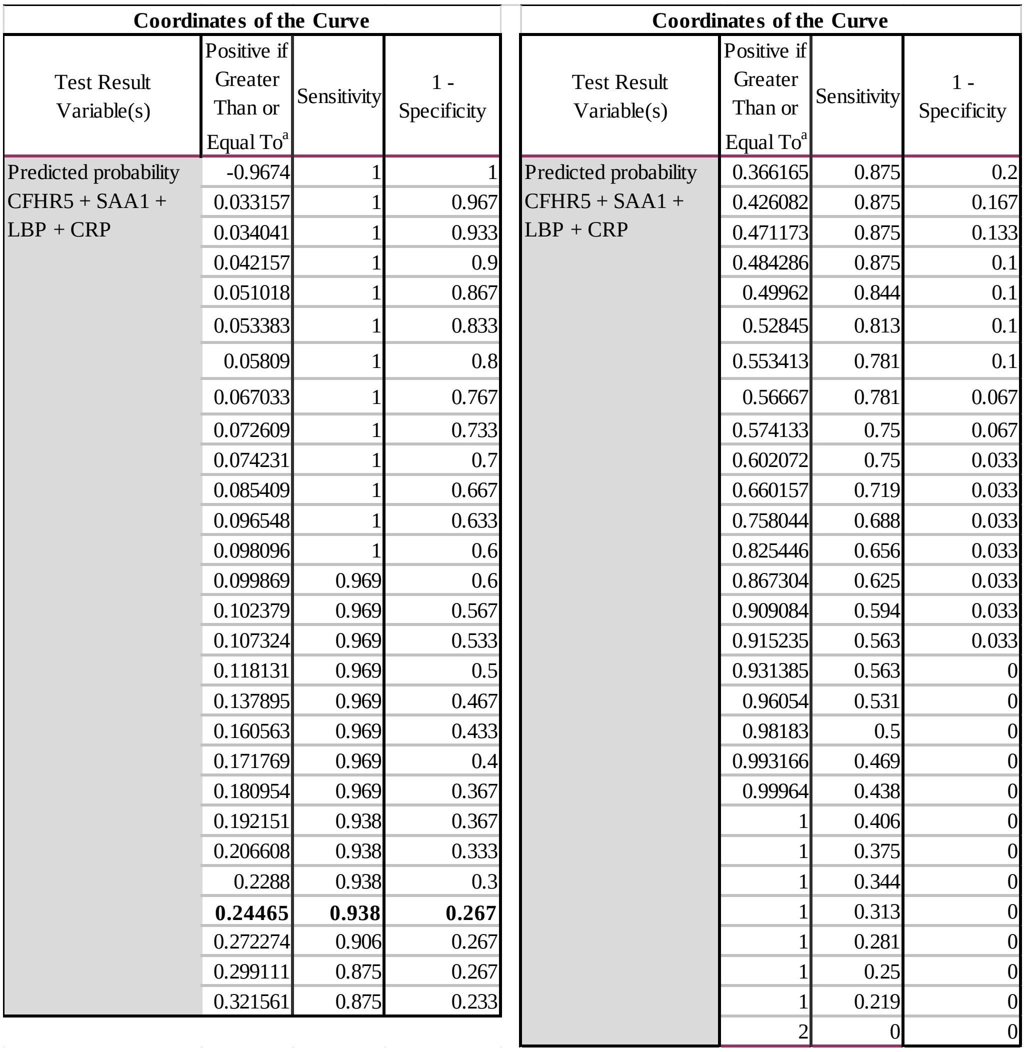
*Figure 64. Receiver operator characteristic curves for validated candidates*

*ROC curves were generated after binary logistic regression using SPSS Statistics v.25. A. ROC curves for the MIMIC cohort calculated for individual analytes LBP, CFHR5, SAA1 and CRP. A combined curve for the combination of the four analytes was also produced. B. ROC curves for the MIMIC cohort calculated for individual analytes IFL2, LRG1 and LTN and combined analytes. C. ROC curves for the South African cohort calculated for individual analytes LBP, CFHR5 and SAA1. A combined curve for the combination of the three analytes was also produced. Table with the statistical description of each ROC curve is included.*

Table 15 presents the coordinates of the ROC curves with the highest AUC including cut-off values, sensitivity and specificity values generated in SPSS Statistics v.25. Table 15A shows the curve coordinates based on fluorescence intensities in the MIMIC cohort, B the curve coordinates based on concentration levels (ng/mL) in the MIMIC cohort and C the curve coordinates based on fluorescence intensities in the South African cohort.

***Table 15. Coordinates of the ROC curves***

A. MIMIC – Fluorescence intensities used for calculations



**Coordinates of the Curve**

| Test Result Variable(s) | Positive if Greater Than or Equal To[a] | Sensitivity | 1 - Specificity |
|---|---|---|---|
| Predicted probability CFHR5 + SAA1 + LBP + CRP | -0.9674 | 1 | 1 |
| | 0.033157 | 1 | 0.967 |
| | 0.034041 | 1 | 0.933 |
| | 0.042157 | 1 | 0.9 |
| | 0.051018 | 1 | 0.867 |
| | 0.053383 | 1 | 0.833 |
| | 0.05809 | 1 | 0.8 |
| | 0.067033 | 1 | 0.767 |
| | 0.072609 | 1 | 0.733 |
| | 0.074231 | 1 | 0.7 |
| | 0.085409 | 1 | 0.667 |
| | 0.096548 | 1 | 0.633 |
| | 0.098096 | 1 | 0.6 |
| | 0.099869 | 0.969 | 0.6 |
| | 0.102379 | 0.969 | 0.567 |
| | 0.107324 | 0.969 | 0.533 |
| | 0.118131 | 0.969 | 0.5 |
| | 0.137895 | 0.969 | 0.467 |
| | 0.160563 | 0.969 | 0.433 |
| | 0.171769 | 0.969 | 0.4 |
| | 0.180954 | 0.969 | 0.367 |
| | 0.192151 | 0.938 | 0.367 |
| | 0.206608 | 0.938 | 0.333 |
| | 0.2288 | 0.938 | 0.3 |
| | **0.24465** | **0.938** | **0.267** |
| | 0.272274 | 0.906 | 0.267 |
| | 0.299111 | 0.875 | 0.267 |
| | 0.321561 | 0.875 | 0.233 |
| Predicted probability CFHR5 + SAA1 + LBP + CRP | 0.366165 | 0.875 | 0.2 |
| | 0.426082 | 0.875 | 0.167 |
| | 0.471173 | 0.875 | 0.133 |
| | 0.484286 | 0.875 | 0.1 |
| | 0.49962 | 0.844 | 0.1 |
| | 0.52845 | 0.813 | 0.1 |
| | 0.553413 | 0.781 | 0.1 |
| | 0.56667 | 0.781 | 0.067 |
| | 0.574133 | 0.75 | 0.067 |
| | 0.602072 | 0.75 | 0.033 |
| | 0.660157 | 0.719 | 0.033 |
| | 0.758044 | 0.688 | 0.033 |
| | 0.825446 | 0.656 | 0.033 |
| | 0.867304 | 0.625 | 0.033 |
| | 0.909084 | 0.594 | 0.033 |
| | 0.915235 | 0.563 | 0.033 |
| | 0.931385 | 0.563 | 0 |
| | 0.96054 | 0.531 | 0 |
| | 0.98183 | 0.5 | 0 |
| | 0.993166 | 0.469 | 0 |
| | 0.99964 | 0.438 | 0 |
| | 1 | 0.406 | 0 |
| | 1 | 0.375 | 0 |
| | 1 | 0.344 | 0 |
| | 1 | 0.313 | 0 |
| | 1 | 0.281 | 0 |
| | 1 | 0.25 | 0 |
| | 1 | 0.219 | 0 |
| | 2 | 0 | 0 |

B. MIMIC – Concentration (ng/mL) used for calculations

| Coordinates of the Curve | | | |
|---|---|---|---|
| Test Result Variable(s) | Positive if Greater Than or Equal To[a] | Sensitivity | 1 - Specificity |
| ILF2 | -0.875 | 1 | 1 |
| | 0.1745 | 1 | 0.941 |
| | 0.377 | 1 | 0.882 |
| | 0.5986 | 1 | 0.824 |
| | 0.7285 | 0.964 | 0.824 |
| | 0.8323 | 0.964 | 0.765 |
| | 0.8764 | 0.964 | 0.706 |
| | 0.8857 | 0.964 | 0.647 |
| | 0.9029 | 0.929 | 0.647 |
| | 0.9632 | 0.929 | 0.588 |
| | 1.0352 | 0.929 | 0.529 |
| | 1.1196 | 0.929 | 0.471 |
| | 1.1899 | 0.893 | 0.471 |
| | 1.2517 | 0.893 | 0.412 |
| | 1.3381 | 0.893 | 0.353 |
| | 1.4819 | 0.857 | 0.353 |
| | 1.6346 | 0.821 | 0.353 |
| | 1.6778 | 0.786 | 0.353 |
| | 1.7605 | 0.75 | 0.353 |
| | 1.8443 | 0.75 | 0.294 |
| | 1.878 | 0.75 | 0.235 |
| | 1.9262 | 0.75 | 0.176 |
| | 2.0497 | 0.714 | 0.176 |

| Coordinates of the Curve | | | |
|---|---|---|---|
| Test Result Variable(s) | Positive if Greater Than or Equal To[a] | Sensitivity | 1 - Specificity |
| ILF2 | 2.241 | 0.679 | 0.176 |
| | 2.5134 | 0.679 | 0.118 |
| | 2.7037 | 0.643 | 0.118 |
| | 2.8017 | 0.643 | 0.059 |
| | 2.9418 | 0.607 | 0.059 |
| | 3.0632 | 0.536 | 0.059 |
| | 3.27 | 0.5 | 0.059 |
| | 3.4061 | 0.464 | 0.059 |
| | 3.4909 | 0.429 | 0.059 |
| | 3.6261 | 0.393 | 0.059 |
| | 4.5222 | 0.357 | 0.059 |
| | 5.3727 | 0.321 | 0.059 |
| | 5.5201 | 0.286 | 0.059 |
| | 5.7353 | 0.25 | 0.059 |
| | 5.939 | 0.214 | 0.059 |
| | 6.3868 | 0.179 | 0.059 |
| | 7.7949 | 0.143 | 0.059 |
| | 10.6927 | 0.107 | 0.059 |
| | 12.6438 | 0.071 | 0.059 |
| | 12.9138 | 0.036 | 0.059 |
| | 13.9786 | 0 | 0.059 |
| | 15.9136 | 0 | 0 |

C. South Africa cohort – Fluorescence intensities used for calculations

| Coordinates of the Curve | | | |
|---|---|---|---|
| Test Result Variable(s) | Positive if Greater Than or Equal To[a] | Sensitivity | 1 - Specificity |
| Predicted probability SAA1 + CHRF5 + LBP | 0 | 1 | 1 |
| | 0.0425226 | 1 | 0.983 |
| | 0.0496764 | 1 | 0.967 |
| | 0.0528383 | 1 | 0.95 |
| | 0.0559323 | 1 | 0.933 |
| | 0.0594681 | 1 | 0.917 |
| | 0.0634368 | 1 | 0.9 |
| | 0.0687227 | 1 | 0.883 |
| | 0.0728766 | 1 | 0.867 |
| | 0.0746193 | 1 | 0.85 |
| | 0.0762439 | 1 | 0.833 |
| | 0.0802747 | 1 | 0.817 |
| | 0.0843991 | 1 | 0.8 |
| | 0.0880026 | 1 | 0.783 |
| | 0.0908821 | 1 | 0.767 |
| | 0.0931761 | 1 | 0.75 |
| | 0.0967707 | 1 | 0.733 |
| | 0.0995513 | 1 | 0.717 |
| | 0.1022588 | 1 | 0.7 |
| | 0.1048321 | 1 | 0.683 |
| | 0.1063257 | 1 | 0.667 |
| | 0.1078842 | 1 | 0.65 |
| | 0.1092587 | 1 | 0.633 |
| | 0.1148208 | 1 | 0.617 |
| | 0.1212178 | 1 | 0.6 |
| | 0.1232555 | 0.973 | 0.6 |
| | 0.1245202 | 0.946 | 0.6 |
| | 0.1257834 | 0.946 | 0.583 |
| | 0.1268544 | 0.946 | 0.567 |
| | 0.1279723 | 0.946 | 0.55 |
| | 0.1305172 | 0.946 | 0.533 |
| | 0.133438 | 0.946 | 0.517 |
| | 0.1352929 | 0.946 | 0.5 |
| | 0.1368367 | 0.919 | 0.5 |
| | 0.1394832 | 0.919 | 0.483 |
| | 0.1476612 | 0.919 | 0.467 |
| | 0.153943 | 0.919 | 0.45 |
| | 0.1550783 | 0.919 | 0.433 |
| | 0.1578296 | 0.892 | 0.433 |
| | 0.1616226 | 0.892 | 0.417 |
| | 0.1636783 | 0.892 | 0.4 |
| | 0.1653895 | 0.892 | 0.383 |
| | 0.1700262 | 0.892 | 0.367 |
| | 0.1737987 | 0.892 | 0.35 |
| | 0.1774076 | 0.892 | 0.333 |
| | 0.1893629 | 0.892 | 0.317 |
| | 0.2004697 | 0.865 | 0.317 |
| | 0.2036134 | 0.865 | 0.3 |
| | 0.2083621 | 0.838 | 0.3 |
| | 0.2188413 | 0.811 | 0.3 |
| | 0.2355137 | 0.811 | 0.283 |
| | 0.247339 | 0.811 | 0.267 |
| | 0.2498731 | 0.811 | 0.25 |
| | 0.2595306 | 0.811 | 0.233 |
| | 0.2732809 | 0.811 | 0.217 |
| | 0.2832491 | 0.784 | 0.217 |
| | 0.290309 | 0.784 | 0.2 |
| | 0.2978909 | 0.757 | 0.2 |
| | 0.3110075 | 0.757 | 0.183 |
| | 0.3206368 | 0.73 | 0.183 |
| | 0.3230623 | 0.73 | 0.167 |
| | 0.3243051 | 0.73 | 0.15 |
| | 0.336836 | 0.73 | 0.133 |
| | 0.3512792 | 0.703 | 0.133 |
| | 0.3596542 | 0.676 | 0.133 |
| | 0.3769818 | 0.676 | 0.117 |
| | 0.4407209 | 0.649 | 0.117 |
| | 0.4964904 | 0.649 | 0.1 |
| | 0.5206475 | 0.649 | 0.083 |
| | 0.5783908 | 0.649 | 0.067 |
| | 0.6162633 | 0.622 | 0.067 |
| | 0.620378 | 0.595 | 0.067 |
| | 0.6243439 | 0.595 | 0.05 |
| | 0.6406425 | 0.595 | 0.033 |
| | 0.673791 | 0.568 | 0.033 |
| | 0.7150463 | 0.568 | 0.017 |
| | 0.7544402 | 0.541 | 0.017 |
| | 0.7782297 | 0.514 | 0.017 |
| | 0.7936893 | 0.514 | 0 |
| | 0.8097073 | 0.486 | 0 |
| | 0.8769995 | 0.459 | 0 |
| | 0.9463703 | 0.432 | 0 |
| | 0.9608297 | 0.405 | 0 |
| | 0.9682376 | 0.378 | 0 |
| | 0.9745492 | 0.351 | 0 |
| | 0.9775219 | 0.324 | 0 |
| | 0.9785362 | 0.297 | 0 |
| | 0.9793822 | 0.27 | 0 |
| | 0.9828458 | 0.243 | 0 |
| | 0.9884791 | 0.216 | 0 |
| | 0.9933741 | 0.189 | 0 |
| | 0.9958237 | 0.162 | 0 |
| | 0.9962689 | 0.135 | 0 |
| | 0.9967276 | 0.108 | 0 |
| | 0.997824 | 0.081 | 0 |
| | 0.9985297 | 0.054 | 0 |
| | 0.9987584 | 0.027 | 0 |
| | 1 | 0 | 0 |

[a.] The smallest cut-off value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

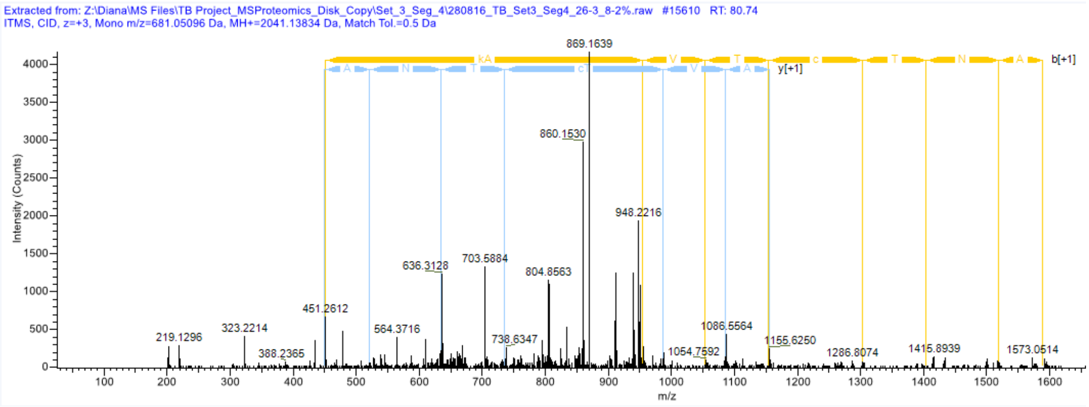
Therefore, the ROC curve analysis potentially gave an optimal AUC of AUC of 0.935 (95%CI, 0.878-0.933) in the MIMIC cohort, by including CFHR5, SAA1, LBP and CRP. In the South Africa cohort the performance was reduced since the CRP data was not available, AUC was 0.886 (95%CI, 0.690 – 0.961). Further analysis of these analytes in diverse cohorts, and the power of adding alternative biomarkers that emerged from the interrogation of segment 4 and further analysis of segment 1 – 3, is needed to define the best diagnostic panel.

### 5.3.5 A potential marker from *Mycobacterium tuberculosis*

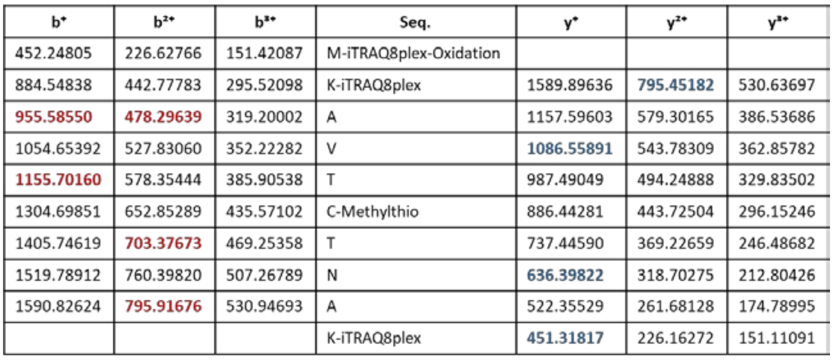
Alternatively to the analysis of the proteomic data based on fully quantified data which led to the selection of the candidates validated above, an alternative approach is to try to identify peptides only present in tuberculosis samples. This analysis approach is necessary to show pathogen-derived peptides. Further analysis of data using *PL3* allowing for missing data indicated the presence of an *Mtb* derived peptide in 10 out of 11 tuberculosis patients. This identification was possible since the spectrum raw data analysis was performed including the reference proteome for *Mycobacterium tuberculosis* strain ATCC 25618/H37Rv (UP000001584) together with the reviewed proteome for *Homo sapiens* (UP000005640). This finding was consistent in the three data sets generated for segment 4. Figure 65 presents the spectrometric evidence for the presence of this peptide derived from the possible dehydrogenase (O69693) product of the gene Rv3726. The peptide sequence MKAVTCTNAK comprises 10 amino acids at the N-terminus of the protein.

Figure 65A shows the list of b and y ions used for sequencing and the CID fragmentation spectrum for the precursor ion 681.05096Da (z = +3) collected in the Set A. Figure 65B depicts the list of b and y ions used for sequencing and the CID fragmentation spectrum for the precursor ion 1021.08215Da (z = +2) collected in the Set B. Figure 65C presents the list of b and y ions used for sequencing and the CID fragmentation spectrum for the precursor ion 681.05188 (z = +3) collected in the Set B. The identification was performed with one high confidence unique peptide and 1 to 3 PSMs with co-isolation ranging 0% to 20%.

A.

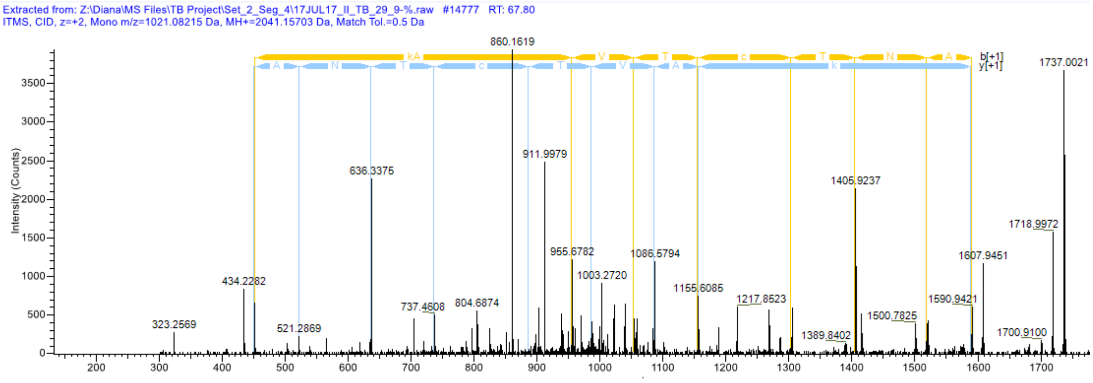


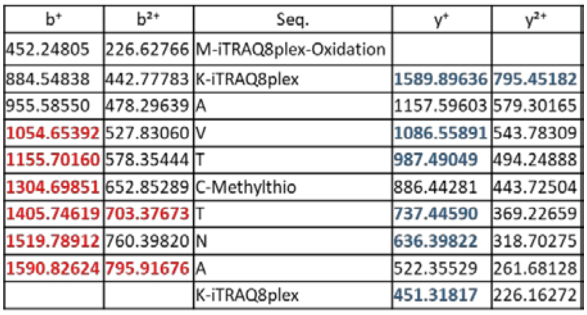
Extracted from: Z:\Diana\MS Files\TB Project_MSProteomics_Disk_Copy\Set_3_Seg_4\280816_TB_Set3_Seg4_26-3_8-2%.raw #15610 RT: 80.74
ITMS, CID, z=+3, Mono m/z=681.05096 Da, MH+=2041.13834 Da, Match Tol.=0.5 Da



| b⁺ | b²⁺ | b³⁺ | Seq. | y⁺ | y²⁺ | y³⁺ |
|---|---|---|---|---|---|---|
| 452.24805 | 226.62766 | 151.42087 | M-iTRAQ8plex-Oxidation | | | |
| 884.54838 | 442.77783 | 295.52098 | K-iTRAQ8plex | 1589.89636 | 795.45182 | 530.63697 |
| 955.58550 | 478.29639 | 319.20002 | A | 1157.59603 | 579.30165 | 386.53686 |
| 1054.65392 | 527.83060 | 352.22282 | V | 1086.55891 | 543.78309 | 362.85782 |
| 1155.70160 | 578.35444 | 385.90538 | T | 987.49049 | 494.24888 | 329.83502 |
| 1304.69851 | 652.85289 | 435.57102 | C-Methylthio | 886.44281 | 443.72504 | 296.15246 |
| 1405.74619 | 703.37673 | 469.25358 | T | 737.44590 | 369.22659 | 246.48682 |
| 1519.78912 | 760.39820 | 507.26789 | N | 636.39822 | 318.70275 | 212.80426 |
| 1590.82624 | 795.91676 | 530.94693 | A | 522.35529 | 261.68128 | 174.78995 |
| | | | K-iTRAQ8plex | 451.31817 | 226.16272 | 151.11091 |

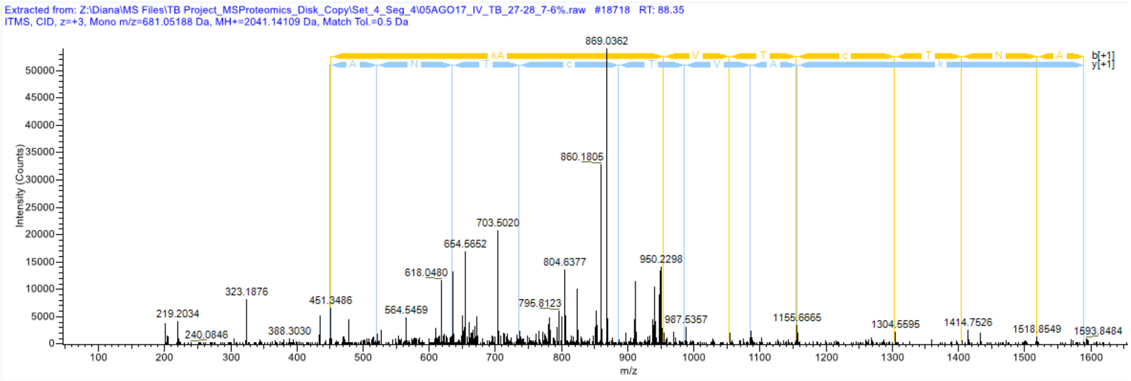
B.

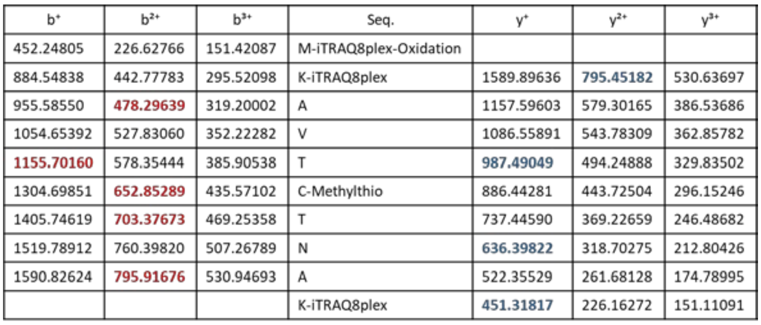




| b⁺ | b²⁺ | Seq. | y⁺ | y²⁺ |
|---|---|---|---|---|
| 452.24805 | 226.62766 | M-iTRAQ8plex-Oxidation | | |
| 884.54838 | 442.77783 | K-iTRAQ8plex | **1589.89636** | **795.45182** |
| 955.58550 | 478.29639 | A | 1157.59603 | 579.30165 |
| **1054.65392** | 527.83060 | V | **1086.55891** | 543.78309 |
| **1155.70160** | 578.35444 | T | **987.49049** | 494.24888 |
| **1304.69851** | 652.85289 | C-Methylthio | 886.44281 | 443.72504 |
| **1405.74619** | **703.37673** | T | **737.44590** | 369.22659 |
| **1519.78912** | 760.39820 | N | **636.39822** | 318.70275 |
| **1590.82624** | **795.91676** | A | 522.35529 | 261.68128 |
| | | K-iTRAQ8plex | **451.31817** | 226.16272 |

C.





| b⁺ | b²⁺ | b³⁺ | Seq. | y⁺ | y²⁺ | y³⁺ |
|---|---|---|---|---|---|---|
| 452.24805 | 226.62766 | 151.42087 | M-iTRAQ8plex-Oxidation | | | |
| 884.54838 | 442.77783 | 295.52098 | K-iTRAQ8plex | 1589.89636 | **795.45182** | 530.63697 |
| 955.58550 | **478.29639** | 319.20002 | A | 1157.59603 | 579.30165 | 386.53686 |
| 1054.65392 | 527.83060 | 352.22282 | V | 1086.55891 | 543.78309 | 362.85782 |
| **1155.70160** | 578.35444 | 385.90538 | T | **987.49049** | 494.24888 | 329.83502 |
| 1304.69851 | **652.85289** | 435.57102 | C-Methylthio | 886.44281 | 443.72504 | 296.15246 |
| 1405.74619 | **703.37673** | 469.25358 | T | 737.44590 | 369.22659 | 246.48682 |
| 1519.78912 | 760.39820 | 507.26789 | N | **636.39822** | 318.70275 | 212.80426 |
| 1590.82624 | **795.91676** | 530.94693 | A | 522.35529 | 261.68128 | 174.78995 |
| | | | K-iTRAQ8plex | **451.31817** | 226.16272 | 151.11091 |

***Figure 65. Mass spectrometric evidence of an Mtb derived peptide***

*CID fragmentation spectra used for the peptide MKAVTCTNAK identification including list of y and b ions. A. Set A. B. Set B and C. Set C.*

The finding of a pathogen-derived peptide in the plasma of HIV-negative individuals is potentially very exciting as it may form the basis of a rule-in test for TB.

### 5.3.6 Bioinformatic mining of proteomic profile from segment four

Further data mining of the proteomic data generated in this work was conducted to interpret biologically relevant patterns of protein expression in plasma of patients with pulmonary tuberculosis. Network-based analysis was applied to the datasets resulting from the detailed profile of segment 4. Specifically, weighted gene co-expression network analysis (WGCNA) was used to explore correlation relationships between clusters of highly correlated proteins (colour modules) and specific sample traits. The batch effect was corrected in order to increase the analysis power. As indicated in the Figure 57A, the main variable explaining the variance of the data is the batch or iTRAQ set experiment. Two different approaches for batch correction were evaluated on the protein expression values from sets A to C: master pool and ComBat. The first approach involved the normalisation of each protein to its expression level in the master pool. Figure 66 presents a PCA with the sample variance distribution after master pool normalisation. Alternatively, ComBat is a well-recognised tool used for batch effect correction initially designed for microarray data but widely used for transcriptomics. ComBat was run as part of the Surrogate Variable Analysis package R package version 3.28.0. Figure 66B presents the sample variance distribution after ComBat correction.  Once the data was batch corrected, the common proteins across the three segments were processed through *PL2* (LIMMA and FDR correction). A direct comparison of the number of significantly regulated proteins using *PL2* resulting from both correction approaches is shown in Figure 66C. In overall, 159 proteins were significantly regulated using ComBat and 72 proteins using master pool correction. Forty-nine proteins were common to both approaches. All *p* values of significantly regulated proteins were FDR corrected.

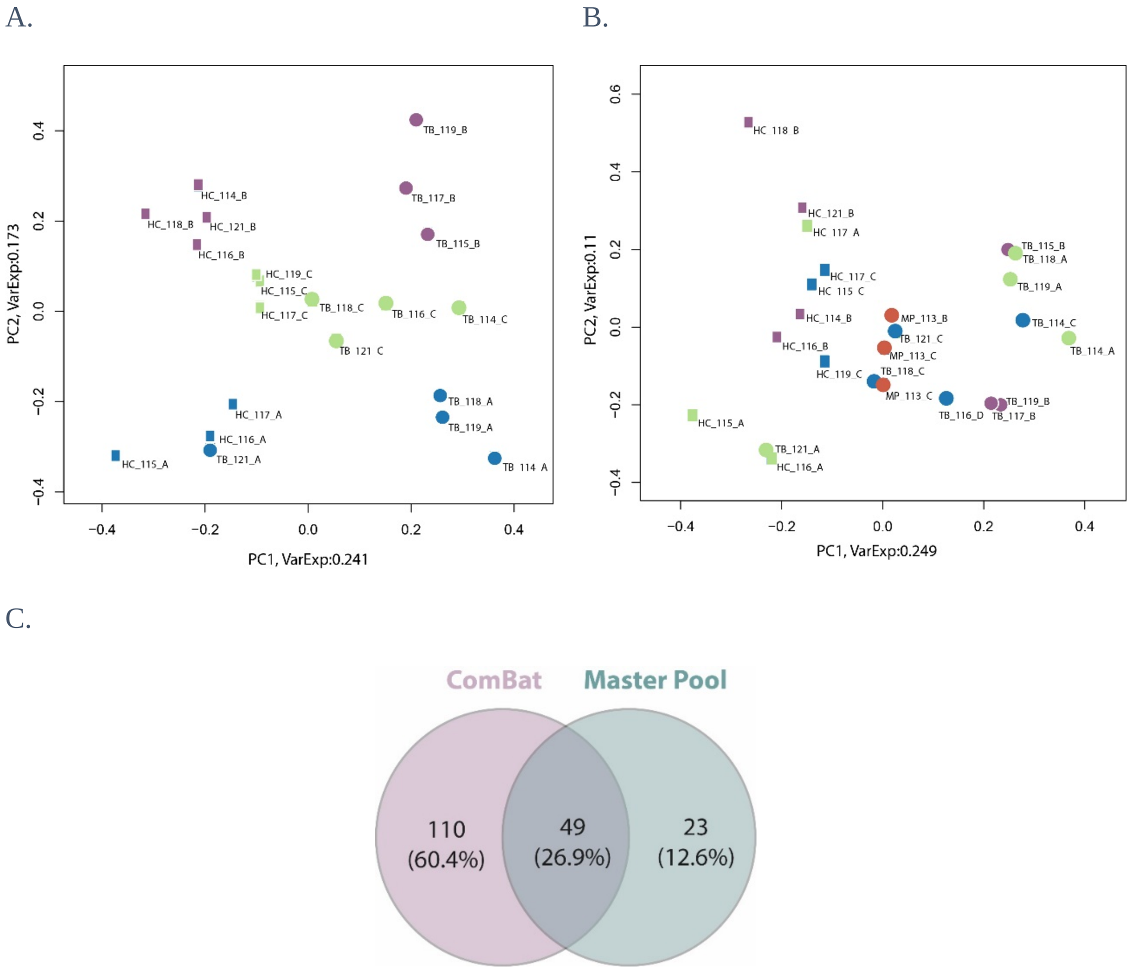


*Figure 66. Batch effect correction*

*Batch effects were corrected to combine the datasets and increase the statistical power of the data analysis. A. PCA datasets from sets A to C batch corrected using master pool. B. PCA datasets from sets A to C batch corrected using ComBat. In blue set A, purple set B, green set C and orange master pool samples. Squares indicate control and circles tuberculosis samples. C. Venn diagram of proteins significantly regulated using PL2 after master pool and ComBat batch correction. Proteins considered significantly regulated when FDR adjusted p value <0.05.*

Considering that ComBat led to better correction of batch effects as demonstrated in Figure 66B and a higher number of proteins significantly regulated, the dataset derived from this approach was used for downstream analysis. WGCNA package for R was run in RStudio. Unsupervised sample hierarchical clustering is shown in Figure 67A. Four covariates are colour coded to assess possible cofounding effect of ethnicity, smoking status and batch. Samples are mainly clustered by disease status (group): healthy controls and pulmonary tuberculosis, excepting two samples TB_121 from set A and TB_118 from set C. Interestingly, samples showed a clustering pattern based on the ethnicity of the participants. Neither smoking status nor batch seemed to affect the clustering of samples. Figure 67B depicts the main protein co-expression modules, the correlation score for the absolute correlation between the protein expression profiles and predefined covariates and finally, significance of correlation. Consistent with the hierarchical clustering, ethnicity exhibits one of the strongest modules (module brown) with a correlation score of 0.81 and *p* vale 0.000006, after the module turquoise correlated to the group (correlation score -0.87 and *p* value 0.0000002). Figure 67C indicates the ethnicity of the participants on the PCA based on common proteins among sets A to C and corrected with ComBat.

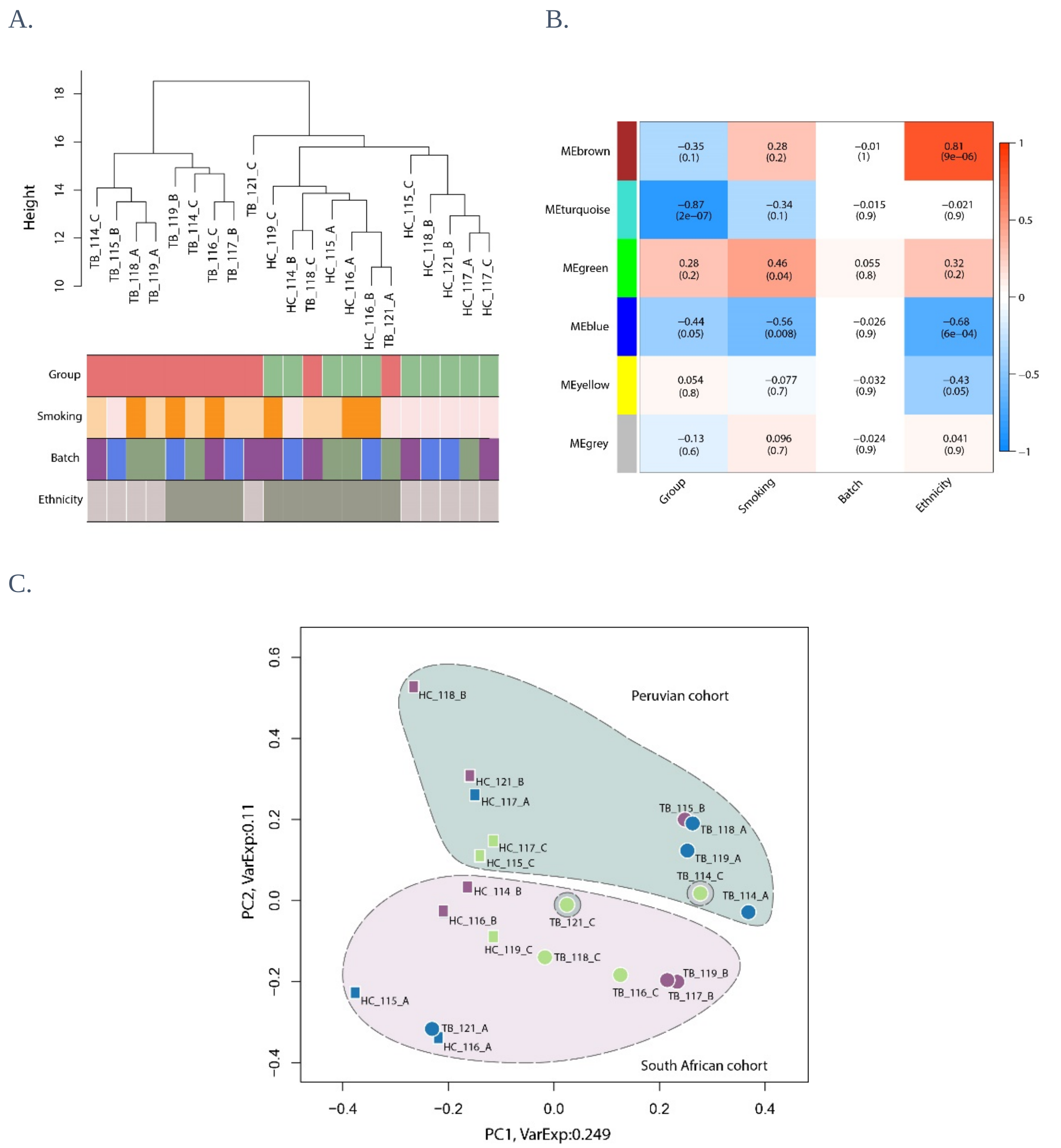


***Figure 67. WGCNA analysis of plasma proteomic profile resulting from segment 4***

*WGCNA analysis based on the common proteins resulting from segment 4 profile across iTRAQ sets A to C. A. Hierarchical clustering of samples with colour-coded predefined covariates. B. Module-traits associations describing protein co-expression modules and predefined covariates. C. PCA of ComBat corrected protein expression indicating ethnicities of each participant. In blue set A, purple set B, green set C and orange master pool samples. Squares indicate control and circles tuberculosis samples. Samples that do not cluster based on ethnicity are indicated with dashed circles.*

Assessment of sample clustering and main co-expression modules suggested that ethnicity might be a biologically relevant covariate to explore further. Samples from the two ethnicities considered in this study were separated and the datasets processed independently using WGCNA. The pulmonary tuberculosis sample labelled with the iTRAQ tag121 from experiment A was removed since this sample exhibited the largest deviation from the group and its removal increased the statistical power: as a result 10 samples were included in each ethnic group.

Figure 68 presents a parallel of the sample clustering, protein co-expression modules and covariates derived from WGCNA between Peruvian and South African samples. Figure 68A and B present the hierarchical clustering of samples with group and batch colour-coded for Peruvian and South African discovery cohorts, respectively. Samples clearly clustered by group, control vs. pulmonary tuberculosis and did not exhibit an obvious batch effect. Figure 68C and D depict the clustering dendograms of the topological overlap measurement (TOP) matrix, which show clusters of highly interconnected proteins with assigned colour modules for Peruvian and South African samples, respectively. Figure 68E and F present the module-traits associations describing protein co-expression modules and two covariates: group and batch for Peruvian and South African individuals, respectively. Three modules were strongly correlated to the disease status in the Peruvian cohort: module pink (correlation score -0.65, *p* value 0.04), module green (correlation score -0.82, *p* value 0.004) and, module turquoise (correlation score -0.94, *p* value 0.00004). In contrast, only one module was significantly associated to the disease status in the South African cohort, the module turquoise (correlation score -0.93, *p* value 0.00009).

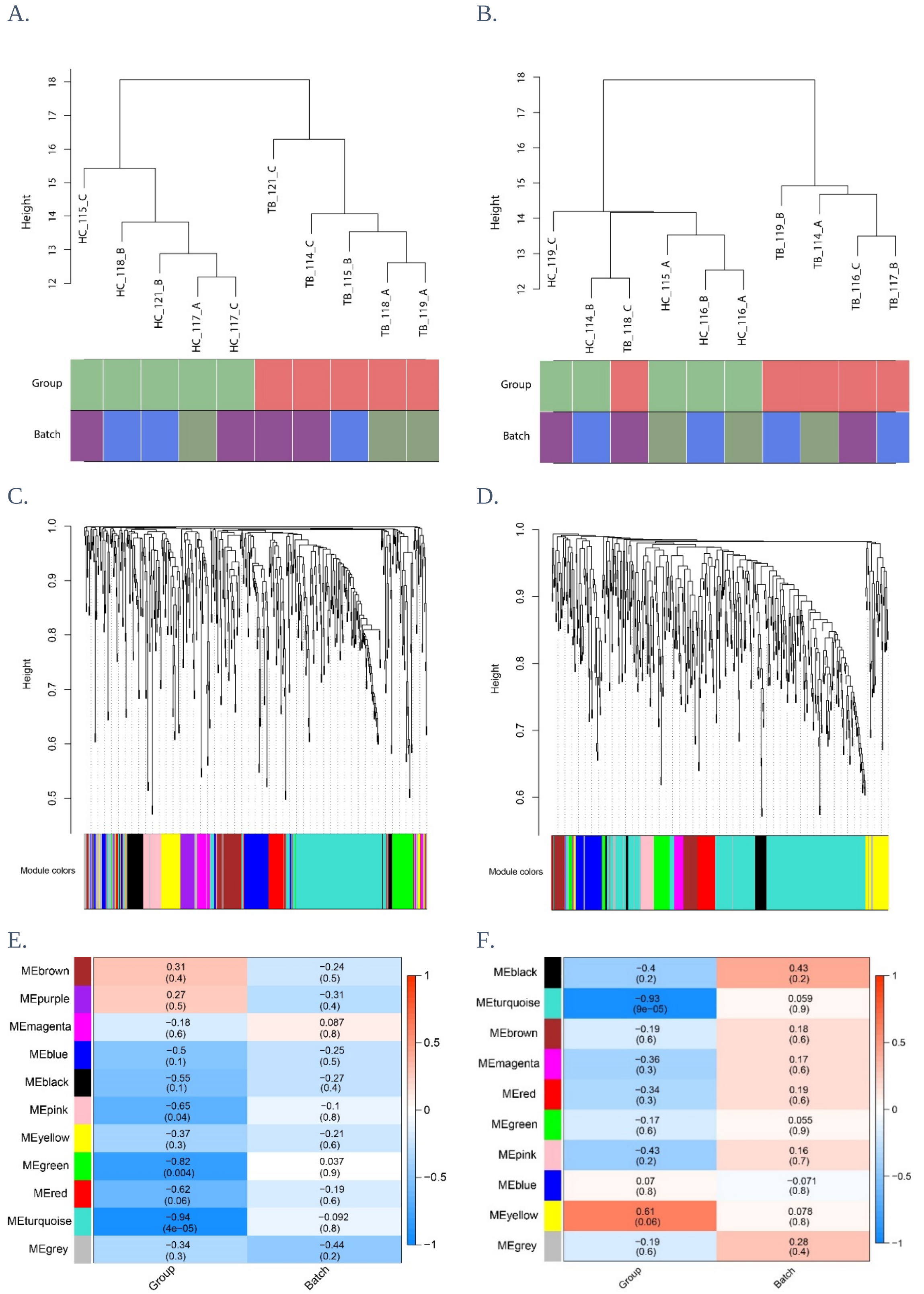
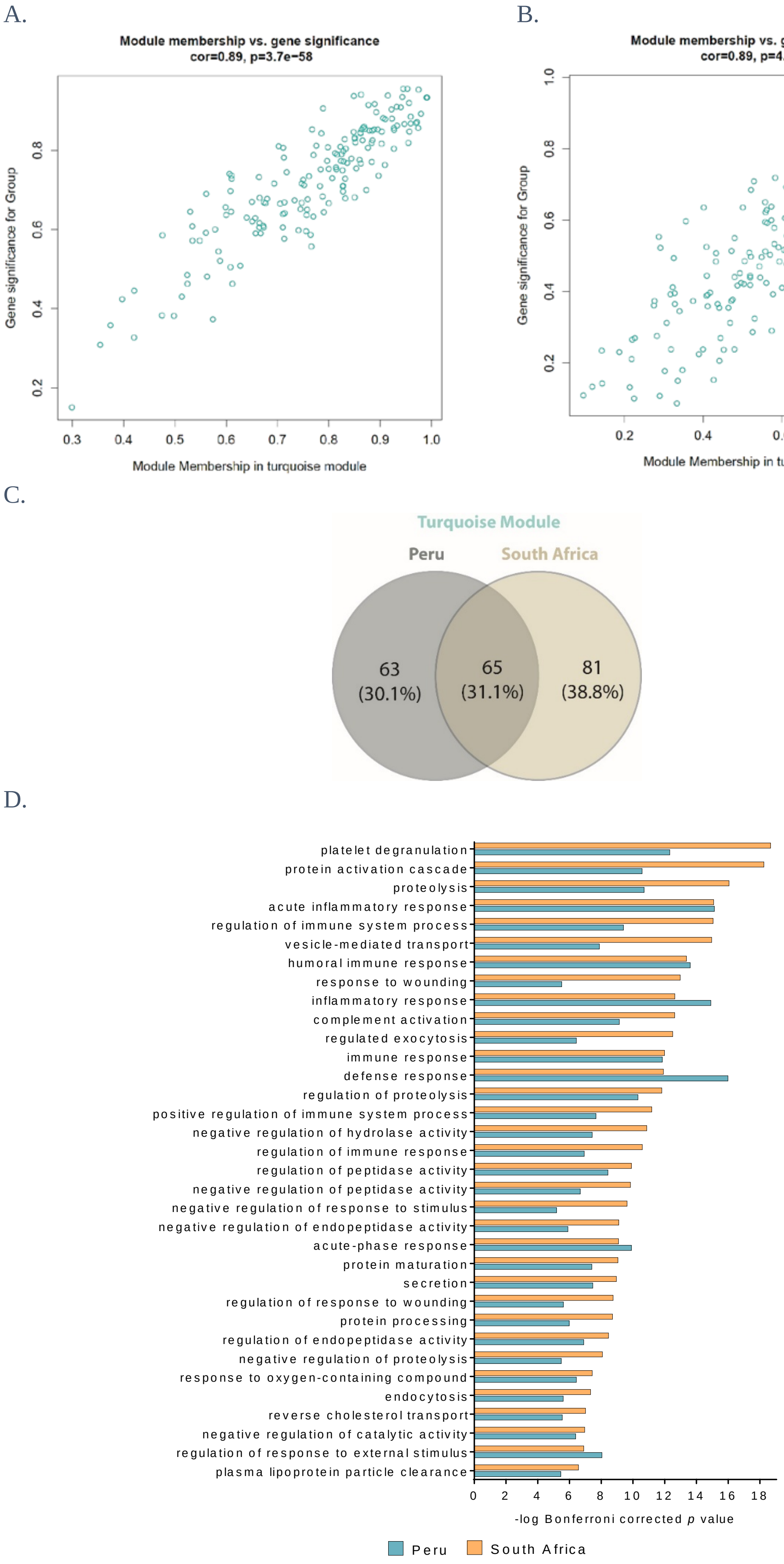


***Figure 68. Independent WGCNA analysis of Peruvian and South African samples***

*Hierarchical clustering of samples with group and batch colour coded. A. Peru and B. South Africa. Module detection of highly correlated proteins based on clustering dendograms of the topological overlap measurement (TOP) matrix for C. Peruvian and D. South African samples. Heatmap of detected modules. Colour scale indicates the correlation score. Significance of correlation in brackets E. Peru and F. South Africa.*

Analysis of highly correlated proteins identified module turquoise as the most strongly associated cluster to the disease status in both ethnicities. Further exploration of this particular module was performed in order to evaluate the commonalities and variations of the response to the tuberculosis infection in South African and Peruvian individuals. Figure 69A and B indicates a very strong correlation between the proteins from the module turquoise and the pulmonary tuberculosis infection in both ethnicities. Noticeably, the correlation score of the module and the disease status was 0.89 with a *p* value $3.7\text{x}10^{-58}$ in the Peruvian group (Figure 69A). Similarly, the correlation score in the South African group was 0.89 with a p value $4.4\text{x}10^{-94}$ (Figure 69B). A comparison of the total number of proteins included in the module turquoise discriminated by ethnicity is presented in the Venn diagram of Figure 69C. Sixty-five proteins were common between both ethnic groups, whereas 63 proteins where exclusive to the Peruvian cohort and 81 to the South African cohort. Gene ontology enrichment analysis of the proteins in each group based on biological processes was performed using the tool ToppFun from the ToppGene Suite (https://toppgene.cchmc.org/enrichment.jsp). Only Bonferroni corrected enriched terms were included in the analysis. Figure 68D illustrates the GO enrichment for the proteins common between Peruvian and South African individuals from the turquoise module. The most significant GO terms common to both groups are platelet degranulation, protein activation cascade, proteolysis, acute inflammatory response and regulation of immune system, while other biological process enriched in both datasets besides immune response were cholesterol transport and plasma lipoprotein particle clearance. Figure 69E and F present the GO enrichment based on the proteins exclusive to each group. The GO enrichment of proteins exclusive to Peru (Figure 69E) identified biological processes such as protein-lipid complex remodelling, response to bacterium and alternative pathway of complement activation. Conversely, the biological process identified for the South African group (Figure 69F) included exocytosis, extracellular matrix organisation, cell adhesion, humoral immune response mediated by circulating immunoglobulin, classical pathway of complement activation and cell migration.


A.
Module membership vs. gene significance
cor=0.89, p=3.7e−58
Gene significance for Group
0.2
0.4
0.6
0.8
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Module Membership in turquoise module
B.
Module membership vs. gene significance
cor=0.89, p=4.4e−94
Gene significance for Group
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
1.0
Module Membership in turquoise module
C.
Turquoise Module
Peru
South Africa
63
(30.1%)
65
(31.1%)
81
(38.8%)
D.
platelet degranulation
protein activation cascade
proteolysis
acute inflammatory response
regulation of immune system process
vesicle-mediated transport
humoral immune response
response to wounding
inflammatory response
complement activation
regulated exocytosis
immune response
defense response
regulation of proteolysis
positive regulation of immune system process
negative regulation of hydrolase activity
regulation of immune response
regulation of peptidase activity
negative regulation of peptidase activity
negative regulation of response to stimulus
negative regulation of endopeptidase activity
acute-phase response
protein maturation
secretion
regulation of response to wounding
protein processing
regulation of endopeptidase activity
negative regulation of proteolysis
response to oxygen-containing compound
endocytosis
reverse cholesterol transport
negative regulation of catalytic activity
regulation of response to external stimulus
plasma lipoprotein particle clearance
0
2
4
6
8
10
12
14
16
18
-log Bonferroni corrected p value
Peru
South Africa

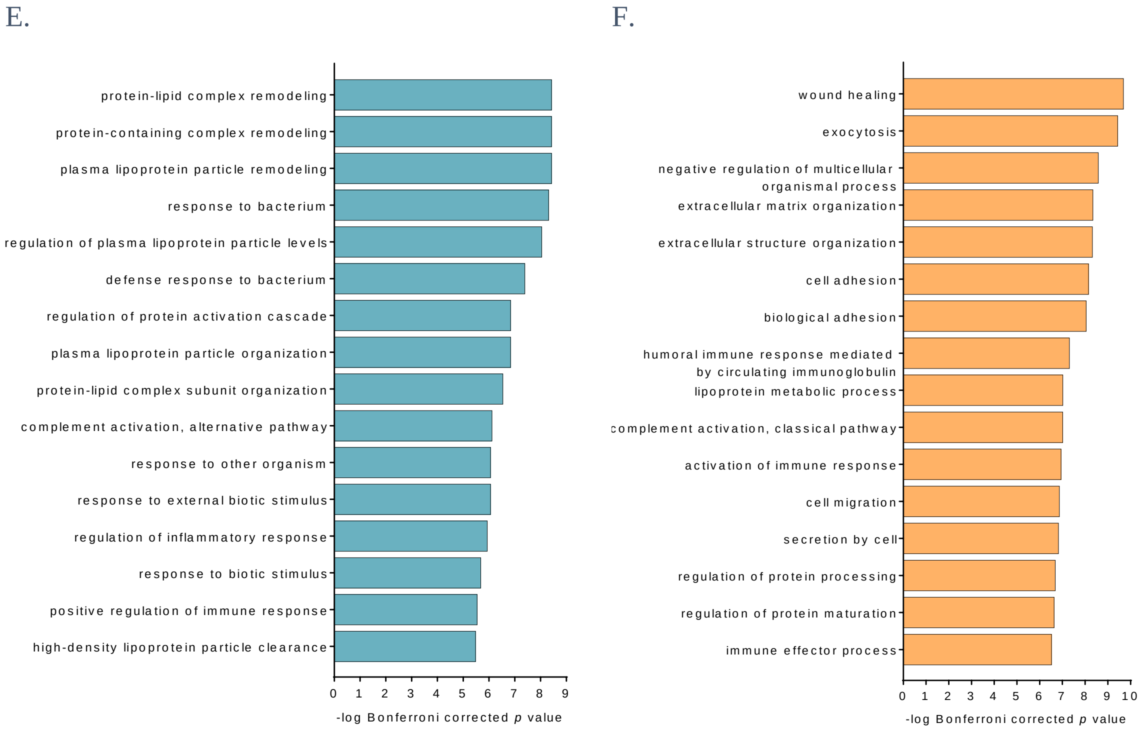
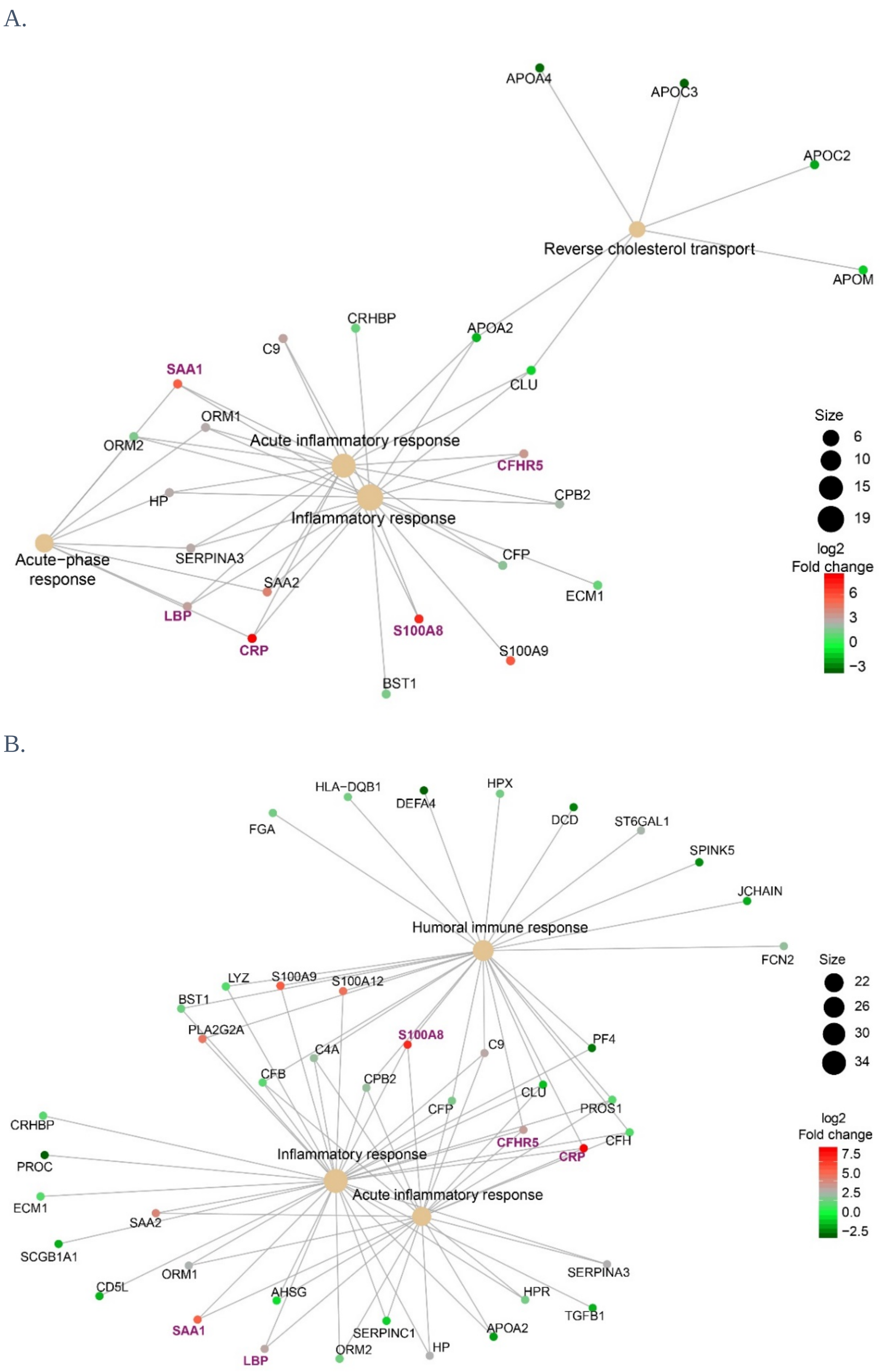


***Figure 69. Strongly correlated proteins to tuberculosis infection: module turquoise***

*Module turquoise exhibited a very strong and significant correlation to pulmonary tuberculosis infection in both ethnicities. A. Scatter plot of the correlation of proteins' significance for disease status vs. module membership for Peru and B. South Africa. C. Venn diagram of proteins included in the turquoise module comparing both ethnicities. Gene ontology enrichment analysis was performed using ToppFun from Toppgene suit https://toppgene.cchmc.org/enrichment.jsp for biological processes. Bar plots show enriched terms and significance (Bonferroni corrected p values). In blue Peruvian individuals and orange South African individuals. D. Significantly enriched biological processes common to both groups. E. Significantly enriched biological processes unique for the Peruvian group and F. South Africa.*

Further bioinformatic analysis of the proteins belonging to the turquoise module was performed using the tool ClusterProfiler 3.8.1 an R package available from Bioconductor (236). Differential profiles for each ethnicity were generated integrating the top biological processes enriched on the gene ontology analysis and the fold changes when comparing controls vs. pulmonary tuberculosis patients at relative expression protein level. Figure 70 presents the GO network with the main proteins associated to each biological process. Figure 70A the GO network common between both ethnicities. Figure 70B exhibits the profile for the Peruvian samples and, Figure 70C the profile for South African samples.


A.
APOA4
APOC3
APOC2
Reverse cholesterol transport
APOM
CRHBP
APOA2
C9
SAA1
CLU
ORM1
ORM2
Acute inflammatory response
CFHR5
HP
CPB2
Inflammatory response
Acute−phase response
SERPINA3
CFP
ECM1
SAA2
LBP
S100A8
CRP
S100A9
BST1
Size
6
10
15
19
log2 Fold change
6
3
0
−3
B.
HLA−DQB1
DEFA4
HPX
FGA
DCD
ST6GAL1
SPINK5
JCHAIN
Humoral immune response
FCN2
LYZ
S100A9
S100A12
BST1
PLA2G2A
S100A8
C4A
C9
PF4
CFB
CPB2
CLU
CFP
PROS1
CRHBP
CFHR5
CFH
PROC
Inflammatory response
CRP
Acute inflammatory response
ECM1
SAA2
SCGB1A1
ORM1
SERPINA3
CD5L
AHSG
HPR
TGFB1
SAA1
SERPINC1
APOA2
LBP
ORM2
HP
Size
22
26
30
34
log2 Fold change
7.5
5.0
2.5
0.0
−2.5

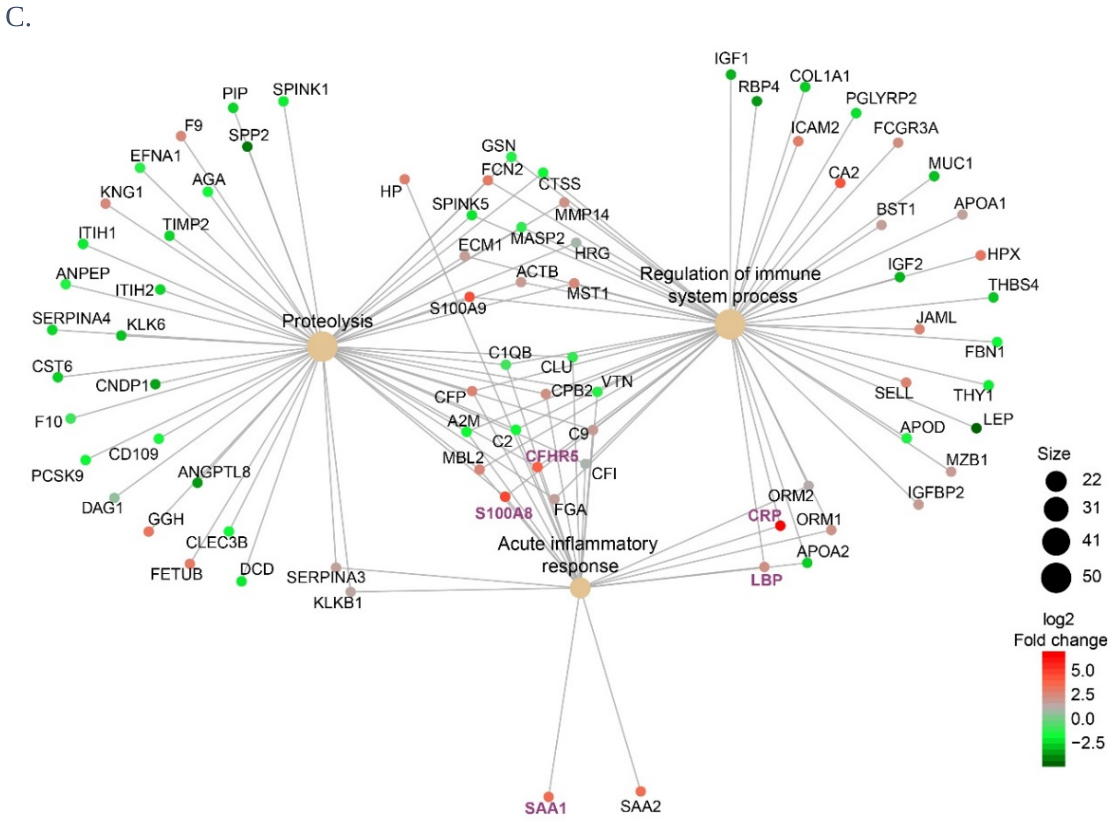


*Figure 70. Gene ontology enrichment analysis for strongly correlated proteins to tuberculosis infection*

*Gene ontology enrichment analysis of proteins strongly correlated to disease status (module turquoise). Networks show top biological processes enriched and the proteins associated to such GO terms. Circles indicates biological processes and the size the number of proteins associated to each GO term. Colour scale indicated the log2 fold change of the relative expression levels of each protein. A. Network based on common proteins between both ethnicities. Network for B. Peruvian cohort and C. South Africa. In purple proteins validated in MIMIC and South African cohorts.*

## 5.4 Discussion

Pulmonary tuberculosis remains as a devastating disease and the development of accurate, inexpensive, rapid and sputum-independent diagnostic tools suitable for resource-limited settings constitutes a critical priority in the fight against tuberculosis transmission. Diverse high-throughput strategies are being applied to detect whole or molecular traces of the *Mtb* bacilli or host responses to tuberculosis infection (103, 237). Such biomarkers or biosignatures have the potential to translate to point-of-care devices and proteins stand a greater opportunity of translation to a rapid test than other biomolecules. Many efforts are concentrated on developing diagnostic tools for tuberculosis that could yield results in minutes, such as lateral-flow or biosensor devices (122-124, 238, 239). However, novel biosignatures with higher performance than those currently available are urgently required to advance on the biomarker's pipeline. iTRAQ quantitative proteomics using the iTRAQ reagents was applied in this work to profile plasma proteins that could be used as a novel biomarkers of pulmonary tuberculosis. Chapter 4 demonstrated a powerful optimised multidimensional proteomic method based on hyper-fractionation of non-depleted whole plasma. This method delivered the most extended deep plasma proteomic coverage reported to date thus increasing the opportunities of novel biomarker discovery.

This chapter presented an extension of such a method demonstrating an increased statistical power. The most significant limitations of the results were discussed in the previous chapter. This detailed profiling was focused on the sub-proteome segment 4, which is primarily enriched for low-molecular weight proteins and protein degradation products. This fraction contains a diverse subproteome and successfully recapitulates multiple biological processes from plasma as demonstrated in a number of plasma discovery studies reported by our group in the context of cancer and nutrition (92, 240-242). In-depth profiling of segment 4 involved 10 healthy controls and 11 pulmonary tuberculosis, which according to statistical power estimations by Levin, Y., *et al.* (2011) (203) and Cohen, F. G., *et al.* (2013) (209) will provide a statistical power >0.9 when a fold change cut-off of 1.5 in logarithm base 2 scale is considered. This powerful approach resulted in the identification and independent validation of known and novel biomarkers of tuberculosis infection, significantly including the identification of a peptide *Mtb*-derived. Validation of this exciting finding will require the development of a method suitable for measurement of this protein in multiple samples, therefore it is considered in the ongoing and future work. A common plasma proteomic profile to both ethnicities was defined which can potentially be translated to diverse host contexts. Additionally, this comprehensive plasma proteomic work has potentially revealed new biological aspects of tuberculosis immunopathology associated to geographically diverse populations using a variety of bioinformatic approaches and tools to mine the datasets generated.

Samples were processed using the optimised settings defined to produce a complete plasma profile of a small sample set. In addition to the information generated for segment 4 in Chapter 4, two extra 8-plex sets were processed for this particular segment. As presented in Figure 55B and C, the experiments run subsequently were reproducible at both peptide and protein. However, these experiments were conducted a year a part from the first plasma profile and the resolution for the HCD fragmentation event was increased from 30000 to 60000 for the second round of experiments. This explains the differences between sets B-C and A in terms of total peptide/protein number and proteins identities. Although shotgun proteomics is a powerful strategy to conduct hypothesis-generating research with high-throughput, it suffers from limited reproducibility. This is mainly explained by the semi-stochastic sampling of the proteome when the number of peptides from complex biological matrixes exceeds significantly the sequencing cycles of the mass spectrometer (243). Although, the sensitivity and speed of the mass spectrometers have improved at fast pace over the last years, generating robust and reproducible quantitative data across a large number of complex samples remains as one of the greatest challenges in clinical proteomics (244).

Concomitantly with proteome subsampling, proteomics datasets are frequently incomplete compared to transcriptomic or genomic data. Particularly, data from labelling-based quantitative proteomic experiments will comprises a variable configuration of peptides totally, partially or non-quantified across the samples over an iTRAQ/TMT experimental batch. Recently, Chen L. S., *et al*. (2018) have suggested that the missing probability depends on the combined batch-level abundances and proposed linear mixed-effect strategy to model partially quantified data which is usually rejected (245). In this work, 426 fully quantified proteins at 1%FDR were common to the three iTRAQ set experiments. This stringent dataset constituted the initial input for downstream processing leading to candidate selection for validation. Further bioinformatic analysis was conducted to explore the complete datasets including missing values, which led to biologically relevant insight on the host response to *Mtb* infection. This result section will be discussed later.

Principal component analysis was performed to explore the variance distribution of the data. Figure 57A demonstrates that the experimental batch explains over 55% of the variance of the data followed by the group effect. Despite the strong batch effect on the variance, the control and tuberculosis groups were clearly distinguished between the components 2 and 3. Similarly, the master pool location for the three experiments was consistent along component 3. The heatmap in Figure 56B allows a general visualisation of the batch relative protein abundance being the abundance order B>C>A. Altogether, these data suggests that the batch effect is mainly dictated by the combined batch-level abundance in concordance with Chen L. S., *et al*. (2018) (245). Statistical analysis involving LIMMA and FDR correction for multiple testing (*PL2*) was applied to the dataset containing fully quantified common proteins across sets A – C to evaluate significant changes in relative protein abundance driven by pulmonary tuberculosis infection. At this stage, no batch effect correction was applied and differentially expressed proteins (DEPs) were defined by the regression coefficient of the linear model, which estimates the group effect. One hundred and seventy four proteins were determined as differentially expressed by this method, representing a large biosignature for pulmonary tuberculosis. Well recognised acute response reactants such as CRP (246-248) and SAA (152, 249) were identified, enhancing our confidence that the methodology has been robust, but a set of completely novel proteins were also profiled, such as Disks large homolog 4 (DLG4), Retinol-binding protein 4 (RBP4), Protein CC2D2B (CC2D2B), and Pulmonary surfactant-associated protein B (SFTPB).

As an orthogonal bioinformatic assessment to differential expression analysis correlation patterns were evaluated to identify clusters of consistently co-expressed proteins resulting from *Mtb* infection. As previously discussed in Chapter 4, differential analysis is constrained not only by small sample sizes and by individual's biological heterogeneity, but also small fold changes are difficult to detect. Co-expression analysis by Biolayout 3D express using a Person correlation cut-off of >0.75 identified two main clusters of strongly correlated proteins related with up and downregulation. This comparison provided compelling evidence of the involvement of 32 proteins (Table 11) in tuberculosis pathogenesis and thus underpinned their value as candidates for validation.

Validation of potential classifiers of pulmonary tuberculosis is of pivotal importance for the qualification of new biomarkers or signatures with diagnostic value and is very often the point that such studies fall. Well-characterised tuberculosis cohorts with complementary profiles and from geographically diverse populations have been suggested as an ideal setting for validation (237). In this work, two geographically diverse cohorts were selected for both discovery and validation. Specifically the validation cohorts included one from South Africa and one from the United Kingdom. Seven out of fifteen potential candidates were successfully validated in both cohorts when comparing healthy donors to pulmonary tuberculosis patients: IFL2, LTN, LRG1, CFHR5, LBP, SAA1 and CRP. Additionally, CFHR5, LBP, SAA1 and CRP distinguished pulmonary tuberculosis in HIV co-infected individuals. Importantly, IFL2, LTN and LRG1 have not yet been tested on the HIV co-infected cohort.

The biomarker pipeline is a challenging path that involves a stepwise process where most candidates fail to reach the bedside. One of the multiple obstacles is the required platform changes from discovery to validation and finally to field-testing (83). Liquid Chromatography hyphenated with mass spectrometry constitutes the gold standard in analytical sensitivity, specificity and selectivity by contrast to immunoassays. Translational use of the novel protein candidates in larger cohorts usually relies on antibody-based assays such as ELISA, WB, IHC, Luminex and targeted IP/LC-MS which represents a challenge to the validation of candidates. The validation rate of shotgun proteomic findings is underreported in the literature and only the number of proteomic candidates currently utilised in clinical settings can be used to estimate translation success. Case in point, strikingly only about ten protein biomarkers were approved for clinical use by the FDA between 2000 and 2012 for cancer diagnosis (250). In my work, some of the main challenges associated to biomarker validation were reflected. For instance, only two (ILF2 and LTN) out eight of the candidates selected from the complete plasma profile based on one iTRAQ 8-plex were verified on an independent validation cohort (MIMIC); RPGRIP1L, FGL1, MYOF, TNFSF11, KCNN2 and COMP failed to reproduce the differential expression detected at the discovery stage (Figures 60 and 62). This points out the critical effect of underpowered experimental designs on the false discovery rate even when strict data processing is put in place.

Equally important to statistical power, the change of platform from mass spectrometry-based discovery to antibody-dependent verification plays a key role in the detection of the differential expression required for validation. ELISA kits and Luminex arrays selected to measure SFTPB, MYOF, TNFSF11 and S100A8 had a lack of sufficient sensitivity and specificity to produce robust quantification data, even after extensive optimisation of the assays. For example, three different ELISA kits were tested to quantify SFTPB but all had insufficient reproducibility. SFTB is one of the pulmonary surfactant-associated proteins; it is critical for lamellar body packaging (251) and relatively hydrophobic. SFTPB detection in plasma is challenging and has been usually addressed using western blot (252, 253), which is impractical for quantification in large cohorts. Although

SFTB is a promising candidate from the biological perspective with strong mass spectrometric evidence, its validation is limited to the current available antibody-based platforms. This protein is an excellent candidate for validation using targeted proteomics. Particularly relevant to assays such as ELISA and Luminex is the presence of heterophilic antibodies in the patient serological sample, which may cause falsely decreased or increased detection (81). During the optimisation of the ELISA and Luminex assays, the composition of the buffer was adjusted to reduce matrix effects, based on spike-in experiments. Nevertheless, matrix effect strongly impaired the sensitivity of the Luminex 2-plex for MYOF and S100A8, an effect that was augmented by a limited antibody pairing design.

Conversely, five (LRG1, CFHR5, LBP, SAA1 and CRP) out of seven prioritised candidates included in the list of the top 32 proteins resulting from the detailed profile of segment 4 were successfully verified in one or both validation cohorts, confirming the potential of a fully powered study. Noticeably, the two candidates, which failed validation, were S100A8 and SFTB. S100A8 was reported and validated in other proteomic studies (254). It has been suggested that targeting S100A8/A9 could decrease organ injury by reducing lung tissue damage during tuberculosis immunopathology (254, 255).  The most likely reason why this particular protein was not validated in this work is a poor performance of the assay, which was unable to consistently detect the protein in plasma/serum samples. Similarly, SFTB will require development of a more sensitive and high-throughput assay as previously mentioned. Altogether, these results indicate that this optimised proteomic platform is a powerful method for biomarker discovery when sufficient statistical power is achieved accompanied by complementary bioinformatic methods.

Previous proteomic studies in this field have proposed alternative biosignatures. For example Xu, D. *et al.,* (2015) proposed the combination of S100A9, SOD3 and MMP9 with an AUC of 0.981 (119). However, validation was performed on the same discovery cohort which limits the diagnostic interpretation. Achkar, J. *et al.,* (2015) described two different signatures for pulmonary tuberculosis for negative and positive HIV coinfection, particularly for HIV negative a signature of  nine proteins (CD14, SEPP1, SELL, TNXB, LUM, PEPD, QSOX1, COMP, APOC1) with high performance (AUC=0.96) distinguishing tuberculosis infections from other respiratory conditions. The validation was conducted in limited number of samples (n=75 from which 18 and 19 samples from tuberculosis and respiratory symptomatic patients were included, respectively). The authors indicated that larger cohorts from diverse geographic regions would be required to assess the diagnostic value of this signature (117), but to date no further publications have been reported . In a more recent study, Chen, C. *et al.,* (2018) employed iTRAQ proteomics to identify a serum signature for tuberculosis co-infected with HIV, the experimental designed compared HIV positive patients negative and positive for tuberculosis coinfection (153). However, only the proteomic findings were validated and no evaluation of the accuracy or diagnostic value of the signature was conducted. Chegou, N. N. *et al.,* (2018) evaluated 12 biomarkers identified in QuantiFERON supernatants and a biosignature comprising unstimulated IFN-γ, MIP-1β, TGF-α and antigen-specific levels of TGF-α and VEGF only provided a limited AUC of 0.81 (95% CI, 0.76–0.86) (validation set n=134 individuals) (256). Furthermore, analysis of cell culture supernatants after overnight stimulation is not a valid approach in resource-poor settings.

Preliminary assessment of the diagnostic performance of the novel signature described in this work was based on the quantitative data from ELISA (ng/mL) and Luminex (Fluorescence intensity) from both validation cohorts and used to produce ROC curves and calculate AUC as an indicator of performance. The best biosignature included four proteins CFHR5, SAA1, LBP and CRP with an AUC of 0.935 (95%CI, 0.878-0.933) in the MIMIC cohort. In the South Africa cohort the performance was reduced since the CRP data was not available, AUC was 0.886 (95%CI, 0.690 – 0.961). Considering the coordinates of the curve for the signature comprising four proteins, a cut-off of 0.484 would reach 90% overall sensitivity and 87.5% specificity. In 2014 the WHO defined the key requirements for a rapid biomarker-based non-sputum-based test for detecting tuberculosis based on a survey conducted with a community of stakeholders. Among the key characteristics the minimal sensitivity should be >80% for a single test when compared with culture (for smear-negative cases it should be >60%; for smear-positive it should be 99%) and specificity >98% against a microbiological reference (25, 257). Therefore, this exploratory evaluation of the novel biosignature described in my work for pulmonary tuberculosis suggests an adequate diagnostic ability and may be potentially further developed into a diagnostic test. Importantly, CFHR5, SAA1, LBP and CRP retained differential expression in HIV co-infected patients, albeit their diagnostic ability in this population will require additional testing.

The analysis of diagnostic performance has been described in this work as preliminary since there is a number of considerations to this biosignature. The calculation of performance was based on the fluorescence intensities resulting from Luminex analysis, further development of the customised arrays produced for this particular work is required to generate quantification values with clinical applicability. ILF2 was only measured in the MIMIC cohort with an AUC of 0.826 (95%CI, 0.690 – 0.961) based on ELISA measurements. This performance for a standalone analyte is promising and makes ILF2 a good candidate for the panel. Further evaluation of this analyte in the South African cohort to explore its diagnostic value will be necessary. This work comprised the validation of a small subset of the 32 top proteins, which constitutes a valuable biosignature for tuberculosis. Verification and validation of the remaining proteins could identify new classifiers with diagnostic utility. Once this biosignature is refined will require further validation to confirm its diagnostic value. This represents a major body of work, as I have only characterised a limited number of the regulated proteins, and have not performed detailed analysis of segments 1 – 3. I aim to address some of these challenges in the future.

A final consideration to the novel biosignature is the fact that the ROC analysis compared healthy donors to pulmonary tuberculosis patients; therefore, its diagnostic value to distinguish tuberculosis from other respiratory conditions is unclear. In this respect, direct detection of *Mtb* antigens in blood would provide a highly specific test that could rule out tuberculosis infection from other respiratory conditions. However, detection of *Mtb* antigens has proved to be problematic even when highly sensitive methods such SOMAscan have been attempted (258). Chang, L. et al., (2017) successfully quantified CFP-10 and ESAT-6 in serum using antibody-labelled and energy-focusing porous discoidal silicon nanoparticles (nano-disks) followed by high-throughput mass spectrometry (MS) to enhance sensitivity and specificity (258). Although this is a significant technological step forward for tuberculosis diagnostics, discovery of novel *Mtb*-derived antigens with diagnostic value remains elusive.

The optimised plasma proteomic method developed in this work allowed me to identify one *Mtb*-derived peptide, a unique peptide from a predicted *Mtb* hydrogenase (O69693). This finding was consistent in three independent experiments and in two different ethnic groups (Figure 63). Statistical analysis of the dataset resulting from segment 4 profile and allowing missing values based on differential expression (*PL2*) identified this protein as one of the top proteins and co-expression network analysis verified this protein as strongly correlated to the disease status. Verification and validation of this promising finding will require development of a targeted proteomic assay or an antibody-based detection method so it presence can be confirmed in a wider range of samples.

As a complementary approach to the data processing and bioinformatic analysis that led to the identification of biomarker candidates for pulmonary tuberculosis, WGCNA (weighted gene co-expression network analysis) was applied on the dataset containing sets A to C and subjected to batch correction. This bioinformatic approach not only identified a cluster of proteins strongly interconnected and significantly correlated to the disease status (module turquoise) but also a confounding effect of the ethnicity variable (module brown). Correspondingly, 25% of the variance of the data is explained by the group effect but 11% is determined by the ethnicity (PC1 and PC2 respectively in PCA Figure 67). Therefore, the proteomic profile derived from each ethnicity was independently evaluated and then compared. The module turquoise was strongly correlated to tuberculosis infection in both groups and 65 proteins were identified common between them. Interestingly, a similar number of proteins were unique for each ethnicity. These results suggest that although there is a common host response to *Mtb* infection, there are specific host-pathogen interactions dictated by the geographical diversity. This complex variable not only encompasses particular features of the genetic background associated to the host, but also exposition to diverse circulating *Mtb* strains in specific geographic regions.

For instance, the *Mtb* linage distribution differs between Peru and South Africa as recently reported by Coll, F. *et al.*, (2018). Coll and colleagues undertook a large genome-wide association study (GWAS) of 6,465 *Mtb* clinical isolates from more than 30 countries. The reported linage composition in Peru (n=78) was mainly constituted by lineage 4 (92.31%) and lineage 2 (7.7%), whereas in South

Africa (n=594) such composition was more diverse; linage 2 (39%), linage 3 (2.5%) and linage 4 (57.2%) (259). Furthermore, direct interactions between mycobacterium lineage and ethnicity may occur since geo-ethnic restrictions significantly contribute to the *Mtb* lineage distributions (260). Additionally, other host's factors such as nutritional and metabolic status, exposition to polluted air and social determinants might associate to the ethnicity variable.

Gene ontology enrichment analysis of the protein expression profile in each ethnic group has pinpointed common aspects of the host response during pulmonary tuberculosis infection. Common responses involve well-known reactants of the acute-phase and inflammatory responses additionally to proteins associated to reverse cholesterol transport (Figure 70). Proteins such as CRP, LBP, SAA1, SAA2, S100A8, S100A9, CFHR5, C9, CRHBP, ORM1/2, CPB2, HP, CFP, ECM1 and SERPINA3 were upregulated. CRP, LBP, SAA1, SAA2, S100A8, S100A9, SERPINA3 and HP are involved in the activation of the acute-phase and inflammatory response and their role in tuberculosis is well recognised (150, 254, 255, 261-263). Similarly, ECM1 the extracellular matrix protein 1 has been reported upregulated in saliva (264) which is consistent with lung destructive immunopathology (52). CFHR5, C9, CFP and CPB2 are part of the complement classical and alternative activation pathways. CFHR5 co-localizes with C3, binds C3b in a dose-dependent manner, and is recruited to tissues damaged by C-reactive protein (265). Involvement of CFHR5 in tuberculosis is for the first time identified in this work. Downregulated proteins included CLU and apopoliproteins APOA4, APOC3, APOC2 and APOCM. The role of the lipids profile and cholesterol in tuberculosis immunopathology is still matter of debate. Cholesterol uptake, catabolism and broader utilization are central for maintenance of the pathogen in the host and contribute to pathogenesis and virulence. However it is not completely clear if the low circulating lipid profiles in pulmonary tuberculosis patients are merely consequence of the disease (266) or have wider biological implications. Apopoliproteins are associated to lipid transport structurally associated to lipoprotein particles such as HDL, LDL and VLDL. Deniz, O. *et al.,* (2007) demonstrated that serum HDL-C concentrations significantly and negatively correlate with the degree of radiological extent of disease and degree of smear positivity in patients with pulmonary tuberculosis (267). Downregulation of apopoliproteins is consistently reported in various serum/plasma proteomic profiles for pulmonary tuberculosis (36, 117, 118, 268) in agreement with the data generated in this work.

In summary, this chapter has described the discovery and initial validation of a novel biosignature for pulmonary tuberculosis with satisfactory diagnostic performance in independent validation cohorts. Additionally, the detection of an *Mtb*-derived peptide in plasma constitutes a very promising biomarker for a highly specific rule in test and its validation must be prioritised. Comprehensive statistical and bioinformatic tools have been applied to mine biologically relevant information from the proteomic data generated. This analysis has revealed differential host responses to *Mtb* infection resulting from ethnic diversity. This work has generated new exciting research avenues in the field of tuberculosis diagnostics and tuberculosis host-pathogen interactions. However, further analysis of the dataset, and additional diagnostic markers identified, may further increase the power of a new TB test and the biological insight into the underlying processes.

# CHAPTER 6

## Final Discussion and Future Directions

Tuberculosis was cited as "one of the most seriously neglected and underestimated health, human rights and poverty problems of our era" by UNICEF in early 2000 (269) and yet in 2016 tuberculosis, a curable disease, caused over 1.7 million deaths worldwide. One of the main challenges for tuberculosis control is the striking diagnostic gap. In 2016 the WHO reported that 39% worldwide cases went undiagnosed or unreported (1); which means that the health care of about 4.1 million people was unknown. Importantly, undiagnosed active pulmonary tuberculosis patients drive ongoing transmission in their communities.

Effective treatment is crucial to any strategy for controlling tuberculosis and biomarkers that benefit early initiation of therapy could reduce transmission. Identification of patients with tuberculosis preferably at level 0/1 of the health care systems and at the community level is a pressing need to improve diagnosis algorithms, especially in settings where chest radiography and Xpert MTB/RIF are not widely available. Although during the last 15 years several studies have been conducted to establish a tuberculosis biosignature relevant for a blood based diagnosis test that is fast, low-cost and high performance, it remains elusive. Challenges are associated with the complex biology of host-pathogen interactions taking place during the natural course of *Mtb* infection and analytical limitations of most plasma proteomic strategies, thereby hindering a universal blood proteomic biosignature.

The ultimate aim of my thesis was to identify and validate a novel plasma biosignature for pulmonary tuberculosis suitable for translation to point-of-care that can be potentially applied in different contexts of the disease. Specifically, a more universal biosignature was aimed by profiling the plasma proteome of individuals from different ethnic background and possibly exposed to different *Mtb* circulating strains. While indeed this is an ambitious target, this approach allowed the identification of common host response traits to active tuberculosis between two different ethnicities that could lead to a biosignature that recapitulates more accurately the heterogeneity of the pathology encompassing variables such as diverse geographical location and with potential adequate diagnostic performance represents valuable progress in the fight against tuberculosis. Validation of potential markers in both a South African and an UK based cohort provided additional evidence of the applicability of these plasma proteins for diagnosis of active tuberculosis in diverse host contexts.

In brief, the four objectives proposed in this thesis to contribute to this biosignature were:

**1.** To produce a comprehensive plasma proteomic profile of pulmonary tuberculosis by optimising a quantitative MudPIT (Multidimensional Protein Identification Technique) approach.

**2**. To identify a common set of biomarkers for pulmonary tuberculosis by systematic statistical analysis and complementary bioinformatic tools.

**3.** Validate the proteomic findings for active tuberculosis using independent validation cohorts.

**4.** Determine multi-marker panel performance in order to evaluate its potential for diagnostic use.

In this chapter, I aim to explore to what extent the above objectives were achieved and the contributions made by this work.

### 6.1 The most comprehensive plasma proteomic profile of pulmonary tuberculosis to date

As I presented throughout this thesis, plasma proteomics imposes significant analytical challenges as a consequence of the wide concentration dynamic range of circulating proteins. Depletion is the most common strategy to address the underrepresentation of the plasma proteome. The initial working hypothesis of this thesis was that depletion yields biased plasma profiles; whereas the most abundant proteins are pulled out, a discrete subpopulation of proteins are inadvertently co-removed (137, 138). Consequently, substitute approaches are required to best recapitulate the immunopathology driven by the tuberculosis infection reflected in the plasma proteome. Orthogonal chromatographic hyper-fractionation of plasma offers an alternative method to reduce the complexity of the matrix and still gain information from the whole spectrum of plasma protein concentrations. Hyper-fractionation provides two additional benefits: (a) Denaturing SEC plasma fractionation using 6M guanidine dissolves blood microparticles such as exosomes and denatures any additional protease activity, and thus increases the identification of proteins associated to such structures. (b) Extensive chromatographic separation has proved to reduce peptide co-isolation, which in turn contributes to alleviate the ratio-compression effect, common in iTRAQ/TMT proteomics (270).

Garbis and colleagues first reported this MudPIT approach in 2011 and in the following years, the method was evolved to generate quantitative profiles using isobaric labelling (35, 92). An early task undertaken in this work was to further optimise this method to increase its analytical capabilities. As presented in Chapter 3 and 4, the optimisation included chromatographic parameters for SEC, use of C4-based fractionation to increase orthogonality between offline and online separation of labelled peptides, development of a rapid solid phase extraction protocol and matched gradients between C4 and C18 chromatographic separations. The optimised method was applied to profile the pulmonary tuberculosis plasma proteome of eight male volunteers from Peru and South Africa by comparing healthy individuals and pulmonary tuberculosis patients.

***Contributions***

The main contributions of the optimised qMudPIT method described in this thesis are:

1. The most comprehensive plasma proteome of pulmonary tuberculosis to date.

2. An optimised protocol that produces unbiased and in-depth profiling of plasma capturing proteins throughout the dynamic range of concentration suitable for biomarker discovery.

Previous plasma proteomic studies focused on pulmonary tuberculosis are limited to a proteomic coverage between 500 and 1000 proteins, of which the vast majority are higher abundant proteins with limited clinical utility to tuberculosis, and therefore the search space for novel biomarkers is restricted from its inception. My work has identified 5022 proteins in plasma from which 3577 were fully quantified. The dynamic range of this proteome ranged 11 orders of magnitude, limited to the abundance levels currently reported in databases and literature. Additionally, 128 proteins were annotated as exosome derived. Over two hundred proteins were differentially expressed by linear modelling assessment. This is the most comprehensive plasma proteome generated to date in the context of pulmonary tuberculosis.

The limitations of this approach are mainly related to a limited throughput due to extensive offline chromatographic fractionation, which is a laborious and experimentally costly process. Additionally, the profile of the complete plasma proteome requires four independent MS experiments, which introduce batch effects and increases the cost of the method. These limitations restricted me to the analysis of seven samples for the profiling of all four segments as presented in Chapter 4 Despite these limitations, the protocol developed in this work has been used to produce extensive profiles not only from plasma but also from tissue in the context of other pathologies. For instance, our group investigated the primary cancer-associated fibroblasts in oesophageal adenocarcinoma, a study recently published in the British Journal of Cancer (271).

### 6.2 Towards a universal biosignature for pulmonary tuberculosis

Tuberculosis pathogenesis is a complex phenomenon determined by heterogeneity of the host-pathogen interaction and the nature of the immune response mounted by the host resulting in a broad spectrum of clinical manifestations, which mimics a variety of other respiratory conditions. The confirmation of the active disease solely relies on the detection of *Mtb* bacilli in sputum by smear test or culture. Although over 77 million smears are performed annually (public sector of 22 high-burden countries) (272), this test is insufficient to address the current diagnostic gap. New proteomic biosignatures for tuberculosis are reported every year in the quest for better diagnostic tests, but the commonalities across them are surprisingly small. Noticeably, the set of differentially expressed proteins profiled in this work overlap with a significant number of candidates previously reported. Annexe 6 highlights the common profiled proteins between this work and other studies. Seven out nine studies focused on pulmonary tuberculosis reported shared candidates with this work. These reports included a variety of proteomic techniques and diversity of ethnicities. This comparison suggests that the biosignature described in this thesis is more universal and encompasses candidates relevant to diverse contexts of the disease.

Chapter 5 demonstrated that an increased statistical power and complementarity of the bioinformatic tools employed for data processing was key to identify a list of 32 strong candidates and to suggest a differential host response between Peruvian and South African ethnicities.

***Contributions***

The main contributions of this thesis to the plasma proteomic exploration and discovery of novel biomarkers for proteomic are:

1. A complementary bioinformatic approach to proteomics data
2. A list of 32 strong candidates from which a subset of 8 proteins were selected for validation
3. Identification of a *Mtb*-derived peptide as a potential highly specific marker of pulmonary tuberculosis
4. Specific plasma proteomic profiles determined by ethnic diversity: Peru and South Africa

Small sample sizes, incomplete datasets and significant batch effects across experiments due to high variability of shotgun proteomic methods create difficulties in the effective detection of abundance changes from proteomic data, as discussed in Chapter 4 and 5. Bioinformatic tools to process data derived from transcriptomic and genomic datasets are more advanced to the ones available for proteomics. Often the statistical assessment of differential expression relies on Mann-Whitney test or t-test which are underpowered by small sample sizes and make assumptions of the variance of the data. Here, I compared four proposed statistical pipelines systematically including linear modelling (LIMMA) and permutation to filter out ratios followed by correction for multiple testing. Complementary, correlation network approaches such as Biolayout 3D Express and WGCNA were applied to evaluate expression patterns. This extensive analysis led to the identification of 32 proteins as strong candidates with a high rate of validation in one or two independent cohorts (6 validated proteins out of 8 candidates). Additionally, an *Mtb*-derived peptide was consistently detected and identified as potential biomarker. Blast against the complete UniProt Knowledgebase (UniProtKB) confirmed 100% identity to a predicted hydrogenase from *Mycobacterium tuberculosis* and *Mycobacterium bovis*. This suggest that this potential marker might be relevant not only for diagnostic of the human pathology but also for diagnosis of tuberculosis in cattle. According to TBFree England 277341 cattle have been culled in England since January of 2008 to control bovine tuberculosis, causing devastating financial consequences to family farming businesses (http://www.tbfreeengland.co.uk/home/). This promising finding will require further investigation to determine diagnostic value.

WGCNA network analysis suggested a differential plasma proteomic profile determined by ethnic diversity. A unique feature of the host response reflected in the proteomic profile from Peruvian individuals suggests divergent control of the antimicrobial activity; proteins directly related to microbial clearance by innate and adaptive responses were significantly downregulated. Proteins with antimicrobial activity such as defensin A4, dermcidin and serine protease inhibitor Kazal-type 5 were downregulated. Similarly, JCHAIN and PF4 had reduced expression. JCHAIN is a polypeptide secreted by plasma and mucosal cells involved in polymer formation of IgA and IgM and subsequent bacterial agglutination (273, 274). PF4, the platelet factor 4, is a chemokine secreted by platelets and binds to polyanions (P) on bacteria. Recently, P4 has been associated to opsonisation of several bacterial species (275).  Although a reduced antimicrobial response promotes bacterial persistence, control over this process might minimise the host-induced damage.

In the case of the South African individuals, the proteolysis process was differentially enriched compared to the Peruvian group. This process was highly regulated by the upregulation of proteins such as KNG1, an inhibitor of thiol proteases, FETUB, a cysteine protease inhibitor associated to the severity of lung damage in COPD (276) and MMP-14. Membrane type 1 matrix metalloproteinase-14 is key for leukocyte migration and collagen destruction. Elevated MMP-14 has been reported in induced sputum of tuberculosis patients (277). Consistently, proteins from the extracellular matrix such as EM1 and DAG1 were also upregulated. Conversely, TIMP-2 a metalloproteinase inhibitor was downregulated.  Although proteolysis is an extended process well recognised during tuberculosis pathogenesis, the enrichment of this process in the South African cohort might be associated to the pathology severity. A more profound understanding of the determinants of the host response to *Mtb* infection would benefit the design of better diagnostic tools and therapeutic treatments. For instance, personalisation of the interventions to specific populations might improve the clinical outcomes.

The main limitation of the signatures generated in this thesis was that the analysis was entirely based on male individuals and although the pathogenesis of tuberculosis exhibits sexual dimorphism biased towards the male population, these biosignatures will require validation on cohorts including both genders. Additionally, two of the three cohorts used (South African cohort and the UK-based MIMIC cohort) in this study were part of a retrospective study, which limited the selection of the groups described in Table 4 to pre-established clinical features such as age, BMI and smoking status.

### 6.3 A novel multi-marker panel for pulmonary tuberculosis diagnosis

A critical aspect of the biomarker pipeline is the adequate validation of the candidate differential expression established during the discovery stage. In the case of diagnostic biomarkers, the validation cohort must be independent to the discovery group, well clinically annotated and the cases defined by the gold standard test. In this thesis, a subset of 15 proteins were selected for validation in independent cohorts using ELISA and/or LUMINEX assays. Additionally, the potential diagnostic performance was explored by ROC analysis and logistic regression for combined classifiers.

***Contributions***

1. Validation of seven proteins in one or two independent validation cohorts including completely novel biomarkers for pulmonary tuberculosis: ILF2, LTN and CFHR5.
2. Potential multi-marker panel with AUC of 0.935 (95%, 0.878 – 0.993)

From the subset of proteins subjected to ELISA or Luminex measurement, seven proteins were successfully validated. LBP, CFHR5 and SAA were significantly upregulated in both cohorts (320 samples). ILF2, LTN and CFHR5 were novel findings. The interleukin enhancer-factor 2 functions mainly as a heterodimeric complex with ILF3, and may be involved in T-cell activation. A recent report identified ILF2 as a potential biomarker in paediatric tuberculosis by bioinformatic mining of publicly available gene expression datasets (278). There are no reports or patents available describing complement factor H-related protein (CFHR5) or E3 ubiquitin-protein ligase listerin (LTN) in tuberculosis to date.

Evaluation of the performance of these markers by ROC analysis demonstrated that the multi-marker panel comprising LBP, CFHR5, CRP and SAA exhibited an AUC of 0.935 (95%, 0.878 – 0.993) in the UK cohort (MIMIC). Although this is a promising diagnostic performance, it is necessary to consider some limitations to the current validation state of these markers in this thesis. The ROC analysis was based on fluoresce intensities, and so further development of Luminex assays are required to improve quantification accuracy of these analytes. The concentration levels expressed in ng/mL are only available for the analytes measured by ELISA. These were not combined with the results produced by Luminex for calculation of the ROC curves. Validation in both cohorts was limited by sample exhaustion. The samples used for validation were part of a retrospective study and were biased towards male population, particularly in the South African cohort. However, this reflects the incidence distribution of the disease in Durban. The analysis presented in this work for the diagnostic performance is preliminary, limited to advance tuberculosis pathology in adults and will require further validation.

Summarising the main contributions of this work, I identified entirely new biomarkers of active tuberculosis including both host and pathogen derived analytes. These findings validate the proteomic and bioinformatic approaches used throughout this work but equally raises further questions about taking the work further as a diagnostic tool in the field. Ultimately this is the goal of this research and it should be regarded as determinant of success. In addition to developing a diagnostic, my work has given new information on the basic biological processes underlying tuberculosis pathology. For example, I demonstrated activation of specific immune response mechanisms such as the complement system and major changes in the lipid pathways once again. Additionally, some specific features of the immune response in Peruvian and South African individuals were proposed. Although the data I found is promising, further experimental confirmation is required to consolidate my findings and confirm the biological relevance. Lastly, I have further refined a methodological platform which has applicability to a wide range of human diseases. Therefore from this point, I am keen to pursue further work to advance my findings and translate to a new diagnostic test and better understanding of human tuberculosis. I outline the future work I propose below.

### 6.4 Future directions

My PhD has generated a large proteomic dataset from patients with tuberculosis and healthy controls, and it is inevitable that considerable further analysis of the dataset can be undertaken from both a biomarker and biological perspective. For example, I have not analysed the majority of the 32 biomarkers that are divergently regulated from detailed analysis of segment 4, due to time, financial and sample exhaustion considerations. Similarly, further analysis of the biological sub-pathways and effect of ethnicity can be performed. I have performed analysis showing the validity of the approach, by validating a significant number of the markers that I first identified, but am fully aware that significant further work is required to build on my thesis.

Ongoing and future work towards the design of multi-marker panel suitable for point-of-care are outlined below.

- We are currently working on the detailed analysis of the remaining segments 1, 2 and 3, which will increase the statistical power of the complete plasma proteome generating the most extensive plasma profile of pulmonary tuberculosis. This will expand our options to find novel biomarkers and potential *Mtb*-derived proteins.

- I will develop a method to validate the *Mtb*-derived peptide (O69693) through targeted proteomics or antibody-based assays. A polyclonal antibody against this peptide has been already produced in rabbit. If we can validate this finding, we feel this has great potential as a near-patient rule in test similar to urinary LAM in HIV co-infected individuals, and will have significant translational potential. The antibodies generated are being used to confirm the presence of the protein O69693 in *Mycobacterium tuberculosis* using free cell lysates from different *Mtb* strains and in plasma from individual with active tuberculosis by immunoprecipitation followed by western blot and/or mass spectrometric analysis.

- I will continue validating the currently discovered markers to complete both cohorts and generate quantitative data to produce a more accurate estimation of the diagnostic performance of the proposed biosignature. Additionally, the clinical relevance of the general biosignature proposed in this work needs to be evaluated in different clinical contexts. Preliminary results of this part of the study will be used to prepare a “diagnostic” paper.

- Further bioinformatics analysis of the biological pathways divergently modulated by tuberculosis will be complemented with immunohistochemical analysis of TB lymph nodes, to form the "biological" paper.

- Specific features of the plasma proteomic profiles resulting from active tuberculosis in Peruvian and South African cohorts suggested in this work may be further evaluated, including *Mtb* strain characterisation and specific ethnicities to explore the role of the host genetic background and exposure to *Mtb* strain in the immunopathology of the disease.

An MRC Confidence in concept (CiC) grant has been awarded with the aim to generate further evidence of the diagnostic value of the mycobacterium protein O69693, and this will be the focus of my laboratory work in the autumn.

### 6.5 Publications and participation in congresses

- UK Patent 1719565.2: This patent protects the intellectual property on the potential diagnostic biomarkers for pulmonary tuberculosis: CHFR5, SFTPB, ILF2 and O69693.

- Participation in congress:

  2018 Keystone Symposia meeting on "Tuberculosis: Translating Scientific Findings for Clinical and Public Health Impact" – Whistler, British Columbia, Canada.
  **Poster:** *Comprehensive Plasma Proteomic Profiling Reveals Novel Biomarkers of Pulmonary Tuberculosis*

  2017 16th Human Proteome Organisation World Congress – Dublin, Ireland
  **Short oral presentation:** *An optimised quantitative Multidimensional Protein Identification Technology for in-depth profiling of the human plasma proteome.*

  2015 British Society for Proteome Research 2015 Meeting: Capturing the Proteome in Space and Time- the 4th Dimension – Reading, UK
  **Poster:** *Quantitative Multidimensional Protein Identification Technology (qMudPIT) for discovery of plasma biomarkers of tuberculosis.*

- Co-author of other proteomic studies using the optimised qMudPIT approach described in this work

  Al-Daghri, N. M., Manousopoulou, A., Alokail, M. S., Yakout, S., Alenad, A., **Garay-Baquero, D. J.**, Fotopoulos, M., Teng, J., Al-Attas, O., Al-Saleh, Y., Sabico, S., Chrousos, G. P., Garbis, S. D. (2018). "Sex-specific correlation of IGFBP-2 and IGFBP-3 with vitamin D status in adults with obesity: a cross-sectional serum proteomics study." Nutr Diabetes 8(1): 54.

  Antigoni Manousopoulou, Annette Hayden, Massimiliano Mellone, **Diana J. Garay-Baquero**, Cory H. White, Fergus Noble, Monette Lopez, Gareth J. Thomas, Timothy J. Underwood & Spiros D. Garbis (2018). "Quantitative proteomic profiling of primary cancer-associated fibroblasts in oesophageal adenocarcinoma." Br J Cancer 118(9): 1200-1207.

Bashar Zeidan, Antigoni Manousopoulou, **Diana J. Garay-Baquero**, Cory H. White, Samantha E. T. Larkin, Kathleen N. Potter, Theodoros I., Roumeliotis, Evangelia K. Papachristou, Ellen Copson, Ramsey I. Cutress, Stephen A. Beers, Diana Eccles, Paul A. Townsend and Spiros D. Garbis (2018). "Increased circulating resistin levels in early-onset breast cancer patients of normal body mass index correlate with lymph node negative involvement and longer disease free survival: a multi-center POSH cohort serum proteomics study." Breast Cancer Res 20(1):19.

Brace, P. T., Tezera, L. B., Bielecka, M. K., Mellows, T., **Garay, D.**, Tian, S., Rand, L., Green, J., Jogai, S., Steele, A. J., Millar, T. M., Sanchez-Elsner, T., Friedland, J. S., Proud, C. G., Elkington, P. T. (2017). "Mycobacterium tuberculosis subverts negative regulatory pathways in human macrophages to drive immunopathology." PLoS Pathog 13(6): e1006367.

# ANNEXES

## Annexe 1. Proteins significantly modulated ($p < 0.05$) common to preliminary experiments 1 and 2.

Fold-changes are presented as $log_2 TB_x - \frac{1}{4}\sum_{i=1}^{4} log_2\, C_i$. FC_01: Fold-change experiment 1 and FC_02: Fold-change experiment 2. Proteins presented in descendant fold-change magnitude.

| UniProt ID | Protein name | FC_01 | FC_01 | FC_01 | FC_01 | FC_02 | FC_02 | FC_02 | FC_02 |
|---|---|---|---|---|---|---|---|---|---|
| | | 117 | 118 | 119 | 121 | 117 | 118 | 119 | 121 |
| P07988 | Pulmonary surfactant-associated protein B | 3.754 | 1.655 | 0.687 | 3.567 | 3.368 | 1.767 | 0.619 | 3.309 |
| P05109 | Protein S100-A8 | 2.981 | 5.043 | 1.139 | 4.349 | 2.661 | 4.758 | 0.971 | 4.163 |
| P02750 | Leucine-rich alpha-2-glycoprotein | 2.804 | 1.434 | 3.202 | 2.845 | 2.848 | 1.580 | 3.371 | 2.885 |
| P02741 | C-reactive protein | 2.724 | 3.465 | 1.907 | 3.219 | 2.593 | 3.187 | 1.987 | 3.088 |
| P0DJI9 | Serum amyloid A-2 protein | 2.692 | 3.018 | 1.331 | 3.302 | 2.385 | 2.620 | 1.376 | 2.972 |
| P06702 | Protein S100-A9 | 2.392 | 4.329 | 0.867 | 3.763 | 1.898 | 3.631 | 0.823 | 3.195 |
| Q92496 | Complement factor H-related protein 4 | 1.611 | 0.343 | 2.125 | 1.883 | 1.869 | 0.473 | 2.580 | 2.137 |
| P0DJI8 | Serum amyloid A-1 protein | 1.466 | 1.741 | 0.829 | 1.645 | 1.394 | 1.542 | 1.051 | 1.553 |
| Q9BXR6 | Complement factor H-related protein 5 | 1.449 | 1.650 | 1.653 | 0.927 | 1.856 | 1.888 | 2.272 | 1.493 |
| P78352 | Disks large | 1.412 | 2.501 | 0.730 | 1.963 | 1.738 | 2.510 | 1.416 | 2.029 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | homolog 4 | | | | | | | | |
| P02763 | Alpha-1-acid glycoprotein 1 | 1.381 | 2.447 | 0.710 | 1.913 | 1.460 | 2.373 | 0.959 | 1.941 |
| P08571 | Monocyte differentiation antigen CD14 | 1.293 | 0.909 | 1.103 | 1.560 | 1.451 | 1.092 | 1.462 | 1.791 |
| P52566 | Rho GDP-dissociation inhibitor 2 | 1.249 | 1.713 | 1.913 | 1.718 | 1.512 | 1.845 | 2.366 | 1.969 |
| Q5TFQ8 | Signal-regulatory protein beta-1 isoform 3 | 1.178 | 1.513 | 0.356 | 1.188 | 1.446 | 1.646 | 0.810 | 1.445 |
| P19652 | Alpha-1-acid glycoprotein 2 | 1.032 | 1.816 | 0.291 | 1.047 | 1.087 | 1.768 | 0.518 | 1.131 |
| P60660 | Myosin light polypeptide 6 | 0.966 | 1.445 | 2.378 | 2.449 | 1.549 | 1.923 | 3.194 | 2.966 |
| Q05682 | Caldesmon | 0.879 | 1.915 | 2.496 | 2.800 | 1.140 | 2.044 | 2.948 | 3.047 |
| P31146 | Coronin-1A | 0.834 | 1.624 | 1.938 | 2.279 | 1.093 | 1.754 | 2.387 | 2.526 |
| Q15485 | Ficolin-2 | 0.709 | 1.177 | 0.172 | 0.826 | 0.445 | 0.967 | 0.381 | 0.725 |
| P02790 | Hemopexin | 0.706 | 0.277 | 0.575 | 0.705 | 0.881 | 0.436 | 0.907 | 0.930 |
| P62328 | Thymosin beta-4 | 0.691 | 1.335 | 1.496 | 2.011 | 0.821 | 1.280 | 1.895 | 2.285 |
| P62937 | Peptidyl-prolyl cis-trans isomerase A | 0.522 | 1.110 | 1.285 | 1.601 | 0.788 | 1.239 | 1.738 | 1.852 |
| O00151 | PDZ and LIM domain protein 1 | 0.505 | 0.897 | 1.886 | 1.898 | 0.767 | 1.028 | 2.338 | 2.147 |
| P02766 | Transthyretin | -0.189 | -0.876 | -1.454 | -0.875 | -0.167 | -0.894 | -1.200 | -0.763 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| P02654 | Apolipoprotein C-I | -0.236 | -1.390 | -1.108 | -0.914 | -0.366 | -1.345 | -1.003 | -0.918 |
| P02656 | Apolipoprotein C-III | -0.563 | -2.107 | -1.368 | -0.672 | -0.576 | -1.899 | -1.085 | -0.587 |
| P02647 | Apolipoprotein A-II | -0.665 | -1.497 | -1.017 | -1.020 | -0.356 | -1.153 | -0.493 | -0.639 |
| Q13103 | Secreted phosphoprotein 24 | -0.711 | -1.474 | -1.513 | -1.897 | -1.060 | -1.735 | -1.973 | -1.612 |
| P02753 | Retinol-binding protein 4 | -0.881 | -1.655 | -1.858 | -1.072 | -0.770 | -1.500 | -1.467 | -0.908 |
| P02652 | Apolipoprotein A-II | -0.907 | -2.013 | -1.553 | -1.203 | -0.768 | -1.837 | -1.204 | -0.995 |
| O95445 | Apolipoprotein M | -0.996 | -0.787 | -1.242 | -1.169 | -0.863 | -0.757 | -0.863 | -0.930 |
| P29622 | Kallistatin | -1.096 | -2.210 | -1.029 | -1.011 | -0.897 | -1.190 | -0.766 | -1.053 |

## Annexe 2. Summary of proteins significantly regulated from independent analysis of SEC segments

*p* values were calculated from PL2– Linear modelling on filtered ratios by permutation. FDR correction for multiple testing.

* Proteins significantly regulated resulting from the four pipelines

HC:TB indicates the proportion of quantified samples in each group when comparing for missingness patterns.

### Segment 1

| No. | Protein Accession | Name | logFC | P.Value | adj.P.Val | All Pipelines | HC_115 | HC_116 | HC_117 | TB_118 | TB_119 | TB_114 | TB_121 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Q68CZ1 | Protein fantom | -2.69 | 6.42E-05 | 8.15E-03 | * | 5187.48 | 4108.07 | 6563.26 | 663.12 | 690.48 | 685.00 | 1348.69 |
| 2 | A6NNF4 | Zinc finger protein 726 | -2.56 | 6.68E-04 | 4.24E-02 |  | 4642.59 | 2149.25 | 2312.11 | 579.69 | 331.08 | NA | 586.38 |
| 3 | Q6NSJ2 | Pleckstrin homology-like domain family B member 3 | 1.78 | 1.50E-03 | 4.95E-02 |  | 178.34 | NA | 225.64 | 823.02 | 823.80 | 726.09 | 454.63 |
| 4 | Q6P1X5 | Transcription initiation factor TFIID subunit 2 | -1.48 | 9.34E-03 | 1.19E-01 | * | 10142.94 | 6605.48 | 4717.68 | 2107.03 | 2471.38 | 4426.17 | 1530.05 |
| 5 | Q8N9V6 | Ankyrin repeat domain-containing protein 53 | 1.59 | 1.05E-02 | 1.19E-01 | * | 1339.86 | 985.40 | 1497.29 | 3229.04 | 4368.74 | 1897.05 | 7637.10 |
| 6 | P62070 | Ras-related protein R-Ras2 | -1.41 | 1.06E-02 | 1.19E-01 |  | NA | 1926.64 | 2161.55 | 1123.61 | 605.43 | NA | 668.67 |
| 7 | Q76LX8 | A disintegrin and metalloproteinase with thrombospondin motifs 13 | -0.96 | 1.63E-02 | 1.19E-01 |  | 13464.90 | 10488.39 | 12073.05 | 5433.98 | 5718.24 | 4470.94 | 10138.16 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | Q9UMZ2 | Synergin gamma | 3.62 | 1.82E-02 | 1.19E-01 | | 927.25 | NA | NA | NA | NA | NA | 11425.50 |
| 9 | O75882 | Attractin | -0.56 | 1.93E-02 | 1.19E-01 | * | 6529.68 | 5509.25 | 5478.98 | 3821.97 | 3561.03 | 3910.56 | 4505.66 |
| 10 | P54802 | Alpha-N-acetylglucosaminidase | -1.37 | 2.18E-02 | 1.19E-01 | * | 7696.40 | 9374.54 | 3548.73 | 2061.38 | 1817.64 | 2006.43 | 4822.39 |
| 11 | P43251 | Biotinidase | -0.83 | 2.33E-02 | 1.19E-01 | * | 4719.11 | 3481.00 | 2292.07 | 1927.03 | 1867.05 | 1572.58 | 2257.34 |
| 12 | P36784 | Regulatory protein E2 | -1.62 | 2.42E-02 | 1.19E-01 | | 1554.16 | 5963.60 | 2774.26 | 1133.11 | 1476.32 | 440.45 | 1144.06 |
| 13 | A6NES4 | Maestro heat-like repeat-containing protein family member 2A | 3.12 | 2.49E-02 | 1.19E-01 | | NA | 921.22 | NA | NA | NA | NA | 7988.55 |
| 14 | Q8TCS8 | Polyribonucleotide nucleotidyltransferase 1, mitochondrial | -1.71 | 2.51E-02 | 1.19E-01 | | NA | 1308.39 | 1041.83 | 356.03 | NA | NA | NA |
| 15 | P01042 | Kininogen-1 | -0.59 | 2.63E-02 | 1.19E-01 | * | 6822.19 | 7775.17 | 6576.49 | 4754.01 | 3952.55 | 4379.39 | 5880.65 |
| 16 | O15144 | Actin-related protein 2/3 complex subunit 2 | -1.63 | 2.63E-02 | 1.19E-01 | | 1740.58 | 1531.13 | 1615.42 | 621.77 | 305.84 | 272.19 | 1489.82 |
| 17 | P9WGI5 | RNA polymerase sigma factor SigB | -1.52 | 2.97E-02 | 1.25E-01 | | NA | 6636.68 | NA | 2207.63 | 2438.59 | NA | NA |
| 18 | P18428 | Lipopolysaccharide-binding protein | 1.44 | 3.10E-02 | 1.25E-01 | * | 888.06 | 2098.30 | 1379.88 | 4025.43 | 4263.26 | 7143.06 | 1574.97 |
| 19 | P01770 | Immunoglobulin heavy variable 3-30 | 0.97 | 3.16E-02 | 1.25E-01 | | 3970.80 | 3412.05 | 2033.71 | 5189.06 | 9677.38 | 5702.63 | 4238.49 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | Q96KN2 | Beta-Ala-His dipeptidase | -0.66 | 3.60E-02 | 1.34E-01 | * | 5857.54 | 5271.87 | 6284.95 | 2807.17 | 3221.23 | 3918.54 | 5148.63 |
| 21 | P03951 | Coagulation factor XI | -0.71 | 3.70E-02 | 1.34E-01 | | 4201.65 | 7411.54 | 5747.75 | 4670.83 | 3390.74 | 2890.60 | 3049.21 |
| 22 | A3DSK5 | RNA-directed RNA polymerase | 1.52 | 4.22E-02 | 1.45E-01 | | 10583.02 | 14179.18 | 14513.12 | 74232.61 | 62902.45 | 34053.91 | 12074.35 |
| 23 | P09172 | Dopamine beta-hydroxylase | -1.00 | 4.22E-02 | 1.45E-01 | | 6219.51 | 1958.56 | 2807.69 | 1243.46 | 1621.38 | 1750.45 | 1963.66 |
| 24 | P01008 | Antithrombin-III | -0.52 | 4.39E-02 | 1.47E-01 | | 3499.77 | 3118.86 | 3277.18 | 2308.22 | 2094.64 | 1939.61 | 2966.76 |
| 25 | P06396 | Gelsolin | -0.54 | 4.80E-02 | 1.55E-01 | | 4742.75 | 4589.95 | 5250.71 | 2983.46 | 3325.72 | 2799.63 | 4494.09 |
| 26 | P43652 | Afamin | -0.79 | 4.98E-02 | 1.55E-01 | | 3867.50 | 3509.83 | 3608.52 | 2105.97 | 1597.29 | 1600.75 | 3774.10 |
| 27 | P62937 | Peptidyl-prolyl cis-trans isomerase A | 4TB:1HC | | | | | | x | | | | |
| 28 | Q8IZF2 | Adhesion G protein-coupled receptor F5 | 4TB:1HC | | | | | x | | | | | |
| 29 | Q14194 | Dihydropyrimidinase-related protein 1 | 1TB:3HC | | | | | | | | | x | |
| 30 | Q68798 | Genome polyprotein | 1TB:3HC | | | | | | | | | x | |
| 31 | Q9ULJ1 | Outer dense fiber protein 2-like | 1TB:3HC | | | | | | | | | x | |
| 32 | Q9UPN9 | E3 ubiquitin-protein ligase TRIM33 | 1TB:3HC | | | | | | | | | x | |

## Segment 2

| No. | Protein Accession | Name | logFC | P.Value | adj.P.Val | All Pipelines | HC_115 | HC_116 | HC_117 | TB_118 | TB_119 | TB_114 | TB_121 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Q9Y5W7 | Sorting nexin-14 | -1.74 | 7.01E-04 | 7.71E-02 | * | 4774.70 | 4591.66 | 5962.30 | 2213.60 | 1461.01 | 984.90 | 1659.72 |
| 2 | Q96A44 | SPRY domain-containing SOCS box protein 4 | -1.71 | 3.05E-03 | 1.10E-01 |  | 2143.94 | 2421.84 | 4386.04 | 1141.37 | 773.66 | 539.98 | 1181.59 |
| 3 | P04180 | Phosphatidylcholine-sterol acyltransferase | -0.82 | 4.71E-03 | 1.10E-01 | * | 6185.98 | 4905.51 | 5500.85 | 3540.67 | 2523.88 | 3113.48 | 3356.73 |
| 4 | P02749 | Beta-2-glycoprotein 1 | -0.70 | 1.17E-02 | 1.10E-01 |  | 4609.33 | 4705.67 | 5193.47 | 2905.42 | 2820.18 | 2500.53 | 3816.25 |
| 5 | Q96N23 | Cilia- and flagella-associated protein 54 | -1.84 | 1.43E-02 | 1.10E-01 |  | 1806.60 | 5769.16 | 9604.23 | 1274.60 | 1315.92 | 935.95 | 1814.16 |
| 6 | Q96PW8 | Putative uncharacterized protein KIAA1920 | -1.32 | 1.72E-02 | 1.10E-01 |  | 1960.01 | 1497.16 | 996.88 | 881.88 | 702.39 | 296.67 | 578.09 |
| 7 | O43610 | Protein sprouty homolog 3 | -1.79 | 2.04E-02 | 1.10E-01 |  | 28838.84 | 97387.59 | 78841.52 | 26213.82 | 32191.35 | 10027.40 | 11160.13 |
| 8 | P0A5T9 | Phosphoribosylformylglycinamidine synthase subunit PurL | 4.15 | 2.06E-02 | 1.10E-01 |  | 54.51 | 276.35 | NA | 553.45 | NA | 3687.36 | 5025.27 |
| 9 | P07358 | Complement component C8 beta chain | 0.70 | 2.17E-02 | 1.10E-01 | * | 2342.02 | 2474.65 | 1893.64 | 4461.79 | 4069.51 | 3372.17 | 2741.54 |
| 10 | P00488 | Coagulation factor XIII A chain | -0.47 | 2.44E-02 | 1.10E-01 | * | 5008.52 | 4959.16 | 4384.30 | 3218.07 | 3457.07 | 3586.55 | 3508.16 |
| 11 | P06276 | Cholinesterase | -0.88 | 2.50E-02 | 1.10E-01 |  | 5974.81 | 3470.62 | 6447.31 | 2698.23 | 2169.84 | 2641.73 | 3874.77 |
| 12 | Q9NZM1 | Myoferlin | -1.56 | 2.55E-02 | 1.10E-01 |  | 936.08 | 572.84 | 614.04 | 233.84 | NA | NA | NA |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 13 | P04431 | Immunoglobulin kappa variable 1-39 | -1.22 | 3.13E-02 | 1.10E-01 | | 7392.10 | 15611.39 | 11965.89 | 6628.49 | 6049.94 | 5736.08 | 2287.72 |
| 14 | Q9H6F5 | Coiled-coil domain-containing protein 86 | 1.41 | 3.15E-02 | 1.10E-01 | | 3848.19 | 6120.37 | 5455.88 | 16814.25 | 21986.90 | 17344.49 | 5053.55 |
| 15 | P02748 | Complement component C9 | 0.81 | 3.22E-02 | 1.10E-01 | * | 3005.98 | 3774.63 | 2922.59 | 6192.79 | 7438.02 | 6301.62 | 3438.64 |
| 16 | Q14568 | Heat shock protein HSP 90-alpha A2 | -2.69 | 3.42E-02 | 1.10E-01 | | NA | NA | 3647.77 | 519.68 | 575.96 | 1374.22 | 246.64 |
| 17 | P0C6H4 | Capsid protein | -1.61 | 3.49E-02 | 1.10E-01 | | 4844.51 | 1479.75 | 1247.90 | 483.13 | 473.04 | 607.87 | 1537.46 |
| 18 | P00738 | Haptoglobin | 1.08 | 3.50E-02 | 1.10E-01 | | 1162.28 | 2218.58 | 1777.31 | 4557.44 | 5223.54 | 3466.69 | 1866.20 |
| 19 | Q9IDV9 | Gag-Pol polyprotein | 2.25 | 3.55E-02 | 1.10E-01 | * | 787.91 | 9474.39 | 4387.00 | 20593.31 | 23270.90 | 16050.56 | 7008.77 |
| 20 | P05362 | Intercellular adhesion molecule 1 | 1.80 | 3.69E-02 | 1.10E-01 | * | 3039.91 | 1321.10 | 1062.57 | 1681.62 | 8669.87 | 10067.40 | 7015.53 |
| 21 | P02790 | Hemopexin | -0.47 | 3.74E-02 | 1.10E-01 | | 5033.68 | 6187.09 | 5752.66 | 4442.90 | 4433.83 | 3833.80 | 3602.33 |
| 22 | P01854 | Ig epsilon chain C region | 1.74 | 3.79E-02 | 1.10E-01 | | 791.95 | 1866.04 | 1285.18 | 2081.98 | 2251.36 | 4951.11 | 12665.74 |
| 23 | P01780 | Immunoglobulin heavy variable 3-7 | -1.17 | 4.06E-02 | 1.10E-01 | | 21948.64 | 10879.23 | 5962.90 | 3479.90 | 6981.36 | 4695.90 | 5425.57 |
| 24 | P04196 | Histidine-rich glycoprotein | -0.74 | 4.19E-02 | 1.10E-01 | | 6219.60 | 4984.19 | 9558.06 | 2963.78 | 4134.28 | 4998.47 | 4174.62 |
| 25 | P19652 | Alpha-1-acid glycoprotein 2 | 0.62 | 4.37E-02 | 1.10E-01 | | 4222.83 | 2803.16 | 3077.62 | 6133.50 | 5553.03 | 5202.46 | 3789.22 |
| 26 | Q08830 | Fibrinogen-like protein 1 | 1.08 | 4.41E-02 | 1.10E-01 | | 1021.88 | 1437.82 | 795.21 | 2780.53 | 2359.91 | 3456.62 | 1072.63 |
| 27 | P02787 | Serotransferrin | -0.67 | 4.50E-02 | 1.10E-01 | | 4287.02 | 3927.25 | 3992.87 | 2520.28 | 2161.83 | 1991.26 | 3963.90 |
| 28 | Q9H628 | Ras-related and estrogen-regulated growth inhibitor-like protein | 2.69 | 4.54E-02 | 1.10E-01 | | 621.67 | 11520.49 | 5610.85 | 35124.74 | 43104.84 | 23829.00 | 6526.37 |
| 29 | Q8IV33 | Uncharacterized protein KIAA0825 | 1.35 | 4.59E-02 | 1.10E-01 | | 888.21 | 260.33 | 470.43 | 971.53 | 1369.79 | 683.36 | 2434.81 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 30 | P07357 | Complement component C8 alpha chain | 0.61 | 4.61E-02 | 1.10E-01 | | 2034.81 | 1683.19 | 2064.82 | 3633.34 | 3407.21 | 2834.27 | 2070.88 |
| 31 | Q8WWF3 | Serine-rich single-pass membrane protein 1 | -1.46 | 4.61E-02 | 1.10E-01 | | 7304.62 | 1609.19 | 2598.13 | 647.00 | 888.77 | 1887.93 | 1538.26 |
| 32 | Q14568 | Heat shock protein HSP 90-alpha A2 | 4TB:1HC | | | | | | x | | | | |
| 33 | Q9C0K0 | B-cell lymphoma/leukemia 11B | 4TB:1HC | | | | | x | | | | | |
| 34 | Q9H9A5 | CCR4-NOT transcription complex subunit 10 | 4TB:1HC | | | | | | x | | | | |
| 35 | Q9UKG9 | Peroxisomal carnitine O-octanoyltransferase | 4TB:1HC | | | | | x | | | | | |
| 36 | O15078 | Centrosomal protein of 290 kDa | 1TB:3HC | | | | | | | | | | x |
| 37 | Q86VY9 | Transmembrane protein 200A | 1TB:3HC | | | | | | | | | | x |
| 38 | Q9NQ66 | 1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-1 | 1TB:3HC | | | | | | | | | | x |
| 39 | Q9NZM1 | Myoferlin | 1TB:3HC | | | | | | | x | | | |

## Segment 3

| No. | Protein Accession | Name | logFC | P.Value | adj.P.Val | All Pipelines | HC_115 | HC_116 | HC_117 | TB_118 | TB_119 | TB_114 | TB_121 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Q9H2S1 | Small conductance calcium-activated potassium channel protein 2 | -3.57 | 9.24E-04 | 1.61E-01 |  | 212985.93 | 83330.56 | 64466.34 | 4541.92 | 10137.26 | 15951.60 | 8079.57 |
| 2 | O14788 | Tumor necrosis factor ligand superfamily member 11 | -2.67 | 1.52E-03 | 1.61E-01 |  | 1069.97 | 1311.37 | 1185.39 | 169.58 | 152.88 | NA | 247.20 |
| 3 | Q12905 | Interleukin enhancer-binding factor 2 | 3.84 | 1.06E-02 | 3.94E-01 |  | NA | NA | 1116.14 | 21980.09 | 15551.29 | NA | 12000.65 |
| 4 | Q8IVL1 | Neuron navigator 2 | -3.39 | 1.40E-02 | 3.94E-01 |  | 664.27 | 714.86 | 983.98 | 74.12 | NA | NA | NA |
| 5 | B2HQK3 | ATP synthase gamma chain (M.marinum) | -1.82 | 1.95E-02 | 3.94E-01 | * | NA | 744.67 | 490.98 | 213.96 | 147.51 | 117.81 | 232.23 |
| 6 | Q9HC77 | Centromere protein J | -2.25 | 2.17E-02 | 3.94E-01 |  | 462.52 | 2344.75 | 921.90 | 258.91 | 158.73 | 225.88 | NA |
| 7 | Q8WUA8 | Tsukushin | -1.53 | 2.99E-02 | 3.94E-01 |  | 13745.84 | 8291.24 | 15497.08 | 3335.15 | 8196.31 | 4318.86 | 2566.24 |
| 8 | Q86TB3 | Alpha-protein kinase 2 | -1.48 | 3.11E-02 | 3.94E-01 | * | 353.57 | 804.17 | 668.25 | 180.45 | 341.54 | 226.21 | 129.90 |
| 9 | A1KJN3 | Protein translocase subunit SecA 2 | 2.33 | 3.55E-02 | 3.94E-01 |  | 344.22 | 717.79 | 130.49 | 2591.46 | 3333.38 | 1667.40 | 455.46 |
| 10 | P49747 | Cartilage oligomeric matrix protein | 2.24 | 4.50E-02 | 3.94E-01 |  | 1061.80 | 1604.12 | 1586.15 | 11230.33 | 13342.61 | 10447.21 | 1187.81 |
| 11 | A0FGR9 | Extended synaptotagmin-3 | 4TB:1HC |  |  |  |  | x |  |  |  |  |  |
| 12 | P30153 | Serine/threonine-protein phosphatase 2A | 4TB:1HC |  |  |  |  | x |  |  |  |  |  |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 65 kDa regulatory 13subunit A alp14ha isoform | | | | | | | | | | | |
| 13 | P62857 | 40S ri15bosomal protein 16S28 | 4TB:1HC | | | | x | | | | | | |
| 14 | P99999 | Cytochrome c | 4TB:1HC | | | | | x | | | | | |
| 15 | B2HE73 | Catalase-peroxidase | 4TB:1HC | | | | | x | | | | | |
| 16 | Q13127 | RE1-silencing transcription factor | 1TB:3HC | | | | | | | | | | x |
| 17 | Q8IVL1 | Neuron navigator 2 | 1TB:3HC | | | | | | | x | | | |
| 18 | Q96HZ4 | Transcription cofactor HES-6 | 1TB:3HC | | | | | | | | | | x |
| 19 | Q96JB1 | Dynein heavy chain 8, axonemal | 1TB:3HC | | | | | | | x | | | |

## Segment 4

| No. | Protein Accession | Name | logFC | P.Value | adj.P.Val | All Pipelines | HC_115 | HC_116 | HC_117 | 118 | 119 | 114 | 121 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | P26543 | Replication protein E1 | 3.03 | 7.50E-04 | 3.99E-02 |  | 168.28 | NA | NA | 1415.60 | 1031.47 | 1773.56 | NA |
| 2 | P26676 | RNA-directed RNA polymerase L | 1.95 | 1.52E-03 | 3.99E-02 |  | 499.65 | 507.36 | 555.13 | NA | NA | 1465.46 | 2753.73 |
| 3 | P02741 | C-reactive protein | 2.55 | 2.13E-03 | 3.99E-02 | * | 413.83 | 558.83 | 446.81 | 5451.67 | 2919.61 | 5087.91 | 709.99 |
| 4 | P07988 | Pulmonary surfactant-associated protein B | 1.29 | 2.63E-03 | 3.99E-02 | * | 3030.92 | 3173.56 | 2704.91 | 7515.61 | 8116.19 | 9413.09 | 4801.81 |
| 5 | Q7KZ85 | Transcription elongation factor SPT6 | -1.63 | 4.04E-03 | 4.18E-02 |  | 8814.45 | 13888.41 | 28467.20 | 5343.37 | 3054.01 | 6416.83 | 5565.08 |
| 6 | Q9H0I3 | Coiled-coil domain-containing protein 113 | 3.41 | 4.05E-03 | 4.18E-02 |  | 476.94 | NA | NA | NA | NA | NA | 5082.55 |
| 7 | Q7Z3J3 | RanBP2-like and GRIP domain-containing protein 4 | 2.13 | 4.40E-03 | 4.18E-02 |  | NA | 353.88 | NA | 1268.19 | 1541.01 | 1903.40 | NA |
| 8 | P78352 | Disks large homolog 4 | 1.47 | 5.64E-03 | 4.51E-02 | * | 8584.83 | 11105.86 | 7569.72 | 27931.07 | 36337.37 | 31733.36 | 11684.46 |
| 9 | O75015 | Low affinity immunoglobulin gamma Fc region receptor III-B | 1.63 | 6.53E-03 | 4.51E-02 |  | 4668.72 | 3608.22 | 4402.01 | 19890.94 | 13867.57 | 20453.97 | 5056.25 |
| 10 | P0DJI8 | Serum amyloid A-1 protein | 1.49 | 7.51E-03 | 4.76E-02 | * | 599.78 | 816.69 | 816.47 | 2810.60 | 1910.54 | 3622.77 | 949.78 |
| 11 | P04545 | Matrix M2-1 | -2.13 | 1.36E-02 | 7.41E-02 |  | 2705.10 | 5435.41 | 3281.40 | 761.96 | 490.99 | 331.79 | 3869.20 |
| 12 | P18065 | Insulin-like growth factor-binding protein 2 | 0.91 | 1.73E-02 | 7.93E-02 |  | 1043.12 | 1267.94 | 742.61 | 2127.76 | 1481.80 | 2022.21 | 1891.57 |
| 13 | P0CG05 | Ig lambda-2 chain C regions | -1.56 | 1.81E-02 | 7.93E-02 |  | 1890.61 | 2299.32 | 7055.06 | 714.72 | 875.35 | 815.80 | 2484.71 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | O94822 | E3 ubiquitin-protein ligase listerin | 3.74 | 2.11E-02 | 7.93E-02 | | 189.54 | 1492.22 | NA | 15386.74 | 6863.41 | 33041.86 | 722.79 |
| 15 | P03950 | Angiogenin | -0.98 | 2.24E-02 | 7.93E-02 | | 2448.29 | 2445.12 | 2299.77 | 742.42 | 1062.73 | 1516.76 | 1833.42 |
| 16 | P46939 | Utrophin | 2.77 | 2.58E-02 | 7.93E-02 | | NA | 1106.14 | 179.88 | 2671.98 | 3876.17 | 12336.37 | 675.86 |
| 17 | P08637 | Low affinity immunoglobulin gamma Fc region receptor III-A | 1.08 | 2.76E-02 | 7.93E-02 | | 1412.32 | 1617.46 | 1196.55 | 5290.68 | 3771.83 | 1618.26 | 2351.84 |
| 18 | P25054 | Adenomatous polyposis coli protein | 2.13 | 2.76E-02 | 7.93E-02 | | 224.11 | NA | NA | NA | NA | 983.19 | NA |
| 19 | P13671 | Complement component C6 | 1.26 | 2.82E-02 | 7.93E-02 | | 1686.64 | 1516.51 | 487.61 | 3267.03 | 2903.70 | 3268.22 | 1418.88 |
| 20 | P01743 | Immunoglobulin heavy variable 1-46 | -1.39 | 2.96E-02 | 8.04E-02 | | 5141.13 | 5088.71 | 1771.56 | 1035.92 | 844.13 | 1160.82 | 3465.01 |
| 21 | Q13103 | Secreted phosphoprotein 24 | -1.16 | 3.32E-02 | 8.13E-02 | | 3985.42 | 3514.43 | 2660.30 | 2390.49 | 1167.95 | 698.10 | 2579.63 |
| 22 | P01781 | Immunoglobulin heavy variable 3-7 | -0.75 | 3.32E-02 | 8.13E-02 | | 1735.52 | 1606.52 | 1351.42 | 981.27 | 849.31 | 757.00 | 1167.58 |
| 23 | O43866 | CD5 antigen-like | -0.84 | 3.42E-02 | 8.13E-02 | | 2775.06 | 1660.24 | 2459.75 | 1309.89 | 1244.36 | 888.04 | 1709.98 |
| 24 | P68871 | Hemoglobin subunit beta | -0.85 | 3.46E-02 | 8.13E-02 | | 1204.20 | 1490.67 | 2710.00 | 1029.48 | 829.51 | 985.60 | 938.75 |
| 25 | P0DJI9 | Serum amyloid A-2 protein | 1.28 | 3.56E-02 | 8.13E-02 | | 444.88 | 755.46 | 743.54 | 1515.19 | 1174.69 | 4160.74 | 747.96 |
| 26 | P01719 | Immunoglobulin lambda variable 3-21 | -0.98 | 3.84E-02 | 8.34E-02 | | 1449.43 | 1540.26 | 2242.72 | 609.49 | 1364.05 | 1261.88 | 533.31 |
| 27 | P17936 | Insulin-like growth factor-binding protein 3 | -0.73 | 4.75E-02 | 9.63E-02 | * | 1796.87 | 1537.38 | 2073.74 | 1438.64 | 1193.45 | 1002.71 | 797.54 |
| 28 | P08670 | Vimentin | 4TB:1HC | | | | | x | | | | | |
| 29 | Q9GZM8 | Nuclear distribution protein nudE-like 1 | 4TB:1HC | | | | x | | | | | | |
| 30 | O00585 | CCL21 | 1TB:3HC | | | | | | | | | | x |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 31 | O75419 | CDC45 | 1TB:3HC | | | | | | | | | | x |
| 32 | O75923 | Dysferlin | 1TB:3HC | | | | | | | | | | x |
| 33 | P00488 | Coagulation factor XIII A chain | 1TB:3HC | | | | | | | | | | x |
| 34 | P01275 | Glucagon | 1TB:3HC | | | | | | | | | | x |
| 35 | P04792 | Heat shock protein beta-1 | 1TB:3HC | | | | | | | x | | | |
| 36 | P09681 | Gastric inhibitory polypeptide | 1TB:3HC | | | | | | | | | | x |
| 37 | P99999 | Cytochrome c | 1TB:3HC | | | | | | | | | | x |
| 38 | Q12769 | Nuclear pore complex protein Nup160 | 1TB:3HC | | | | | | | | | x | |
| 39 | Q13948 | Protein CASP | 1TB:3HC | | | | | | | | | | x |
| 40 | Q5T6F2 | Ubiquitin-associated protein 2 | 1TB:3HC | | | | | | | | x | | |
| 41 | Q6UY70 | RNA-directed RNA polymerase L | 1TB:3HC | | | | | | | | | | x |
| 42 | Q86VW0 | SEC14 domain and spectrin repeat-containing protein 1 | 1TB:3HC | | | | | | | | | | x |
| 43 | Q96JM7 | Lethal(3)malignant brain tumor-like protein 3 | 1TB:3HC | | | | | | | | | | x |
| 44 | Q9HDC9 | Adipocyte plasma membrane-associated protein | 1TB:3HC | | | | | | | | | | x |
| 45 | Q9NZT1 | Calmodulin-like protein 5 | 1TB:3HC | | | | | | | | | | x |

## Annexe 3. Summary of proteins significantly regulated from multiconsensus report and *PL3* analysis

Table presenting UniProt accession ID and parameters from LIMMA fit. Relative expression of each protein per sample included as well.

| Protein Accession | logFC | AveExpr | t | P.Value | adj.P.Val | B | MP_113 | TB_114 | HC_115 | HC_116 | HC_117 | TB_118 | TB_119 | TB_121 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P9WGP3 | 2.417462 | 9.727561 | 2.874197 | 0.023295 | 0.551897 | -3.07851 | 1150.772 | 1849.704 | 340.5787 | 692.7804 | 128.1278 | 3027.214 | 3062.104 | 447.3947 |
| Q86SQ4 | 2.223481 | 11.69616 | 3.916241 | 0.005534 | 0.551897 | -1.84126 | 11074.91 | 3518.69 | 1032.183 | 1354.16 | 1110.441 | 10610.09 | 2423.75 | 9447.413 |
| P43490 | 1.877177 | 11.93286 | 2.359881 | 0.04959 | 0.551897 | -3.73958 | 3124.397 | 5123.893 | 1928.03 | 2904.575 | 1262.592 | 31459.29 | 3185.785 | 4812.751 |
| Q7KZF4 | 1.870984 | 12.11476 | 2.402821 | 0.046527 | 0.551897 | -3.68392 | 3442.132 | 5426.861 | 2282.713 | 2708.32 | 1703.031 | 37493.93 | 4413.619 | 4599.913 |
| P05109 | 1.707977 | 11.01586 | 2.938039 | 0.021245 | 0.551897 | -2.99799 | 2520.496 | 6984.227 | 729.2144 | 1717.232 | 856.3205 | 4105.536 | 3042.155 | 1433.401 |
| P02741 | 1.625375 | 10.08319 | 3.322674 | 0.012323 | 0.551897 | -2.52439 | 1247.518 | 2934.013 | 498.3314 | 635.3758 | 550.4155 | 2453.429 | 1645.756 | 744.4996 |
| P05164 | 1.59594 | 11.04472 | 2.844706 | 0.024311 | 0.551897 | -3.11586 | 1286.726 | 1467.007 | 1931.687 | 967.5374 | 936.4003 | 5990.125 | 5565.648 | 3600.876 |
| Q8N9V6 | 1.590722 | 11.07559 | 3.232609 | 0.013975 | 0.551897 | -2.63334 | 1211.875 | 1817.813 | 1347.689 | 1000.901 | 1441.437 | 3136.036 | 4667.963 | 7505.806 |
| P18669 | 1.589438 | 9.686979 | 2.516007 | 0.039346 | 0.551897 | -3.53736 | 621.2132 | 1169.353 | 622.1307 | 642.7353 | 239.1718 | 3843.375 | 881.2276 | 905.6485 |
| Q9Y5E7 | 1.553931 | 10.2979 | 3.513719 | 0.009471 | 0.551897 | -2.2977 | 1089.669 | 1169.353 | 939.2031 | 808.5709 | 440.973 | 2176.429 | 2014.805 | 3370.663 |
| Q15149 | 1.458478 | 10.21631 | 2.681062 | 0.03086 | 0.551897 | -3.32468 | 615.1029 | 1296.919 | 870.2516 | 681.9863 | 665.1966 | 5361.929 | 2004.83 | 1184.944 |
| Q9H6F5 | 1.440924 | 13.10499 | 2.57604 | 0.03601 | 0.551897 | -3.45982 | 8595.147 | 17327.69 | 3625.178 | 6034.843 | 5701.685 | 17114.64 | 23140.33 | 4934.373 |
| Q8N720 | 1.431107 | 11.29291 | 3.173909 | 0.015179 | 0.551897 | -2.70502 | 3989.513 | 3849.829 | 1149.192 | 762.4509 | 2698.157 | 2732.902 | 3912.411 | 4045.23 |
| Q9NP58 | 1.425772 | 9.92497 | 2.564094 | 0.03665 | 0.551897 | -3.47524 | 1079.485 | 1392.593 | 202.6756 | 785.0202 | 1014.345 | 1533.393 | 1725.55 | 1242.28 |
| P14555 | 1.414323 | 9.110333 | 2.425981 | 0.044956 | 0.551897 | -3.6539 | 635.4706 | 2455.641 | 295.6557 | 306.6485 | 326.7258 | 612.3679 | 721.1404 | 426.5453 |
| Q9Y6V0 | 1.409127 | 16.44552 | 2.424671 | 0.045043 | 0.551897 | -3.6556 | 39120.14 | 250879.4 | 46803.45 | 73716.34 | 54988.16 | 172213.4 | 240380.1 | 52216.61 |
| P04843 | 1.393883 | 11.66016 | 2.416241 | 0.04561 | 0.551897 | -3.66652 | 2209.89 | 3757.876 | 2173.017 | 2296.184 | 1526.856 | 15235 | 4054.545 | 3079.639 |
| Q8IV33 | 1.383717 | 9.790568 | 2.489358 | 0.040927 | 0.551897 | -3.57183 | 1564.744 | 680.8824 | 834.2088 | 256.6035 | 491.1564 | 988.2964 | 1436.296 | 2380.314 |
| P01877 | 1.30832 | 11.14265 | 3.510079 | 0.009518 | 0.551897 | -2.30196 | 2291.36 | 3842.919 | 1036.362 | 1354.16 | 1729.725 | 2413.857 | 2498.557 | 5294.895 |
| Q9BY89 | 1.294592 | 9.149526 | 3.936645 | 0.005389 | 0.551897 | -1.81893 | 602.8824 | 1040.724 | 399.083 | 277.7009 | 345.9449 | 782.525 | 1067.248 | 538.6111 |
| P06702 | 1.257019 | 10.748 | 2.771887 | 0.027025 | 0.551897 | -3.20848 | 1904.375 | 4347.867 | 740.184 | 1305.096 | 1131.795 | 3017.321 | 2114.547 | 1324.81 |
| Q14764 | 1.237704 | 10.62209 | 2.606167 | 0.034448 | 0.551897 | -3.42098 | 1619.228 | 2041.053 | 1535.738 | 1108.841 | 522.1206 | 3996.714 | 1665.705 | 1945.95 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P30046 | 1.167051 | 9.70309 | 2.456819 | 0.042947 | 0.551897 | -3.61396 | 605.9375 | 1261.838 | 399.083 | 591.709 | 702.5672 | 1442.873 | 2169.406 | 586.8255 |
| P07900 | 1.149931 | 9.554208 | 2.629059 | 0.033307 | 0.551897 | -3.39151 | 519.375 | 577.2352 | 429.3798 | 648.623 | 455.9213 | 1948.893 | 1266.734 | 1085.909 |
| P02750 | 1.146861 | 11.56515 | 3.053257 | 0.018012 | 0.551897 | -2.85394 | 2953.309 | 4927.229 | 1634.987 | 2610.192 | 1687.015 | 5139.339 | 5426.008 | 2432.437 |
| O75015 | 1.144259 | 13.02735 | 2.668527 | 0.031433 | 0.551897 | -3.34077 | 14715.63 | 15945.73 | 5850.431 | 3797.535 | 5274.593 | 12910.18 | 12567.59 | 5290.551 |
| Q96II8 | 1.144186 | 10.9635 | 3.593924 | 0.008493 | 0.551897 | -2.2044 | 2963.493 | 3731.299 | 987.2602 | 1462.1 | 1195.859 | 2077.5 | 2084.624 | 3057.921 |
| P01133 | 1.119007 | 14.31428 | 2.36436 | 0.049261 | 0.551897 | -3.73377 | 16294.12 | 29021.22 | 6905.598 | 20606.78 | 17297.25 | 37098.21 | 38201.49 | 17982.66 |
| Q9BXR6 | 1.085508 | 11.78053 | 2.946107 | 0.021 | 0.551897 | -2.98785 | 2917.665 | 7563.588 | 1567.08 | 2772.103 | 2989.648 | 5203.643 | 3780.252 | 4161.205 |
| P06576 | 1.071517 | 12.03651 | 2.907587 | 0.022198 | 0.551897 | -3.03634 | 3645.809 | 4204.356 | 2511.506 | 4140.982 | 2120.514 | 8864 | 5316.291 | 6089.78 |
| P18428 | 1.055326 | 11.73107 | 2.368957 | 0.048926 | 0.551897 | -3.72781 | 3238.456 | 6665.313 | 1687.223 | 2914.388 | 2327.654 | 5609.25 | 5775.108 | 2228.286 |
| Q9UPY3 | 1.007305 | 10.1099 | 2.970923 | 0.020264 | 0.551897 | -2.95671 | 2077.5 | 1056.67 | 695.7834 | 501.4317 | 891.5556 | 1434.464 | 1994.856 | 1138.033 |
| P78352 | 0.992514 | 12.43058 | 2.740345 | 0.028297 | 0.551897 | -3.24875 | 5886.25 | 8483.125 | 3875.911 | 4768.998 | 2722.715 | 8379.25 | 9246.157 | 4430.511 |
| P0DJI8 | 0.979326 | 10.16666 | 2.789327 | 0.026347 | 0.551897 | -3.18625 | 1247.518 | 2354.652 | 645.6369 | 836.0466 | 848.3125 | 1627.375 | 1456.245 | 955.6004 |
| Q9HAU6 | 0.97391 | 9.768267 | 2.528461 | 0.038629 | 0.551897 | -3.52126 | 796.375 | 1796.552 | 664.9642 | 801.2113 | 406.8056 | 1147.571 | 1223.844 | 767.9552 |
| Q8IXL6 | 0.964658 | 9.850328 | 2.571637 | 0.036245 | 0.551897 | -3.4655 | 639.5441 | 1615.833 | 662.3524 | 1049.965 | 420.6861 | 1365.214 | 1336.553 | 955.6004 |
| Q8WU39 | 0.917377 | 12.5559 | 2.400088 | 0.046716 | 0.551897 | -3.68746 | 5865.882 | 8132.319 | 3278.853 | 4371.581 | 5178.497 | 4634.804 | 12313.25 | 8552.623 |
| Q9NS69 | 0.904797 | 10.45556 | 2.722299 | 0.029053 | 0.551897 | -3.27183 | 1344.265 | 2253.662 | 1021.214 | 1079.403 | 873.4042 | 1256.393 | 2683.081 | 1537.648 |
| Q9P225 | 0.893104 | 10.28268 | 2.617589 | 0.033874 | 0.551897 | -3.40627 | 1731.25 | 953.5543 | 826.3734 | 675.1174 | 1041.038 | 1671.893 | 1944.985 | 1859.077 |
| Q96FL9 | 0.886417 | 12.05135 | 3.368915 | 0.011557 | 0.551897 | -2.46896 | 4297.574 | 4900.652 | 3907.252 | 2826.073 | 2402.395 | 5579.571 | 5805.031 | 5820.475 |
| P51826 | 0.884731 | 12.29495 | 2.502292 | 0.040152 | 0.551897 | -3.5551 | 7189.779 | 4326.606 | 2904.321 | 3169.519 | 4132.12 | 7973.643 | 9036.698 | 4769.315 |
| P28062 | 0.845498 | 10.39952 | 2.470988 | 0.042055 | 0.551897 | -3.59561 | 1512.298 | 1828.443 | 719.812 | 748.2224 | 1596.258 | 1825.232 | 1660.718 | 1537.648 |
| P07988 | 0.725376 | 12.42482 | 2.603802 | 0.034568 | 0.551897 | -3.42403 | 6069.559 | 8966.812 | 4262.457 | 3660.157 | 4313.634 | 6361.107 | 7141.584 | 5021.246 |
| P13796 | 0.635005 | 11.2064 | 2.456167 | 0.042989 | 0.551897 | -3.6148 | 2607.059 | 3561.212 | 1681.999 | 1834.985 | 1927.255 | 2740.321 | 2922.464 | 2197.881 |
| P08294 | -0.58485 | 10.60326 | -2.58484 | 0.035546 | 0.551897 | -3.44847 | 1822.904 | 1302.234 | 1796.918 | 1933.112 | 2028.689 | 1108 | 1356.502 | 1363.902 |
| Q92859 | -0.60171 | 10.35747 | -2.68598 | 0.030639 | 0.551897 | -3.31837 | 1385 | 974.8152 | 1713.341 | 1560.228 | 1687.015 | 1182.196 | 1047.299 | 1164.095 |
| Q9H4G4 | -0.64265 | 9.998795 | -2.77008 | 0.027096 | 0.551897 | -3.21078 | 1056.062 | 739.3501 | 1379.03 | 1305.096 | 1259.923 | 815.1714 | 832.8524 | 999.0367 |
| P05451 | -0.6664 | 9.947508 | -2.61059 | 0.034224 | 0.551897 | -3.41528 | 845.2574 | 812.1689 | 1159.639 | 1491.538 | 1313.309 | 666.2839 | 866.7649 | 1003.38 |
| Q4LDE5 | -0.66853 | 10.49442 | -2.3823 | 0.047966 | 0.551897 | -3.71052 | 1171.14 | 1056.67 | 1953.626 | 1785.921 | 2082.076 | 1315.75 | 950.5489 | 1667.957 |
| P10721 | -0.67626 | 13.47542 | -2.5985 | 0.034838 | 0.551897 | -3.43086 | 10861.05 | 12703.43 | 15357.38 | 14670.07 | 14948.24 | 9091.536 | 7909.604 | 8474.438 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q8WZ42 | -0.67861 | 9.541532 | -2.67123 | 0.031308 | 0.551897 | -3.3373 | 865.1158 | 661.7475 | 934.5019 | 1049.965 | 886.217 | 530.2571 | 494.7243 | 728.4281 |
| O60880 | -0.68555 | 12.59612 | -2.78434 | 0.02654 | 0.551897 | -3.19261 | 5091.912 | 4592.368 | 7971.212 | 8713.725 | 8392.368 | 4669.429 | 5106.831 | 6645.766 |
| P10646 | -0.70617 | 8.892241 | -2.79889 | 0.025984 | 0.551897 | -3.17407 | 386.9853 | 459.7684 | 801.3001 | 560.7988 | 603.2682 | 404.6179 | 353.0895 | 377.0278 |
| Q9HA90 | -0.70656 | 9.542743 | -2.54695 | 0.037589 | 0.551897 | -3.49737 | 950.1507 | 451.7955 | 780.9281 | 927.3052 | 1195.859 | 619.2929 | 608.4311 | 683.6886 |
| Q9BWP8 | -0.75392 | 10.65944 | -3.03926 | 0.018375 | 0.551897 | -2.87135 | 1721.066 | 1116.201 | 2298.384 | 1903.674 | 2306.3 | 1177.25 | 1290.672 | 1589.772 |
| P17936 | -0.76371 | 10.6261 | -3.05005 | 0.018094 | 0.551897 | -2.85792 | 1812.721 | 1037.535 | 2193.912 | 1727.045 | 2434.427 | 1295.964 | 1346.528 | 1285.717 |
| P68871 | -0.76854 | 10.81701 | -2.74877 | 0.027951 | 0.551897 | -3.23798 | 2235.349 | 1488.268 | 1880.496 | 2134.274 | 3325.983 | 1419.625 | 1256.759 | 1416.026 |
| Q9HBJ7 | -0.80325 | 10.89077 | -2.99116 | 0.019684 | 0.551897 | -2.93137 | 1965.478 | 1371.332 | 2423.75 | 2718.133 | 2658.651 | 1226.714 | 1416.348 | 2058.884 |
| P29622 | -0.8047 | 11.71934 | -2.5327 | 0.038388 | 0.551897 | -3.51578 | 3330.11 | 2056.998 | 4393.047 | 4435.364 | 5146.465 | 2443.536 | 2333.981 | 4265.453 |
| Q96DX7 | -0.83406 | 10.62398 | -2.67169 | 0.031287 | 0.551897 | -3.33671 | 2006.213 | 917.4106 | 2068.545 | 1864.423 | 2477.137 | 1671.893 | 935.5874 | 1398.651 |
| P41229 | -0.84168 | 11.88276 | -2.38094 | 0.048063 | 0.551897 | -3.71227 | 4511.434 | 1754.03 | 4753.475 | 5220.385 | 5466.784 | 3333.893 | 2613.261 | 4421.823 |
| O94923 | -0.86845 | 10.01533 | -2.5355 | 0.038229 | 0.551897 | -3.51217 | 993.4322 | 580.9559 | 1812.589 | 1280.565 | 1364.027 | 853.2589 | 663.2896 | 1272.686 |
| O76096 | -0.8873 | 12.95924 | -3.34019 | 0.012027 | 0.551897 | -2.50336 | 8666.434 | 6091.266 | 9945.733 | 10205.26 | 13773.73 | 7291.036 | 4847.5 | 6202.715 |
| Q12860 | -0.90355 | 10.32289 | -3.0401 | 0.018353 | 0.551897 | -2.8703 | 1395.184 | 904.1225 | 2021.533 | 1422.849 | 2060.721 | 773.6214 | 927.1093 | 1350.871 |
| A0A0C4DH32 | -0.95526 | 10.32704 | -2.69878 | 0.03007 | 0.551897 | -3.30195 | 1211.875 | 835.5559 | 2486.433 | 1422.849 | 1911.239 | 1360.268 | 668.7755 | 1190.157 |
| Q76LX8 | -0.96291 | 13.0073 | -2.83183 | 0.02477 | 0.551897 | -3.13221 | 9231.636 | 4278.769 | 13529.12 | 10646.84 | 11584.88 | 5272.893 | 6109.246 | 9946.931 |
| Q8NCX0 | -0.9788 | 10.39058 | -4.07263 | 0.004521 | 0.551897 | -1.67208 | 1059.118 | 1190.614 | 2183.464 | 2099.929 | 1868.53 | 893.325 | 945.5617 | 1155.408 |
| P51161 | -0.99295 | 10.37753 | -2.3755 | 0.048453 | 0.551897 | -3.71933 | 1395.184 | 666.5313 | 1535.738 | 2148.993 | 2274.268 | 678.65 | 1157.016 | 1789.579 |
| Q96RP7 | -0.99561 | 11.06666 | -2.39675 | 0.046948 | 0.551897 | -3.69179 | 2016.397 | 1520.159 | 3290.867 | 1727.045 | 5819.135 | 1691.679 | 1526.065 | 1711.393 |
| Q9NQX5 | -1.00194 | 10.3927 | -2.7955 | 0.026112 | 0.551897 | -3.17839 | 1751.618 | 804.7275 | 2235.7 | 1589.666 | 2007.335 | 765.7071 | 817.8909 | 1694.019 |
| Q15166 | -1.01925 | 10.47052 | -3.56312 | 0.008855 | 0.551897 | -2.2401 | 1700.699 | 697.3596 | 2178.241 | 2119.555 | 1921.916 | 1058.536 | 1186.939 | 1242.28 |
| O95502 | -1.0581 | 11.77935 | -2.35561 | 0.049907 | 0.551897 | -3.74512 | 5030.809 | 1254.397 | 5025.102 | 5769.899 | 4516.503 | 2364.393 | 4767.706 | 2501.935 |
| P03950 | -1.08584 | 10.6694 | -2.48567 | 0.041151 | 0.551897 | -3.5766 | 1889.099 | 1498.898 | 2580.458 | 2418.844 | 2359.686 | 606.9268 | 949.5514 | 2058.884 |
| Q9BRX9 | -1.08807 | 8.839274 | -2.44406 | 0.043767 | 0.551897 | -3.63048 | 526.5037 | 229.6184 | 524.4494 | 486.7126 | 1291.955 | 383.8429 | 271.3004 | 466.5067 |
| Q96PW8 | -1.10338 | 9.531773 | -2.51138 | 0.039616 | 0.551897 | -3.54335 | 743.4191 | 262.0414 | 1065.614 | 1118.654 | 1259.923 | 897.2821 | 490.2359 | 700.1944 |
| Q2NL82 | -1.11205 | 11.40523 | -2.62698 | 0.033409 | 0.551897 | -3.39418 | 2851.471 | 1966.639 | 2188.688 | 4626.713 | 7228.541 | 2354.5 | 1730.538 | 1750.486 |
| P08047 | -1.13368 | 11.70685 | -3.01454 | 0.019036 | 0.551897 | -2.90216 | 1028.566 | 1690.247 | 4617.662 | 6633.421 | 7773.084 | 2868.929 | 3161.847 | 4152.518 |
| P51570 | -1.13949 | 12.56494 | -2.46165 | 0.042641 | 0.551897 | -3.6077 | 9481.14 | 5187.675 | 15775.27 | 7330.127 | 6150.132 | 2314.929 | 3919.892 | 5724.915 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A6NFK2 | -1.17376 | 12.89546 | -2.59961 | 0.034782 | 0.551897 | -3.42943 | 9185.809 | 5612.895 | 5547.462 | 17270.45 | 17190.47 | 4372.643 | 4837.526 | 6324.337 |
| O14777 | -1.18577 | 10.71694 | -2.38577 | 0.047719 | 0.551897 | -3.70601 | 5468.713 | 727.125 | 4157.985 | 1138.279 | 2487.814 | 979.3929 | 878.7341 | 1598.459 |
| P39059 | -1.18783 | 10.96711 | -2.36378 | 0.049304 | 0.551897 | -3.73452 | 2464.485 | 1275.658 | 3437.128 | 2463.001 | 3555.545 | 1108 | 759.0427 | 3240.354 |
| P61916 | -1.19849 | 9.881909 | -2.41125 | 0.045949 | 0.551897 | -3.673 | 895.1581 | 812.1689 | 783.5399 | 3136.647 | 1452.115 | 445.1786 | 716.1533 | 759.2679 |
| Q6RI45 | -1.21172 | 11.32533 | -2.47927 | 0.041543 | 0.551897 | -3.58489 | 2932.941 | 1498.898 | 5610.145 | 2531.69 | 4740.727 | 2305.036 | 924.6157 | 2979.736 |
| P00441 | -1.22805 | 11.50451 | -4.82242 | 0.001803 | 0.40679 | -0.92573 | 7536.029 | 1732.769 | 4628.109 | 3140.081 | 4826.145 | 1859.857 | 1715.576 | 1737.455 |
| Q86TB3 | -1.2451 | 8.921541 | -3.23522 | 0.013924 | 0.551897 | -2.63015 | 476.6029 | 251.9424 | 885.4001 | 565.7052 | 1006.871 | 277.4946 | 313.1924 | 580.7444 |
| Q13103 | -1.27031 | 11.1577 | -2.70507 | 0.029794 | 0.551897 | -3.29389 | 2515.404 | 892.9605 | 4502.742 | 3915.288 | 2936.261 | 2028.036 | 1062.261 | 2962.361 |
| O00408 | -1.28048 | 12.32268 | -2.71322 | 0.029441 | 0.551897 | -3.28345 | 6079.743 | 2625.729 | 5463.885 | 17859.21 | 5861.845 | 3551.536 | 3181.795 | 4595.569 |
| P06732 | -1.28281 | 11.11089 | -2.4018 | 0.046597 | 0.551897 | -3.68524 | 3564.338 | 1067.832 | 2648.365 | 4180.233 | 3656.98 | 943.7786 | 1040.317 | 3783.309 |
| O94991 | -1.29368 | 10.09292 | -3.49296 | 0.009743 | 0.551897 | -2.32203 | 1313.713 | 553.8481 | 1264.111 | 1893.861 | 2338.331 | 601.4857 | 699.197 | 1181.47 |
| P54802 | -1.29894 | 11.81218 | -2.70532 | 0.029784 | 0.551897 | -3.29358 | 4144.816 | 1796.552 | 7010.07 | 8591.065 | 3400.724 | 1998.357 | 1944.985 | 4717.191 |
| Q9ULM0 | -1.30008 | 11.06693 | -2.48943 | 0.040923 | 0.551897 | -3.57174 | 2469.577 | 1275.658 | 3897.85 | 5426.452 | 2059.654 | 843.8607 | 1286.682 | 3010.141 |
| P35443 | -1.32188 | 11.33017 | -2.55966 | 0.03689 | 0.551897 | -3.48096 | 3004.228 | 1674.301 | 5223.599 | 3620.906 | 4062.718 | 850.7857 | 1466.219 | 4004.834 |
| Q9HD43 | -1.33429 | 10.68969 | -3.01903 | 0.018914 | 0.551897 | -2.89655 | 1659.963 | 1037.535 | 4931.078 | 2355.061 | 1889.884 | 738.9964 | 1755.473 | 1129.346 |
| Q9UMX5 | -1.36077 | 11.17676 | -4.72951 | 0.002012 | 0.40679 | -1.01244 | 2978.768 | 1302.234 | 4189.327 | 3537.497 | 3785.107 | 1266.286 | 1356.502 | 2206.568 |
| P0C221 | -1.36387 | 14.37479 | -2.54061 | 0.037942 | 0.551897 | -3.50557 | 33301.1 | 12437.66 | 54116.49 | 20116.14 | 36729.96 | 7894.5 | 10871.97 | 29189.25 |
| Q96JM2 | -1.37084 | 10.51077 | -2.62146 | 0.033681 | 0.551897 | -3.40128 | 1405.368 | 829.1776 | 3144.607 | 2148.993 | 2381.041 | 544.1071 | 812.9038 | 2475.874 |
| I6Y481 | -1.4317 | 10.61025 | -2.76651 | 0.027238 | 0.551897 | -3.21533 | 1965.478 | 837.682 | 2235.7 | 2531.69 | 3352.676 | 822.0964 | 560.5545 | 2475.874 |
| Q8N3C0 | -1.43469 | 10.98243 | -2.39042 | 0.047391 | 0.551897 | -3.69999 | 1455.778 | 1573.311 | 9642.764 | 2767.196 | 1965.693 | 1131.743 | 921.1247 | 2241.752 |
| Q8WUA8 | -1.44793 | 12.82963 | -3.31477 | 0.012459 | 0.551897 | -2.53389 | 13356.08 | 4799.663 | 13633.59 | 7982.675 | 15268.56 | 3897.786 | 7525.594 | 2523.654 |
| Q6P1X5 | -1.47866 | 11.8287 | -3.31528 | 0.012451 | 0.551897 | -2.53329 | 2882.022 | 4241.563 | 10186.02 | 6682.485 | 4527.181 | 2047.821 | 2643.184 | 1502.899 |
| P08123 | -1.48559 | 9.507023 | -3.10485 | 0.016737 | 0.551897 | -2.79 | 846.2757 | 477.3087 | 1009.199 | 1687.794 | 1238.568 | 313.6036 | 294.2413 | 999.0367 |
| P37198 | -1.52557 | 11.08567 | -3.44113 | 0.010461 | 0.551897 | -2.38311 | 2230.257 | 942.9238 | 3813.227 | 3306.898 | 4932.919 | 1058.536 | 1216.862 | 2953.674 |
| Q9UN79 | -1.52747 | 10.20014 | -2.51293 | 0.039525 | 0.551897 | -3.54134 | 1191.507 | 909.9693 | 1358.136 | 6348.851 | 1153.15 | 550.0429 | 810.909 | 762.7428 |
| Q08378 | -1.59658 | 10.62868 | -2.43946 | 0.044066 | 0.551897 | -3.63645 | 1659.963 | 451.7955 | 2173.017 | 2865.324 | 4164.152 | 1048.643 | 642.3436 | 3014.485 |
| Q8N584 | -1.601 | 9.766489 | -2.87028 | 0.023427 | 0.551897 | -3.08347 | 1334.081 | 444.3542 | 1211.875 | 1717.232 | 1772.434 | 502.5571 | 234.3956 | 1285.717 |
| A0A087WSY4 | -1.62528 | 11.42847 | -2.69164 | 0.030386 | 0.551897 | -3.31111 | 3472.684 | 2338.706 | 4356.482 | 3355.962 | 8947.588 | 612.3679 | 2703.03 | 1893.826 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q96A44 | -1.68049 | 10.54281 | -4.00146 | 0.004954 | 0.551897 | -1.7485 | 1904.375 | 537.9024 | 2016.309 | 2384.499 | 4580.567 | 1157.464 | 812.9038 | 1155.408 |
| Q9Y5W7 | -1.71177 | 11.37289 | -5.01489 | 0.001443 | 0.40679 | -0.75116 | 3493.051 | 982.2566 | 4492.295 | 4523.679 | 6235.551 | 2255.571 | 1536.039 | 1624.521 |
| P53618 | -1.7391 | 9.931582 | -2.45847 | 0.042842 | 0.551897 | -3.61182 | 1002.088 | 424.1563 | 2831.191 | 1197.156 | 2146.14 | 395.7143 | 317.6808 | 2128.383 |
| O43610 | -1.75924 | 14.44902 | -2.77447 | 0.026924 | 0.551897 | -3.20518 | 2973.676 | 9992.654 | 27162.72 | 95968.72 | 82428.85 | 26710.71 | 33812.81 | 10859.09 |
| Q7KZ85 | -1.77999 | 13.0542 | -3.70579 | 0.007307 | 0.551897 | -2.0762 | 15937.68 | 6229.463 | 9412.926 | 14179.43 | 29202.45 | 4521.036 | 2688.068 | 5824.819 |
| Q07507 | -1.79807 | 12.32535 | -2.36036 | 0.049555 | 0.551897 | -3.73896 | 5942.261 | 2514.109 | 12458.28 | 9955.038 | 8669.978 | 1785.661 | 1311.618 | 12792.01 |
| Q96N23 | -1.80847 | 11.04777 | -3.01729 | 0.018961 | 0.551897 | -2.89873 | 1395.184 | 933.3564 | 1702.893 | 5691.397 | 10036.67 | 1295.964 | 1386.425 | 1772.204 |
| P23280 | -1.83233 | 11.35245 | -2.67705 | 0.031043 | 0.551897 | -3.32983 | 4058.254 | 531.5241 | 10024.09 | 4312.705 | 3021.68 | 2750.214 | 1157.016 | 2436.781 |
| Q15645 | -2.13049 | 9.714773 | -4.31393 | 0.003333 | 0.465287 | -1.42019 | 526.5037 | 955.6804 | 2998.346 | 2296.184 | 1323.987 | 396.7036 | 497.7166 | 274.5179 |
| A0A075B6H9 | -2.445 | 11.86787 | -5.10057 | 0.001309 | 0.40679 | -0.67562 | 2871.838 | 1998.531 | 17551.29 | 5877.839 | 10351.65 | 1434.464 | 1336.553 | 3240.354 |
| Q68CZ1 | -2.68301 | 10.68265 | -7.76597 | 9.83E-05 | 0.127721 | 1.108014 | 942.0037 | 655.9008 | 5213.152 | 4160.607 | 6299.615 | 643.0357 | 738.0967 | 1329.153 |
| Q9BZG8 | -2.90215 | 12.87638 | -3.71596 | 0.007209 | 0.551897 | -2.06466 | 8411.838 | 808.9797 | 23715.14 | 43764.88 | 12278.91 | 5045.357 | 5934.697 | 3935.336 |
| Q9H2S1 | -3.48832 | 14.51558 | -6.25184 | 0.000388 | 0.234641 | 0.219017 | 12118.75 | 17752.91 | 211033.4 | 80366.45 | 63423.24 | 5302.571 | 9306.003 | 7931.483 |

## Annexe 4. Proteins significantly regulated from segment 4 fully quantified and common to sets A, B and C by *PL2*

| No. | Protein Accession | Protein Name | |
|---|---|---|---|
| 1 | P02741 | C-reactive protein(CRP) | 2 |
| 2 | P05109 | S100 calcium binding protein A8(S100A8) | 1 |
| 3 | P06702 | S100 calcium binding protein A9(S100A9) | 1 |
| 4 | P02750 | Leucine rich alpha-2-glycoprotein 1(LRG1) | 1 |
| 5 | P0DJI8 | Serum amyloid A1(SAA1) | 1 |
| 6 | P0DJI9 | Serum amyloid A2(SAA2) | 1 |
| 7 | P07988 | Surfactant protein B(SFTPB) | 1 |
| 8 | P02774 | GC, vitamin D binding protein(GC) | 0 |
| 9 | P00738 | Haptoglobin(HP) | 0 |
| 10 | P02763 | Orosomucoid 1(ORM1) | 0 |
| 11 | P02790 | Hemopexin(HPX) | 0 |
| 12 | P01011 | Serpin family A member 3(SERPINA3) | 0 |
| 13 | P25311 | Alpha-2-glycoprotein 1, zinc-binding(AZGP1) | 0 |
| 14 | P01009 | Serpin family A member 1(SERPINA1) | 0 |
| 15 | P19652 | Orosomucoid 2(ORM2) | 0 |
| 16 | P02647 | Apolipoprotein A1(APOA1) | 0 |
| 17 | O95445 | Apolipoprotein M(APOM) | -0 |
| 18 | P02766 | Transthyretin(TTR) | -0 |
| 19 | P02652 | Apolipoprotein A2(APOA2) | -0 |
| 20 | P02655 | Apolipoprotein C2(APOC2) | -0 |
| 21 | P01871 | Immunoglobulin heavy constant mu(IGHM) | -0 |
| 22 | P02654 | Apolipoprotein C1(APOC1) | -0 |
| 23 | P06727 | Apolipoprotein A4(APOA4) | -0 |
| 24 | P02656 | apolipoprotein C3(APOC3) | -0 |
| 25 | P02753 | retinol binding protein 4(RBP4) | -1 |

## Annexe 5. Proteins significantly regulated from segment 4 fully quantified and common to sets A, B and C by *PL3*

| No. | Protein Accession | Protein Name | logFC | AveExpr | t | *p* value | Adj. *p* value | B |
|---|---|---|---|---|---|---|---|---|
| 1 | P02741 | C-reactive protein(CRP) | 2.155535 | 10.76869 | 8.63115 | 1.10E-08 | 4.67E-06 | 10.10343 |
| 2 | P05109 | S100 calcium binding protein A8(S100A8) | 1.730152 | 11.36471 | 5.828895 | 6.02E-06 | 0.000336 | 3.896342 |
| 3 | P06702 | S100 calcium binding protein A9(S100A9) | 1.489748 | 11.05066 | 5.642339 | 9.47E-06 | 0.000367 | 3.449582 |
| 4 | P02750 | Leucine rich alpha-2-glycoprotein 1(LRG1) | 1.377648 | 11.53352 | 7.451894 | 1.38E-07 | 2.93E-05 | 7.619397 |
| 5 | P0DJI8 | Serum amyloid A1(SAA1) | 1.308297 | 10.5955 | 5.710012 | 8.03E-06 | 0.000342 | 3.611995 |
| 6 | P78352 | Discs large MAGUK scaffold protein 4(DLG4) | 1.282465 | 13.86823 | 6.174604 | 2.62E-06 | 0.000223 | 4.715464 |
| 7 | P80511 | S100 calcium binding protein A12(S100A12) | 1.255448 | 11.75491 | 3.376156 | 0.002593 | 0.015559 | -2.02476 |
| 8 | P14555 | Phospholipase A2 group IIA(PLA2G2A) | 1.168357 | 11.14641 | 3.67997 | 0.001234 | 0.009192 | -1.31223 |
| 9 | P0DJI9 | Serum amyloid A2(SAA2) | 1.118803 | 10.57383 | 4.146233 | 0.000388 | 0.00435 | -0.19192 |
| 10 | Q9BXR6 | Complement factor H related 5(CFHR5) | 1.09874 | 10.68526 | 5.841979 | 5.83E-06 | 0.000336 | 3.927559 |
| 11 | P07988 | Surfactant protein B(SFTPB) | 1.080633 | 12.24775 | 6.335411 | 1.79E-06 | 0.000191 | 5.092154 |
| 12 | P02774 | GC, vitamin D binding protein(GC) | 0.869177 | 11.28326 | 5.484812 | 1.39E-05 | 0.000448 | 3.070102 |
| 13 | Q02985 | Complement factor H related 3(CFHR3) | 0.846279 | 11.88127 | 2.160717 | 0.041325 | 0.106693 | -4.59584 |
| 14 | P00738 | Haptoglobin(HP) | 0.816506 | 10.90906 | 5.269922 | 2.36E-05 | 0.000559 | 2.549611 |
| 15 | P18428 | Lipopolysaccharide binding protein(LBP) | 0.773199 | 11.47646 | 4.211746 | 0.00033 | 0.003916 | -0.0328 |
| 16 | Q9NPH3 | Interleukin 1 receptor accessory protein(IL1RAP) | 0.771159 | 10.72791 | 3.553975 | 0.001681 | 0.010853 | -1.60984 |
| 17 | Q15485 | Ficolin 2(FCN2) | 0.732214 | 11.32835 | 3.898952 | 0.000718 | 0.006508 | -0.7892 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 18 | P62805 | Histone cluster 1 H4 family member i(HIST1H4I) | 0.721421 | 12.18652 | 2.303763 | 0.030569 | 0.088586 | -4.32602 |
| 19 | P02748 | Complement C9(C9) | 0.71171 | 11.73737 | 5.786215 | 6.67E-06 | 0.000336 | 3.794406 |
| 20 | P08637 | Fc fragment of igg receptor IIia(FCGR3A) | 0.701374 | 11.39923 | 3.934121 | 0.000658 | 0.006229 | -0.70463 |
| 21 | Q9UGM5 | Fetuin B(FETUB) | 0.696432 | 10.46962 | 4.689032 | 1.00E-04 | 0.001775 | 1.132228 |
| 22 | P02763 | Orosomucoid 1(ORM1) | 0.693382 | 11.17915 | 5.434071 | 1.58E-05 | 0.000448 | 2.947473 |
| 23 | P27918 | Complement factor properdin(CFP) | 0.675854 | 10.66218 | 3.700404 | 0.001173 | 0.009089 | -1.26371 |
| 24 | P50281 | Matrix metallopeptidase 14(MMP14) | 0.671028 | 11.3425 | 3.26178 | 0.003418 | 0.018631 | -2.28798 |
| 25 | P15907 | ST6 beta-galactoside alpha-2,6-sialyltransferase 1(ST6GAL1) | 0.670154 | 12.04035 | 4.210051 | 0.000331 | 0.003916 | -0.03692 |
| 26 | P02790 | Hemopexin(HPX) | 0.663327 | 11.4246 | 4.409915 | 0.000201 | 0.002863 | 0.450013 |
| 27 | P14151 | Selectin L(SELL) | 0.655106 | 11.43208 | 4.408364 | 0.000202 | 0.002863 | 0.446227 |
| 28 | P01011 | Serpin family A member 3(SERPINA3) | 0.639662 | 11.78433 | 5.405874 | 1.69E-05 | 0.00045 | 2.879251 |
| 29 | P25311 | Alpha-2-glycoprotein 1, zinc-binding(AZGP1) | 0.632819 | 11.19054 | 4.121613 | 0.000413 | 0.004507 | -0.25163 |
| 30 | P18065 | Insulin like growth factor binding protein 2(IGFBP2) | 0.621648 | 11.24948 | 3.874109 | 0.000764 | 0.006778 | -0.84886 |
| 31 | P28062 | Proteasome subunit beta 8(PSMB8) | 0.600077 | 11.96528 | 3.655144 | 0.001312 | 0.009471 | -1.37108 |
| 32 | Q9Y275 | Tumor necrosis factor superfamily member 13b(TNFSF13B) | 0.598276 | 12.64274 | 3.638659 | 0.001366 | 0.009525 | -1.4101 |
| 33 | P00740 | Coagulation factor IX(F9) | 0.595895 | 10.52735 | 2.641103 | 0.014567 | 0.04925 | -3.64931 |
| 34 | P68032 | Actin, alpha, cardiac muscle 1(ACTC1) | 0.591307 | 11.12393 | 2.6904 | 0.01303 | 0.045875 | -3.54617 |
| 35 | P13473 | Lysosomal associated membrane protein 2(LAMP2) | 0.585016 | 12.18901 | 3.115937 | 0.004843 | 0.023446 | -2.6188 |
| 36 | P60709 | Actin beta(ACTB) | 0.56933 | 11.16013 | 2.991225 | 0.006503 | 0.028975 | -2.89687 |
| 37 | O75015 | Fc fragment of igg receptor iiib(FCGR3B) | 0.561708 | 12.90186 | 2.333731 | 0.028668 | 0.084224 | -4.26812 |

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| 38 | P01042 | Kininogen 1(KNG1) | 0.561187 | 11.38211 | 3.246314 | 0.003547 | 0.018655 | -2.32332 |
| 39 | Q86YT9 | Junction adhesion molecule like(JAML) | 0.545115 | 11.45958 | 2.337772 | 0.02842 | 0.084076 | -4.26027 |
| 40 | Q8WU39 | Marginal zone B and B1 cell specific protein(MZB1) | 0.538274 | 11.86275 | 2.707977 | 0.01252 | 0.0452 | -3.50916 |
| 41 | Q92820 | Gamma-glutamyl hydrolase(GGH) | 0.521311 | 11.36357 | 2.161998 | 0.041215 | 0.106693 | -4.59347 |
| 42 | P13598 | Intercellular adhesion molecule 2(ICAM2) | 0.503563 | 11.8001 | 2.214747 | 0.036913 | 0.101518 | -4.49523 |
| 43 | Q10588 | Bone marrow stromal cell antigen 1(BST1) | 0.497329 | 11.40364 | 3.394763 | 0.002479 | 0.015174 | -1.98165 |
| 44 | P02671 | Fibrinogen alpha chain(FGA) | 0.479315 | 11.24331 | 4.895872 | 5.97E-05 | 0.001211 | 1.637972 |
| 45 | P24387 | Corticotropin releasing hormone binding protein(CRHBP) | 0.457691 | 11.3672 | 3.153235 | 0.004432 | 0.021983 | -2.53474 |
| 46 | O95998 | Interleukin 18 binding protein(IL18BP) | 0.447294 | 10.77794 | 2.167402 | 0.040754 | 0.106511 | -4.58348 |
| 47 | Q06033 | Inter-alpha-trypsin inhibitor heavy chain 3(ITIH3) | 0.421305 | 10.89745 | 2.819184 | 0.009705 | 0.037264 | -3.27221 |
| 48 | P01009 | Serpin family A member 1(SERPINA1) | 0.413837 | 11.16872 | 3.191709 | 0.004043 | 0.020751 | -2.44763 |
| 49 | P19652 | Orosomucoid 2(ORM2) | 0.398525 | 11.28305 | 2.853505 | 0.008966 | 0.035365 | -3.19816 |
| 50 | Q16610 | Extracellular matrix pnrotein 1(ECM1) | 0.375813 | 11.21064 | 2.59339 | 0.016215 | 0.053201 | -3.74816 |
| 51 | Q14847 | LIM and SH3 protein 1(LASP1) | 0.374206 | 10.9941 | 2.28947 | 0.031515 | 0.090712 | -4.35347 |
| 52 | P03952 | Kallikrein B1(KLKB1) | 0.306673 | 11.38794 | 2.186812 | 0.039138 | 0.104204 | -4.54745 |
| 53 | P07225 | Protein S (alpha)(PROS1) | 0.301303 | 11.85841 | 2.875339 | 0.008523 | 0.03458 | -3.15082 |
| 54 | P02647 | Apolipoprotein A1(APOA1) | 0.286719 | 11.627 | 2.283386 | 0.031926 | 0.090997 | -4.36512 |
| 55 | P61626 | Lysozyme(LYZ) | 0.281266 | 11.8665 | 2.139231 | 0.043208 | 0.108498 | -4.63539 |
| 56 | P00751 | Complement factor B(CFB) | 0.235266 | 10.99869 | 2.173109 | 0.040273 | 0.105903 | -4.57291 |
| 57 | P35542 | Serum amyloid A4, constitutive(SAA4) | 0.229656 | 11.97836 | 2.211594 | 0.037158 | 0.101518 | -4.50114 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 58 | Q14118 | Dystroglycan 1(DAG1) | 0.21608 | 10.57222 | 2.143468 | 0.04283 | 0.108498 | -4.62761 |
| 59 | Q09666 | AHNAK nucleoprotein(AHNAK) | -0.2314 | 10.50802 | -2.13948 | 0.043185 | 0.108498 | -4.63493 |
| 60 | O00187 | Mannan binding lectin serine peptidase 2(MASP2) | -0.2435 | 10.92151 | -2.20724 | 0.037499 | 0.101749 | -4.5093 |
| 61 | P08493 | Matrix Gla protein(MGP) | -0.2625 | 11.15887 | -2.14603 | 0.042604 | 0.108498 | -4.62291 |
| 62 | P00746 | Complement factor D(CFD) | -0.26675 | 10.92193 | -2.90071 | 0.008035 | 0.033235 | -3.09561 |
| 63 | P06396 | Gelsolin(GSN) | -0.27285 | 11.17761 | -2.79963 | 0.010152 | 0.037938 | -3.31421 |
| 64 | Q96NZ9 | Proline rich acidic protein 1(PRAP1) | -0.28702 | 11.8321 | -2.26079 | 0.033496 | 0.094498 | -4.40824 |
| 65 | P05090 | Apolipoprotein D(APOD) | -0.28712 | 11.35434 | -2.42691 | 0.02343 | 0.073392 | -4.0852 |
| 66 | A0A087WSY6 | Immunoglobulin kappa variable 3D-15 (gene/pseudogene)(IGKV3D-15) | -0.28941 | 10.84584 | -2.37939 | 0.02598 | 0.07794 | -4.17902 |
| 67 | Q9P2X0 | Dolichyl-phosphate mannosyltransferase subunit 3(DPM3) | -0.29717 | 11.07508 | -2.41339 | 0.024131 | 0.073955 | -4.112 |
| 68 | P01023 | Alpha-2-macroglobulin(A2M) | -0.29749 | 11.33114 | -2.18239 | 0.039501 | 0.104518 | -4.55568 |
| 69 | P02760 | Alpha-1-microglobulin/bikunin precursor(AMBP) | -0.30014 | 11.48111 | -2.66768 | 0.013718 | 0.047512 | -3.59382 |
| 70 | P36955 | Serpin family F member 1(SERPINF1) | -0.30885 | 10.89062 | -2.24687 | 0.034497 | 0.096647 | -4.43465 |
| 71 | A0A075B6J9 | Immunoglobulin lambda variable 2-18(IGLV2-18) | -0.31973 | 12.41039 | -2.18812 | 0.039031 | 0.104204 | -4.54502 |
| 72 | P39060 | Collagen type XVIII alpha 1 chain(COL18A1) | -0.32113 | 12.1038 | -2.87639 | 0.008502 | 0.03458 | -3.14853 |
| 73 | P19022 | Cadherin 2(CDH2) | -0.32662 | 11.50907 | -2.81125 | 0.009884 | 0.037264 | -3.28928 |
| 74 | P17900 | GM2 ganglioside activator(GM2A) | -0.32679 | 11.27881 | -2.81698 | 0.009755 | 0.037264 | -3.27696 |
| 75 | P02751 | Fibronectin 1(FN1) | -0.32924 | 10.89849 | -2.34427 | 0.028026 | 0.083489 | -4.24765 |
| 76 | P01703 | Immunoglobulin lambda variable 1-40(IGLV1-40) | -0.33542 | 13.03563 | -2.0918 | 0.047641 | 0.117876 | -4.72178 |
| 77 | P10909 | Clusterin(CLU) | -0.33546 | 11.14296 | -4.21244 | 0.000329 | 0.003916 | -0.03112 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 78 | P20700 | Lamin B1(LMNB1) | -0.34369 | 12.20052 | -2.30464 | 0.030511 | 0.088586 | -4.32432 |
| 79 | Q13232 | NME/NM23 nucleoside diphosphate kinase 3(NME3) | -0.34463 | 12.17341 | -2.45273 | 0.022144 | 0.069877 | -4.03377 |
| 80 | P43251 | Biotinidase(BTD) | -0.34765 | 11.81349 | -2.09937 | 0.046906 | 0.116855 | -4.70808 |
| 81 | P02042 | Hemoglobin subunit delta(HBD) | -0.36494 | 11.88266 | -2.51719 | 0.019213 | 0.061539 | -3.90399 |
| 82 | O43866 | CD5 molecule like(CD5L) | -0.36722 | 11.45606 | -2.96038 | 0.006991 | 0.030081 | -2.9649 |
| 83 | Q6YHK3 | CD109 molecule(CD109) | -0.37159 | 9.393202 | -2.45715 | 0.021931 | 0.06972 | -4.02494 |
| 84 | P55056 | Apolipoprotein C4(APOC4) | -0.3749 | 12.05332 | -2.28169 | 0.032041 | 0.090997 | -4.36836 |
| 85 | Q9H8L6 | Multimerin 2(MMRN2) | -0.37516 | 11.36591 | -3.15267 | 0.004438 | 0.021983 | -2.53603 |
| 86 | P98160 | Heparan sulfate proteoglycan 2(HSPG2) | -0.38147 | 10.72717 | -2.08945 | 0.04787 | 0.117876 | -4.72602 |
| 87 | Q13332 | Protein tyrosine phosphatase, receptor type S(PTPRS) | -0.38193 | 12.8865 | -2.67085 | 0.01362 | 0.047512 | -3.58719 |
| 88 | P51884 | Lumican(LUM) | -0.3848 | 11.92239 | -2.72374 | 0.012079 | 0.04398 | -3.47587 |
| 89 | P30086 | Phosphatidylethanolamine binding protein 1(PEBP1) | -0.38503 | 11.34566 | -2.19248 | 0.038677 | 0.104204 | -4.53689 |
| 90 | P80748 | Immunoglobulin lambda variable 3-21(IGLV3-21) | -0.39907 | 10.79213 | -2.58241 | 0.016618 | 0.054041 | -3.77078 |
| 91 | A0A0C4DH38 | Immunoglobulin heavy variable 5-51(IGHV5-51) | -0.40421 | 11.09357 | -2.86127 | 0.008806 | 0.03507 | -3.18135 |
| 92 | Q9H4G4 | GLI pathogenesis related 2(GLIPR2) | -0.40651 | 11.11526 | -3.15293 | 0.004435 | 0.021983 | -2.53542 |
| 93 | O00468 | Agrin(AGRN) | -0.40845 | 10.02615 | -2.64759 | 0.014355 | 0.048923 | -3.6358 |
| 94 | P68871 | Hemoglobin subunit beta(HBB) | -0.40941 | 11.18293 | -2.93633 | 0.007395 | 0.031191 | -3.01773 |
| 95 | P58166 | Inhibin beta E subunit(INHBE) | -0.41266 | 10.4746 | -2.07451 | 0.049354 | 0.120833 | -4.75293 |
| 96 | A0A0B4J1V0 | Immunoglobulin heavy variable 3-15(IGHV3-15) | -0.4145 | 10.54641 | -2.13823 | 0.043297 | 0.108498 | -4.63723 |
| 97 | P01591 | Joining chain of multimeric iga and igm(JCHAIN) | -0.41533 | 10.91625 | -2.63458 | 0.014783 | 0.049586 | -3.66288 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 98 | P19827 | Inter-alpha-trypsin inhibitor heavy chain 1(ITIH1) | -0.41568 | 11.4254 | -3.95223 | 0.000629 | 0.00609 | -0.66102 |
| 99 | P05452 | C-type lectin domain family 3 member B(CLEC3B) | -0.41938 | 11.01441 | -3.77292 | 0.000981 | 0.008195 | -1.09102 |
| 100 | O95445 | Apolipoprotein M(APOM) | -0.42956 | 11.15343 | -3.40282 | 0.002431 | 0.015174 | -1.96296 |
| 101 | A0A0C4DH68 | Immunoglobulin kappa variable 2-24(IGKV2-24) | -0.43433 | 11.47595 | -2.6956 | 0.012877 | 0.045714 | -3.53523 |
| 102 | A0A0C4DH25 | Immunoglobulin kappa variable 3D-20(IGKV3D-20) | -0.43765 | 12.97624 | -2.21137 | 0.037176 | 0.101518 | -4.50156 |
| 103 | P61916 | NPC intracellular cholesterol transporter 2(NPC2) | -0.43857 | 10.78742 | -2.75776 | 0.011176 | 0.041043 | -3.40368 |
| 104 | A0A0A0MRZ8 | Immunoglobulin kappa variable 3D-11(IGKV3D-11) | -0.44507 | 12.43701 | -3.28191 | 0.003256 | 0.018495 | -2.24187 |
| 105 | P04211 | Immunoglobulin lambda variable 7-43(IGLV7-43) | -0.45476 | 11.28938 | -2.39986 | 0.024852 | 0.075621 | -4.13875 |
| 106 | Q92954 | Proteoglycan 4(PRG4) | -0.4569 | 11.6058 | -2.24394 | 0.034711 | 0.096647 | -4.4402 |
| 107 | P05154 | Serpin family A member 5(SERPINA5) | -0.45941 | 11.2661 | -2.97546 | 0.006748 | 0.029636 | -2.93169 |
| 108 | P19823 | Inter-alpha-trypsin inhibitor heavy chain 2(ITIH2) | -0.45991 | 11.47941 | -3.67737 | 0.001242 | 0.009192 | -1.3184 |
| 109 | P00441 | Superoxide dismutase 1, soluble(SOD1) | -0.46826 | 11.01585 | -3.06525 | 0.005462 | 0.025852 | -2.73237 |
| 110 | P24592 | Insulin like growth factor binding protein 6(IGFBP6) | -0.47325 | 10.75637 | -4.01159 | 0.000543 | 0.005505 | -0.51786 |
| 111 | P24593 | Insulin like growth factor binding protein 5(IGFBP5) | -0.48962 | 12.4793 | -4.01157 | 0.000543 | 0.005505 | -0.51792 |
| 112 | P20933 | Aspartylglucosaminidase(AGA) | -0.51091 | 11.81848 | -3.92424 | 0.000674 | 0.006245 | -0.72842 |
| 113 | Q92876 | Kallikrein related peptidase 6(KLK6) | -0.51216 | 10.78403 | -3.63077 | 0.001393 | 0.009525 | -1.42877 |
| 114 | P02766 | Transthyretin(TTR) | -0.51538 | 10.75924 | -4.6228 | 0.000118 | 0.00201 | 0.970238 |
| 115 | Q96PD5 | Peptidoglycan recognition protein 2(PGLYRP2) | -0.52017 | 11.04553 | -3.35774 | 0.002711 | 0.016043 | -2.06734 |
| 116 | P29622 | Serpin family A member 4(SERPINA4) | -0.52774 | 11.38048 | -2.96628 | 0.006895 | 0.029971 | -2.95191 |
| 117 | O75340 | Programmed cell death 6(PDCD6) | -0.5313 | 10.95382 | -2.78873 | 0.01041 | 0.038562 | -3.33756 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 118 | Q9NQ38 | Serine peptidase inhibitor, Kazal type 5(SPINK5) | -0.53208 | 11.33316 | -3.25391 | 0.003483 | 0.018631 | -2.30597 |
| 119 | A0A0B4J1X8 | Immunoglobulin heavy variable 3-43(IGHV3-43) | -0.53417 | 10.06729 | -3.52258 | 0.001816 | 0.011545 | -1.68355 |
| 120 | Q96S96 | Phosphatidylethanolamine binding protein 4(PEBP4) | -0.53703 | 10.49968 | -2.59284 | 0.016235 | 0.053201 | -3.7493 |
| 121 | P16035 | TIMP metallopeptidase inhibitor 2(TIMP2) | -0.55365 | 11.18436 | -3.7252 | 0.001104 | 0.008872 | -1.20476 |
| 122 | P43652 | Afamin(AFM) | -0.55367 | 10.47774 | -4.55923 | 0.000138 | 0.002266 | 0.814806 |
| 123 | P17936 | Insulin like growth factor binding protein 3(IGFBP3) | -0.55573 | 11.92865 | -3.67427 | 0.001251 | 0.009192 | -1.32574 |
| 124 | P01743 | Immunoglobulin heavy variable 1-46(IGHV1-46) | -0.56793 | 11.205 | -2.81301 | 0.009844 | 0.037264 | -3.28549 |
| 125 | P02652 | Apolipoprotein A2(APOA2) | -0.58763 | 10.77286 | -4.92668 | 5.53E-05 | 0.001192 | 1.713237 |
| 126 | P28827 | Protein tyrosine phosphatase, receptor type M(PTPRM) | -0.58912 | 12.7266 | -3.2845 | 0.003236 | 0.018495 | -2.23595 |
| 127 | O75071 | EF-hand calcium binding domain 14(EFCAB14) | -0.59242 | 10.65819 | -3.39234 | 0.002493 | 0.015174 | -1.98726 |
| 128 | Q16819 | Meprin A subunit alpha(MEP1A) | -0.59883 | 11.42098 | -4.19609 | 0.000343 | 0.003945 | -0.07085 |
| 129 | A0A0B4J1X5 | Immunoglobulin heavy variable 3-74(IGHV3-74) | -0.6033 | 11.00144 | -3.06871 | 0.005417 | 0.025852 | -2.72465 |
| 130 | P02655 | Apolipoprotein C2(APOC2) | -0.60503 | 10.88546 | -3.70153 | 0.00117 | 0.009089 | -1.26103 |
| 131 | P29279 | Connective tissue growth factor(CTGF) | -0.61798 | 11.69531 | -2.69598 | 0.012866 | 0.045714 | -3.53444 |
| 132 | P05019 | Insulin like growth factor 1(IGF1) | -0.62037 | 11.9662 | -2.86113 | 0.008809 | 0.03507 | -3.18164 |
| 133 | P01871 | Immunoglobulin heavy constant mu(IGHM) | -0.62351 | 11.57409 | -3.21532 | 0.003821 | 0.01985 | -2.39398 |
| 134 | P04070 | Protein C, inactivator of coagulation factors Va and viiia(PROC) | -0.62696 | 11.30821 | -2.843 | 0.009186 | 0.035902 | -3.22088 |
| 135 | Q16661 | Guanylate cyclase activator 2B(GUCA2B) | -0.62902 | 10.9391 | -3.03671 | 0.005843 | 0.027054 | -2.796 |
| 136 | A0A0C4DH34 | Immunoglobulin heavy variable 4-28(IGHV4-28) | -0.6303 | 11.62495 | -3.00626 | 0.006277 | 0.028709 | -2.8636 |
| 137 | O94874 | UFM1 specific ligase 1(UFL1) | -0.64395 | 12.50055 | -2.38472 | 0.025682 | 0.077591 | -4.16855 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 138 | Q9GZM5 | Yip1 domain family member 3(YIPF3) | -0.6469 | 13.28264 | -3.96615 | 0.000608 | 0.00602 | -0.62749 |
| 139 | P02452 | Collagen type I alpha 1 chain(COL1A1) | -0.65444 | 11.47643 | -2.60784 | 0.015698 | 0.052246 | -3.71834 |
| 140 | P55290 | Cadherin 13(CDH13) | -0.65518 | 11.64084 | -4.76546 | 8.26E-05 | 0.00153 | 1.319147 |
| 141 | P35443 | Thrombospondin 4(THBS4) | -0.6732 | 11.07112 | -4.06477 | 0.000475 | 0.005063 | -0.38931 |
| 142 | P81605 | Dermcidin(DCD) | -0.6733 | 11.02346 | -5.27171 | 2.35E-05 | 0.000559 | 2.553952 |
| 143 | P02654 | Apolipoprotein C1(APOC1) | -0.69471 | 11.35837 | -3.27424 | 0.003317 | 0.018591 | -2.25945 |
| 144 | Q08554 | Desmocollin 1(DSC1) | -0.6954 | 11.37751 | -3.00237 | 0.006335 | 0.028709 | -2.87221 |
| 145 | P39059 | Collagen type XV alpha 1 chain(COL15A1) | -0.70432 | 11.43651 | -3.86169 | 0.000788 | 0.006847 | -0.87864 |
| 146 | P23083 | Immunoglobulin heavy variable 1/OR15-1 (non-functional)(IGHV1OR15-1) | -0.71466 | 11.9046 | -3.7455 | 0.00105 | 0.008601 | -1.15642 |
| 147 | P02144 | Myoglobin(MB) | -0.73068 | 11.62824 | -2.98948 | 0.00653 | 0.028975 | -2.90074 |
| 148 | Q76M96 | Coiled-coil domain containing 80(CCDC80) | -0.73838 | 12.02684 | -2.55971 | 0.017482 | 0.056419 | -3.81736 |
| 149 | P07602 | Prosaposin(PSAP) | -0.7491 | 10.98681 | -2.41544 | 0.024024 | 0.073955 | -4.10795 |
| 150 | Q15828 | Cystatin E/M(CST6) | -0.75097 | 10.80648 | -4.28858 | 0.000272 | 0.003619 | 0.154158 |
| 151 | P12273 | Prolactin induced protein(PIP) | -0.75136 | 10.57002 | -3.12039 | 0.004792 | 0.023446 | -2.60878 |
| 152 | P01766 | Immunoglobulin heavy variable 3-13(IGHV3-13) | -0.75766 | 12.01079 | -2.9479 | 0.007198 | 0.030663 | -2.99233 |
| 153 | O60888 | Cuta divalent cation tolerance homolog(CUTA) | -0.7684 | 11.56088 | -4.52419 | 0.000151 | 0.002381 | 0.729151 |
| 154 | Q15166 | Paraoxonase 3(PON3) | -0.76959 | 10.96242 | -3.82051 | 0.000872 | 0.007431 | -0.9773 |
| 155 | Q9UMX5 | Neudesin neurotrophic factor(NENF) | -0.77554 | 11.79971 | -5.76092 | 7.09E-06 | 0.000336 | 3.733913 |
| 156 | P01614 | Immunoglobulin kappa variable 2-40(IGKV2-40) | -0.82224 | 11.38278 | -3.25202 | 0.003499 | 0.018631 | -2.31029 |
| 157 | Q07507 | Dermatopontin(DPT) | -0.83765 | 12.76099 | -2.41487 | 0.024053 | 0.073955 | -4.10907 |

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| 158 | P06727 | Apolipoprotein A4(APOA4) | -0.83936 | 11.74539 | -4.32277 | 0.00025 | 0.003431 | 0.237458 |
| 159 | Q6UXH0 | Angiopoietin like 8(ANGPTL8) | -0.8562 | 12.12427 | -3.26175 | 0.003418 | 0.018631 | -2.28805 |
| 160 | Q14623 | Indian hedgehog(IHH) | -0.88224 | 11.28212 | -2.9007 | 0.008036 | 0.033235 | -3.09564 |
| 161 | P30990 | Neurotensin(NTS) | -0.88902 | 11.8986 | -3.62616 | 0.001409 | 0.009525 | -1.43967 |
| 162 | A0A0C4DH31 | Immunoglobulin heavy variable 1-18(IGHV1-18) | -0.92856 | 10.17756 | -3.28915 | 0.0032 | 0.018495 | -2.22527 |
| 163 | P02656 | Apolipoprotein C3(APOC3) | -0.96146 | 11.10896 | -5.4383 | 1.56E-05 | 0.000448 | 2.957709 |
| 164 | Q7KZ85 | SPT6 homolog, histone chaperone(SUPT6H) | -0.99713 | 13.77179 | -4.27649 | 0.00028 | 0.003619 | 0.124715 |
| 165 | A0A0A0MS14 | Immunoglobulin heavy variable 1-45(IGHV1-45) | -1.0075 | 11.33805 | -3.60953 | 0.001467 | 0.009767 | -1.47896 |
| 166 | P02753 | Retinol binding protein 4(RBP4) | -1.03432 | 11.32415 | -6.72774 | 7.14E-07 | 0.000101 | 5.998467 |
| 167 | Q13103 | Secreted phosphoprotein 2(SPP2) | -1.04046 | 11.80969 | -5.46475 | 1.46E-05 | 0.000448 | 3.021642 |
| 168 | A0A0C4DH29 | Immunoglobulin heavy variable 1-3(IGHV1-3) | -1.07394 | 10.80289 | -3.58914 | 0.001543 | 0.01011 | -1.52705 |
| 169 | A0A0B4J1V2 | Immunoglobulin heavy variable 2-26(IGHV2-26) | -1.08847 | 12.20977 | -2.65767 | 0.014033 | 0.048208 | -3.61476 |
| 170 | A0A075B6H9 | Immunoglobulin lambda variable 4-69(IGLV4-69) | -1.12357 | 11.12487 | -4.4115 | 0.0002 | 0.002863 | 0.453873 |
| 171 | P01742 | Immunoglobulin heavy variable 1-69(IGHV1-69) | -1.19345 | 11.96599 | -3.04963 | 0.005667 | 0.026529 | -2.76722 |
| 172 | P23280 | Carbonic anhydrase 6(CA6) | -1.20511 | 11.18524 | -4.84817 | 6.72E-05 | 0.001302 | 1.521391 |
| 173 | A0A0B4J1U7 | Immunoglobulin heavy variable 6-1(IGHV6-1) | -1.22754 | 13.11368 | -3.63733 | 0.00137 | 0.009525 | -1.41325 |
| 174 | Q6DHV5 | Coiled-coil and C2 domain containing 2B(CC2D2B) | -1.36697 | 13.59498 | -4.92173 | 5.60E-05 | 0.001192 | 1.701144 |

## Annexe 6. Common proteins with reported signatures

| Authors | Year | Ethnicity | Proposed Candidates |
|---|---|---|---|
| Agranoff, D., *et al.* **(114)** | 2006 | African | **SAA1**<br>**TTR**<br>**CRP**<br>Neopterin |
| Liu, J., *et al.* **(121)** | 2013 | Chinese | Four peaks (2554,6; 4824,4; 5325,7; and 8606,8 Da) One of them identified as **Fibrinogen** |
| Song, S. H., *et al.* **(150)** | 2014 | No information Samples collected in Seoul | **SERPINA1**<br>SERPINC1 |
| Xu, D. D., *et. al.* **(36)** | 2014 | Chinese | **APOCII**<br>CD5L<br>HABP2<br>**RBP4** |
| Xu, D., *et al.* **(119)** | 2015 | Chinese | **S100-A9**<br>SOD3<br>MMP9 |
| Achkar, J. M., *et al.* **(117)** | 2015 | Asian/Black/Hispanic/ Caucasian (NYC) | CD14, SEPP1, **SELL**, TNXB, LUM, PEPD, QSOX1, **COMP**, **APOC1** (HIV-)<br>CD14, SEPP1, PGLYRP2, PFN1, VASN, CPN2, TAGLN2, **IGFBP6** (HIV+) |
| Li, C., *et al.* **(151)** | 2015 | Chinese | SHBG |
| Jiang, T., *et al.*,**(152)** | 2017 | Chinese | **SAA**<br>PROZ<br>C4BPB |
| Cheng, C., *et al.* **(153)** | 2018 | Chinese | ENG (HIV+) |

# REFERENCES

1. World Health Organization. Global Tuberculosis Report 2017. Geneva, Switzerland: WHO; 2017.

2. Pai M. Innovations in Tuberculosis Diagnostics: Progress and Translational Challenges. EBioMedicine. 2015;2(3):182-3.

3. Lawn SD, Mwaba P, Bates M, Piatek A, Alexander H, Marais BJ, et al. Advances in tuberculosis diagnostics: the Xpert MTB/RIF assay and future prospects for a point-of-care test. Lancet Infect Dis. 2013;13(4):349-61.

4. Keeler E, Perkins MD, Small P, Hanson C, Reed S, Cunningham J, et al. Reducing the global burden of tuberculosis: the contribution of improved diagnostics. Nature. 2006;444 Suppl 1:49-57.

5. Maertzdorf J, Kaufmann SH, Weiner J, 3rd. Toward a unified biosignature for tuberculosis. Cold Spring Harb Perspect Med. 2015;5(1):a018531.

6. Lonnroth K, Castro KG, Chakaya JM, Chauhan LS, Floyd K, Glaziou P, et al. Tuberculosis control and elimination 2010-50: cure, care, and social development. Lancet. 2010;375(9728):1814-29.

7. Pai M, Schito M. Tuberculosis diagnostics in 2015: landscape, priorities, needs, and prospects. J Infect Dis. 2015;211 Suppl 2:S21-8.

8. Houben RM, Dodd PJ. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. PLoS Med. 2016;13(10):e1002152.

9. O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MP. The immune response in tuberculosis. Annu Rev Immunol. 2013;31:475-527.

10. Walzl G, Ronacher K, Hanekom W, Scriba TJ, Zumla A. Immunological biomarkers of tuberculosis. Nat Rev Immunol. 2011;11(5):343-54.

11. Lin PL, Flynn JL. Understanding latent tuberculosis: a moving target. J Immunol. 2010;185(1):15-22.

12. Elkington PT, Ugarte-Gil CA, Friedland JS. Matrix metalloproteinases in tuberculosis. Eur Respir J. 2011;38(2):456-64.

13. World Health Organization. Trade, foreign policy, diplomacy and health: 3. Tuberculosis control 2018 [Available from: http://www.who.int/trade/distance_learning/gpgh/gpgh3/en/index7.html.

14. Sathyamoorthy T, Sandhu G, Tezera LB, Thomas R, Singhania A, Woelk CH, et al. Gender-dependent differences in plasma matrix metalloproteinase-8 elevated in pulmonary tuberculosis. PLoS One. 2015;10(1):e0117605.

15. Coussens AK, Wilkinson RJ, Nikolayevskyy V, Elkington PT, Hanifa Y, Islam K, et al. Ethnic variation in inflammatory profile in tuberculosis. PLoS Pathog. 2013;9(7):e1003468.

16. Neyrolles O, Quintana-Murci L. Sexual inequality in tuberculosis. PLoS Med. 2009;6(12):e1000199.

17. Moller M, Hoal EG. Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. Tuberculosis (Edinb). 2010;90(2):71-83.

18. Nhamoyebonde S, Leslie A. Biological Differences Between the Sexes and Susceptibility to Tuberculosis. Journal of Infectious Diseases. 2014;209:S100-S6.

19. Zhao Y, Ying H, Demei J, Xie J. Tuberculosis and sexual inequality: the role of sex hormones in immunity. Crit Rev Eukaryot Gene Expr. 2012;22(3):233-41.

20.Noppert GA, Wilson ML, Clarke P, Ye W, Davidson P, Yang Z. Race and nativity are major determinants of tuberculosis in the U.S.: evidence of health disparities in tuberculosis incidence in Michigan, 2004-2012. BMC Public Health. 2017;17(1):538.
21.Nahid P, Horne DJ, Jarlsberg LG, Reiner AP, Osmond D, Hopewell PC, et al. Racial differences in tuberculosis infection in United States communities: the coronary artery risk development in young adults study. Clin Infect Dis. 2011;53(3):291-4.
22.Stead WW, Senner JW, Reddick WT, Lofgren JP. Racial differences in susceptibility to infection by Mycobacterium tuberculosis. N Engl J Med. 1990;322(7):422-7.
23.Kushigemachi M, Schneiderman LJ, Barrett-Connor E. Racial differences in susceptibility to tuberculosis: risk of disease after infection. J Chronic Dis. 1984;37(11):853-62.
24.Weyer K. Discovery, innovation, and new frontiers in tuberculosis diagnostics: reflections and expectations. J Infect Dis. 2015;211 Suppl 2:S78-80.
25.Denkinger CM, Kik SV, Cirillo DM, Casenghi M, Shinnick T, Weyer K, et al. Defining the needs for next generation assays for tuberculosis. J Infect Dis. 2015;211 Suppl 2:S29-38.
26.Lin HH, Dowdy D, Dye C, Murray M, Cohen T. The impact of new tuberculosis diagnostics on transmission: why context matters. Bull World Health Organ. 2012;90(10):739-47A.
27.Kik SV, Denkinger CM, Casenghi M, Vadnais C, Pai M. Tuberculosis diagnostics: which target product profiles should be prioritised? Eur Respir J. 2014;44(2):537-40.
28.Churchyard G, Kim P, Shah NS, Rustomjee R, Gandhi N, Mathema B, et al. What We Know About Tuberculosis Transmission: An Overview. Journal of Infectious Diseases. 2017;216:S629-S35.
29.World Health Organization. The End TB Strategy. Geneva, Switzerland: WHO; 2014.
30.Wallis RS, Kim P, Cole S, Hanna D, Andrade BB, Maeurer M, et al. Tuberculosis biomarkers discovery: developments, needs, and challenges. Lancet Infect Dis. 2013;13(4):362-72.
31.Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR, 3rd. Protein analysis by shotgun/bottom-up proteomics. Chem Rev. 2013;113(4):2343-94.
32.Parker CE, Borchers CH. Mass spectrometry based biomarker discovery, verification, and validation--quality assurance and control of protein biomarker assays. Mol Oncol. 2014;8(4):840-58.
33.Dayona LaK, M. Proteomics of human plasma: A critical comparison of analytical workflows in terms of effort, throughput and outcome. EuPa Open Proteomics I. 2013:8-16.
34.Cunningham R, Ma D, Li L. Mass Spectrometry-based Proteomics and Peptidomics for Systems Biology and Biomarker Discovery. Front Biol (Beijing). 2012;7(4):313-35.
35.Garbis SD, Roumeliotis TI, Tyritzis SI, Zorpas KM, Pavlakis K, Constantinides CA. A novel multidimensional protein identification technology approach combining protein size exclusion prefractionation, peptide zwitterion-ion hydrophilic interaction chromatography, and nano-ultraperformance RP chromatography/nESI-MS2 for the in-depth analysis of the serum proteome and phosphoproteome: application to clinical sera derived from humans with benign prostate hyperplasia. Anal Chem. 2011;83(3):708-18.
36.Xu DD, Deng DF, Li X, Wei LL, Li YY, Yang XY, et al. Discovery and identification of serum potential biomarkers for pulmonary tuberculosis using iTRAQ-coupled two-dimensional LC-MS/MS. Proteomics. 2014;14(2-3):322-31.
37.Xu D, Li Y, Li X, Wei LL, Pan Z, Jiang TT, et al. Serum protein S9, SOD3 and MMP9 as new diagnostic biomarkers for pulmonary tuberculosis by iTRAQ-coupled two-dimensional LC-MS/MS. Proteomics. 2014.
38.Doherty M, Wallis RS, Zumla A, group WH-TDRECjec. Biomarkers for tuberculosis disease status and diagnosis. Curr Opin Pulm Med. 2009;15(3):181-7.

39.Hunter DJ, Losina E, Guermazi A, Burstein D, Lassere MN, Kraus V. A pathway and approach to biomarker validation and qualification for osteoarthritis clinical trials. Curr Drug Targets. 2010;11(5):536-45.
40.Doust J. Qualification versus validation of biomarkers. Scand J Clin Lab Invest Suppl. 2010;242:40-3.
41.Daniel TM. The history of tuberculosis. Respir Med. 2006;100(11):1862-70.
42.Smith I. Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. Clin Microbiol Rev. 2003;16(3):463-96.
43.Hershkovitz I, Donoghue HD, Minnikin DE, Besra GS, Lee OY, Gernaey AM, et al. Detection and molecular characterization of 9,000-year-old Mycobacterium tuberculosis from a Neolithic settlement in the Eastern Mediterranean. PLoS One. 2008;3(10):e3426.
44.Gordon SV, Bottai D, Simeone R, Stinear TP, Brosch R. Pathogenicity in the tubercle bacillus: molecular and evolutionary determinants. Bioessays. 2009;31(4):378-88.
45.Patnership ST. The stop TB strategy Geneva2015 [Available from: http://www.stoptb.org/about/contact.asp.
46.Lawn SD, Zumla AI. Tuberculosis. Lancet. 2011;378(9785):57-72.
47.Prapruttam D, Hedgire SS, Mani SE, Chandramohan A, Shyamkumar NK, Harisinghani M. Tuberculosis--the great mimicker. Semin Ultrasound CT MR. 2014;35(3):195-214.
48.Elkington PT. Tuberculosis: time for a new perspective? J Infect. 2013;66(4):299-302.
49.Philips JA, Ernst JD. Tuberculosis pathogenesis and immunity. Annu Rev Pathol. 2012;7:353-84.
50.Nunes-Alves C, Booty MG, Carpenter SM, Jayaraman P, Rothchild AC, Behar SM. In search of a new paradigm for protective immunity to TB. Nat Rev Microbiol. 2014;12(4):289-99.
51.Ramakrishnan L. Revisiting the role of the granuloma in tuberculosis. Nat Rev Immunol. 2012;12(5):352-66.
52.Elkington PT, Friedland JS. Permutations of time and place in tuberculosis. Lancet Infectious Diseases. 2015;15(11):1357-60.
53.Al Shammari B, Shiomi T, Tezera L, Bielecka MK, Workman V, Sathyamoorthy T, et al. The Extracellular Matrix Regulates Granuloma Necrosis in Tuberculosis. J Infect Dis. 2015.
54.Wallis RS, Pai M, Menzies D, Doherty TM, Walzl G, Perkins MD, et al. Biomarkers and diagnostics for tuberculosis: progress, needs, and translation into practice. Lancet. 2010;375(9729):1920-37.
55.McNerney R, Maeurer M, Abubakar I, Marais B, McHugh TD, Ford N, et al. Tuberculosis diagnostics and biomarkers: needs, challenges, recent advances, and opportunities. J Infect Dis. 2012;205 Suppl 2:S147-58.
56.Trajman A, Steffen RE, Menzies D. Interferon-Gamma Release Assays versus Tuberculin Skin Testing for the Diagnosis of Latent Tuberculosis Infection: An Overview of the Evidence. Pulm Med. 2013;2013:601737.
57.Pai M, Behr MA, Dowdy D, Dheda K, Divangahi M, Boehme CC, et al. Tuberculosis. Nat Rev Dis Primers. 2016;2:16076.
58.World Health Organization. Global Tuberculosis Report 2014. Geneva, Switzerland: WHO; 2014.
59.Steingart KR, Henry M, Ng V, Hopewell PC, Ramsay A, Cunningham J, et al. Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. Lancet Infect Dis. 2006;6(9):570-81.
60.WHO. Xpert MTB/RIF Implementation Manual: Technical and Operational 'How-To'; Practical Considerations. WHO Guidelines Approved by the Guidelines Review Committee. Geneva2014.

61.Cox HS, Mbhele S, Mohess N, Whitelaw A, Muller O, Zemanay W, et al. Impact of Xpert MTB/RIF for TB diagnosis in a primary care clinic with high TB and HIV prevalence in South Africa: a pragmatic randomised trial. PLoS Med. 2014;11(11):e1001760.
62.Piatek AS, Van Cleeff M, Alexander H, Coggin WL, Rehr M, Van Kampen S, et al. GeneXpert for TB diagnosis: planned and purposeful implementation. Glob Health Sci Pract. 2013;1(1):18-23.
63.Huh HJ, Jeong BH, Jeon K, Koh WJ, Ki CS, Lee NY. Performance evaluation of the Xpert MTB/RIF assay according to its clinical application. BMC Infect Dis. 2014;14:589.
64.WHO. Systematic Screening for Active Tuberculosis: Principles and Recommendations. Geneva2013.
65.Yuen CM, Amanullah F, Dharmadhikari A, Nardell EA, Seddon JA, Vasilyeva I, et al. Turning off the tap: stopping tuberculosis transmission through active case-finding and prompt effective treatment. Lancet. 2015.
66.McNerney R, Daley P. Towards a point-of-care test for active tuberculosis: obstacles and opportunities. Nat Rev Microbiol. 2011;9(3):204-13.
67.Cox H, Nicol MP. Tuberculosis eradication: renewed commitment and global investment required. Lancet Infect Dis. 2018;18(3):228-9.
68.Huddart S, MacLean E, Pai M. Location, location, location: tuberculosis services in highest burden countries. Lancet Glob Health. 2016;4(12):e907-e8.
69.Lawn SD. Advances in Diagnostic Assays for Tuberculosis. Cold Spring Harb Perspect Med. 2015;5(12).
70.Organization WH. Global tuberculosis report 2013. Geneva, Switzerland: World Health Organization; 2013.
71.Paris L, Magni R, Zaidi F, Araujo R, Saini N, Harpole M, et al. Urine lipoarabinomannan glycan in HIV-negative patients with pulmonary tuberculosis correlates with disease severity. Sci Transl Med. 2017;9(420).
72.Steingart KR, Flores LL, Dendukuri N, Schiller I, Laal S, Ramsay A, et al. Commercial serological tests for the diagnosis of active pulmonary and extrapulmonary tuberculosis: an updated systematic review and meta-analysis. PLoS Med. 2011;8(8):e1001062.
73.Poste G. Bring on the biomarkers. Nature. 2011;469(7329):156-7.
74.Drucker E, Krapfenbauer K. Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. EPMA J. 2013;4(1):7.
75.Mayeux R. Biomarkers: potential uses and limitations. NeuroRx. 2004;1(2):182-8.
76.Biomarkers Definitions Working G. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69(3):89-95.
77.Naylor S. Biomarkers: current perspectives and future prospects. Expert Rev Mol Diagn. 2003;3(5):525-9.
78.Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. Nat Rev Drug Discov. 2003;2(7):566-80.
79.Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst. 2001;93(14):1054-61.
80.Pavlou MP, Diamandis EP, Blasutig IM. The long journey of cancer biomarkers from the bench to the clinic. Clin Chem. 2013;59(1):147-57.
81.Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol. 2006;24(8):971-83.
82.Issaq HJ, Waybright TJ, Veenstra TD. Cancer biomarker discovery: Opportunities and pitfalls in analytical methods. Electrophoresis. 2011;32(9):967-75.
83.Phillips KA, Van Bebber S, Issa AM. Diagnostics and biomarker development: priming the pipeline. Nat Rev Drug Discov. 2006;5(6):463-9.

84.Puntmann VO. How-to guide on biomarkers: biomarker definitions, validation and applications with examples from cardiovascular disease. Postgrad Med J. 2009;85(1008):538-45.
85.Bujang MA, Adnan TH. Requirements for Minimum Sample Size for Sensitivity and Specificity Analysis. J Clin Diagn Res. 2016;10(10):YE01-YE6.
86.Warnock DG, Peck CC. A roadmap for biomarker qualification. Nat Biotechnol. 2010;28(5):444-5.
87.Jani D, Allinson J, Berisha F, Cowan KJ, Devanarayan V, Gleason C, et al. Recommendations for Use and Fit-for-Purpose Validation of Biomarker Multiplex Ligand Binding Assays in Drug Development. AAPS J. 2016;18(1):1-14.
88.Goodsaid FM, Frueh FW, Mattes W. Strategic paths for biomarker qualification. Toxicology. 2008;245(3):219-23.
89.Frantzi M, Bhat A, Latosinska A. Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. Clin Transl Med. 2014;3(1):7.
90.Cummings J, Raynaud F, Jones L, Sugar R, Dive C. Fit-for-purpose biomarker method validation for application in clinical trials of anticancer drugs. Br J Cancer. 2010;103(9):1313-7.
91.Liu Y, Buil A, Collins BC, Gillet LC, Blum LC, Cheng LY, et al. Quantitative variability of 342 plasma proteins in a human twin population. Mol Syst Biol. 2015;11(1):786.
92.Al-Daghri NM, Al-Attas OS, Johnston HE, Singhania A, Alokail MS, Alkharfy KM, et al. Whole Serum 3D LC-nESI-FTMS Quantitative Proteomics Reveals Sexual Dimorphism in the Milieu Interieur of Overweight and Obese Adults. J Proteome Res. 2014;13(11):5094-105.
93.Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861-74.
94.Zumla A, Kim P, Maeurer M, Schito M. Zero deaths from tuberculosis: progress, reality, and hope. Lancet Infect Dis. 2013;13(4):285-7.
95.Ottenhoff TH, Ellner JJ, Kaufmann SH. Ten challenges for TB biomarkers. Tuberculosis (Edinb). 2012;92 Suppl 1:S17-20.
96.Weiner J, 3rd, Kaufmann SH. Recent advances towards tuberculosis control: vaccines and biomarkers. J Intern Med. 2014;275(5):467-80.
97.Dove A. To build better tuberculosis diagnostics, look for 'biosignatures'. Nat Med. 2015;21(7):662.
98.Niemz A, Boyle DS. Nucleic acid testing for tuberculosis at the point-of-care in high-burden countries. Expert Rev Mol Diagn. 2012;12(7):687-701.
99.Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ, et al. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. J Clin Microbiol. 2015;53(7):2230-7.
100. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. 2013;13(2):137-46.
101. Png E, Alisjahbana B, Sahiratmadja E, Marzuki S, Nelwan R, Balabanova Y, et al. A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. BMC Med Genet. 2012;13:5.
102. Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. Lancet Respir Med. 2016;4(3):213-24.
103. Haas CT, Roe JK, Pollara G, Mehta M, Noursadeghi M. Diagnostic 'omics' for active tuberculosis. BMC Med. 2016;14(1):37.
104. Maertzdorf J, Weiner J, 3rd, Kaufmann SH. Enabling biomarkers for tuberculosis control. Int J Tuberc Lung Dis. 2012;16(9):1140-8.
105. Maertzdorf J, McEwen G, Weiner J, 3rd, Tian S, Lader E, Schriek U, et al. Concise gene signature for point-of-care classification of tuberculosis. EMBO Mol Med. 2015;8(2):86-95.

106. Zak DE, Penn-Nicholson A, Scriba TJ, Thompson E, Suliman S, Amon LM, et al. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. Lancet. 2016;387(10035):2312-22.
107. Baughn AD, Rhee KY. Metabolomics of Central Carbon Metabolism in Mycobacterium tuberculosis. Microbiol Spectr. 2014;2(3).
108. Swanepoel CC, Loots du T. The use of functional genomics in conjunction with metabolomics for Mycobacterium tuberculosis research. Dis Markers. 2014;2014:124218.
109. Eoh H. Metabolomics: A window into the adaptive physiology of Mycobacterium tuberculosis. Tuberculosis (Edinb). 2014;94(6):538-43.
110. du Preez I, Loots du T. Altered fatty acid metabolism due to rifampicin-resistance conferring mutations in the rpoB Gene of Mycobacterium tuberculosis: mapping the potential of pharmaco-metabolomics for global health and personalized medicine. OMICS. 2012;16(11):596-603.
111. Das MK, Bishwal SC, Das A, Dabral D, Badireddy VK, Pandit B, et al. Deregulated tyrosine-phenylalanine metabolism in pulmonary tuberculosis patients. J Proteome Res. 2015;14(4):1947-56.
112. Isa F, Collins S, Lee MH, Decome D, Dorvil N, Joseph P, et al. Mass Spectrometric Identification of Urinary Biomarkers of Pulmonary Tuberculosis. EBioMedicine. 2018.
113. Tyanova S, Albrechtsen R, Kronqvist P, Cox J, Mann M, Geiger T. Proteomic maps of breast cancer subtypes. Nat Commun. 2016;7:10259.
114. Agranoff D, Fernandez-Reyes D, Papadopoulos MC, Rojas SA, Herbster M, Loosemore A, et al. Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum. Lancet. 2006;368(9540):1012-21.
115. Schubert OT, Ludwig C, Kogadeeva M, Zimmermann M, Rosenberger G, Gengenbacher M, et al. Absolute Proteome Composition and Dynamics during Dormancy and Resuscitation of Mycobacterium tuberculosis. Cell Host Microbe. 2015;18(1):96-108.
116. Esterhuyse MM, Weiner J, 3rd, Caron E, Loxton AG, Iannaccone M, Wagman C, et al. Epigenetics and Proteomics Join Transcriptomics in the Quest for Tuberculosis Biomarkers. MBio. 2015;6(5).
117. Achkar JM, Cortes L, Croteau P, Yanofsky C, Mentinova M, Rajotte I, et al. Host Protein Biomarkers Identify Active Tuberculosis in HIV Uninfected and Co-infected Individuals. EBioMedicine. 2015;2(9):1160-8.
118. Zhang X, Liu F, Li Q, Jia H, Pan L, Xing A, et al. A proteomics approach to the identification of plasma biomarkers for latent tuberculosis infection. Diagn Microbiol Infect Dis. 2014;79(4):432-7.
119. Xu D, Li Y, Li X, Wei LL, Pan Z, Jiang TT, et al. Serum protein S9, SOD3 and MMP9 as new diagnostic biomarkers for pulmonary tuberculosis by iTRAQ-coupled two-dimensional LC-MS/MS. Proteomics. 2015.
120. Zhou F, Xu X, Wu S, Cui X, Fan L, Pan W. Protein array identification of protein markers for serodiagnosis of Mycobacterium tuberculosis infection. Sci Rep. 2015;5:15349.
121. Liu J, Jiang T, Wei L, Yang X, Wang C, Zhang X, et al. The discovery and identification of a candidate proteomic biomarker of active tuberculosis. BMC Infect Dis. 2013;13:506.
122. Golichenari B, Velonia K, Nosrati R, Nezami A, Farokhi-Fard A, Abnous K, et al. Label-free nano-biosensing on the road to tuberculosis detection. Biosens Bioelectron. 2018;113:124-35.
123. Tsai TT, Huang CY, Chen CA, Shen SW, Wang MC, Cheng CM, et al. Diagnosis of Tuberculosis Using Colorimetric Gold Nanoparticles on a Paper-Based Analytical Device. Acs Sensors. 2017;2(9):1345-54.
124. Sun W, Yuan S, Huang H, Liu N, Tan Y. A label-free biosensor based on localized surface plasmon resonance for diagnosis of tuberculosis. J Microbiol Methods. 2017;142:41-5.

125. Zeidan BA, Townsend PA, Garbis SD, Copson E, Cutress RI. Clinical proteomics and breast cancer. Surgeon. 2015.

126. Haleem. J. IaTDV. Proteomic and Metabolomic Approaches to Biomarker Discovery. Elsevier, editor2013.

127. Jaros JA, Guest PC, Bahn S, Martins-de-Souza D. Affinity depletion of plasma and serum for mass spectrometry-based proteome analysis. Methods Mol Biol. 2013;1002:1-11.

128. Zubarev RA. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics. 2013;13(5):723-6.

129. Inc TFS. Plasma and Serum Preparation 2015 [Available from: www.thermofisher.com/uk/en/home/references/protocols/cell-and-tissue-analysis/elisa-protocol/elisa-sample-preparation-protocols/plasma-and-serum-preparation.html.

130. Rai AJ, Gelfand CA, Haywood BC, Warunek DJ, Yi J, Schuchard MD, et al. HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. Proteomics. 2005;5(13):3262-77.

131. Such-Sanmartin G, Ventura-Espejo E, Jensen ON. Depletion of abundant plasma proteins by poly(N-isopropylacrylamide-acrylic acid) hydrogel particles. Anal Chem. 2014;86(3):1543-50.

132. Simpson RJ, Bernhard OK, Greening DW, Moritz RL. Proteomics-driven cancer biomarker discovery: looking to the future. Curr Opin Chem Biol. 2008;12(1):72-7.

133. Surinova S, Schiess R, Huttenhain R, Cerciello F, Wollscheid B, Aebersold R. On the development of plasma protein biomarkers. J Proteome Res. 2011;10(1):5-16.

134. Scientific T. Thermo Scientific Orbitrap Fusion Lumos Tribrid Mass Spectrometer San Jose, CA USA: Thermo Fisher Scientific Inc; 2015 [updated 2015. Available from: https://www.thermoscientific.com/content/dam/tfs/ATG/CMD/cmd-documents/bro/spec/ms/lcms/orbitrap/PS-64391-LC-MS-Orbitrap-Fusion-Lumos-Tribrid-PS64391-EN.pdf.

135. Geyer PE, Holdt LM, Teupser D, Mann M. Revisiting biomarker discovery by plasma proteomics. Mol Syst Biol. 2017;13(9):942.

136. Hakimi A, Auluck J, Jones GD, Ng LL, Jones DJ. Assessment of reproducibility in depletion and enrichment workflows for plasma proteomics using label-free quantitative data-independent LC-MS. Proteomics. 2014;14(1):4-13.

137. Gundry RL, White MY, Nogee J, Tchernyshyov I, Van Eyk JE. Assessment of albumin removal from an immunoaffinity spin column: critical implications for proteomic examination of the albuminome and albumin-depleted samples. Proteomics. 2009;9(7):2021-8.

138. Yadav AK, Bhardwaj G, Basak T, Kumar D, Ahmad S, Priyadarshini R, et al. A systematic analysis of eluted fraction of plasma post immunoaffinity depletion: implications in biomarker discovery. PLoS One. 2011;6(9):e24442.

139. Koutroukides TA, Guest PC, Leweke FM, Bailey DM, Rahmoune H, Bahn S, et al. Characterization of the human serum depletome by label-free shotgun proteomics. J Sep Sci. 2011;34(13):1621-6.

140. Righetti PG, Boschetti E, Lomas L, Citterio A. Protein Equalizer (TM) Technology: The quest for a "democratic proteome". Proteomics. 2006;6(14):3980-92.

141. Bio-Rad Laboratories I. ProteoMiner Protein Enrichment Kits 2016 [Available from: http://www.bio-rad.com/en-uk/product/proteominer-protein-enrichment-kits.

142. Zhao Y, Chang C, Qin P, Cao Q, Tian F, Jiang J, et al. Mining the human plasma proteome with three-dimensional strategies by high-resolution Quadrupole Orbitrap Mass Spectrometry. Anal Chim Acta. 2016;904:65-75.

143. Kostanski LK, Keller DM, Hamielec AE. Size-exclusion chromatography-a review of calibration methodologies. J Biochem Biophys Methods. 2004;58(2):159-86.

144. GE H. Size Exclusion Chromatography: Principles and Methods. Sweden: GE Healthcare Bio-Sciences AB; 2014.

145. Pocsfalvi G, Stanly C, Fiume I, Vekey K. Chromatography and its hyphenation to mass spectrometry for extracellular vesicle analysis. Journal of Chromatography A. 2016;1439:26-41.

146. Dorsey JG, Dill KA. The Molecular Mechanism of Retention in Reversed-Phase Liquid-Chromatography. Chemical Reviews. 1989;89(2):331-46.

147. Camenzuli M, Schoenmakers PJ. A new measure of orthogonality for multi-dimensional chromatography. Anal Chim Acta. 2014;838:93-101.

148. Gilar M, Fridrich J, Schure MR, Jaworski A. Comparison of orthogonality estimation methods for the two-dimensional separations of peptides. Anal Chem. 2012;84(20):8722-32.

149. Yerlikaya S, Broger T, MacLean E, Pai M, Denkinger CM. A tuberculosis biomarker database: the key to novel TB diagnostics. Int J Infect Dis. 2017;56:253-7.

150. Song SH, Han M, Choi YS, Dan KS, Yang MG, Song J, et al. Proteomic profiling of serum from patients with tuberculosis. Ann Lab Med. 2014;34(5):345-53.

151. Li C, He X, Li H, Zhou Y, Zang N, Hu S, et al. Discovery and verification of serum differential expression proteins for pulmonary tuberculosis. Tuberculosis (Edinb). 2015.

152. Jiang TT, Shi LY, Wei LL, Li X, Yang S, Wang C, et al. Serum amyloid A, protein Z, and C4b-binding protein beta chain as new potential biomarkers for pulmonary tuberculosis. PLoS One. 2017;12(3):e0173304.

153. Chen C, Yan T, Liu L, Wang J, Jin Q. Identification of a Novel Serum Biomarker for Tuberculosis Infection in Chinese HIV Patients by iTRAQ-Based Quantitative Proteomics. Front Microbiol. 2018;9:330.

154. Scriba TJ, Penn-Nicholson A, Shankar S, Hraha T, Thompson EG, Sterling D, et al. Sequential inflammatory processes define human progression from M. tuberculosis infection to tuberculosis disease. Plos Pathogens. 2017;13(11).

155. Hamilton K, Tolfree R, Mytton J. A systematic review of active case-finding strategies for tuberculosis in homeless populations. Int J Tuberc Lung D. 2018;22(10):1135-+.

156. Switzar L, Giera M, Niessen WM. Protein digestion: an overview of the available techniques and recent developments. J Proteome Res. 2013;12(3):1067-77.

157. Yates JR, 3rd. Mass spectrometry and the age of the proteome. J Mass Spectrom. 1998;33(1):1-19.

158. Vandermarliere E, Mueller M, Martens L. Getting intimate with trypsin, the leading protease in proteomics. Mass Spectrom Rev. 2013;32(6):453-65.

159. Brownridge P, Beynon RJ. The importance of the digest: proteolysis and absolute quantification in proteomics. Methods. 2011;54(4):351-60.

160. Tsiatsiani L, Heck AJR. Proteomics beyond trypsin. Febs Journal. 2015;282(14):2612-26.

161. Swaney DL, Wenger CD, Coon JJ. Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. Journal of Proteome Research. 2010;9(3):1323-9.

162. Giansanti P, Tsiatsiani L, Low TY, Heck AJ. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. Nat Protoc. 2016;11(5):993-1006.

163. Lin Y, Zhou J, Bi D, Chen P, Wang XC, Liang SP. Sodium-deoxycholate-assisted tryptic digestion and identification of proteolytically resistant proteins. Analytical Biochemistry. 2008;377(2):259-66.

164. LLC. S-AC. Trypsin: Physical Properties and In Vivo Processing 2016 [Available from: http://www.sigmaaldrich.com/technical-documents/articles/biology/trypsin.html.

165. Getz EB, Xiao M, Chakrabarty T, Cooke R, Selvin PR. A comparison between the sulfhydryl reductants tris(2-carboxyethyl)phosphine and dithiothreitol for use in protein biochemistry. Analytical Biochemistry. 1999;273(1):73-80.

166. Burns JA, Butler JC, Moran J, Whitesides GM. Selective Reduction of Disulfides by Tris(2-Carboxyethyl)Phosphine. Journal of Organic Chemistry. 1991;56(8):2648-50.
167. Laremore T. TCEP or DTT? : Pennsylvania State University; 2014 [Available from: http://sites.psu.edu/msproteomics/2014/05/30/tcep-or-dtt/.
168. Jiang XH, Shamshurin D, Spicer V, Krokhin OV. The effect of various S-alkylating agents on the chromatographic behavior of cysteine-containing peptides in reversed-phase chromatography. Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences. 2013;915:57-63.
169. Liang HPH, Brophy TM, Hogg PJ. Redox properties of the tissue factor Cys(186)-Cys(209) disulfide bond. Biochemical Journal. 2011;437:455-60.
170. Fang P, Liu M, Xue Y, Yao J, Zhang Y, Shen H, et al. Controlling nonspecific trypsin cleavages in LC-MS/MS-based shotgun proteomics using optimized experimental conditions. Analyst. 2015;140(22):7613-21.
171. Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nat Clin Pract Oncol. 2008;5(10):588-99.
172. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198-207.
173. SelectScience. Mass Spectrometry Buying Guide United Kingdom2015 [Available from: http://www.selectscience.net/mass_spectrometry_buying_guide.aspx.
174. Bantscheff M, Lemeer S, Savitski MM, Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal Bioanal Chem. 2012;404(4):939-65.
175. Banerjee S, Mazumdar S. Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte. Int J Anal Chem. 2012;2012:282574.
176. Paizs B, Suhai S. Fragmentation pathways of protonated peptides. Mass Spectrom Rev. 2005;24(4):508-48.
177. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science. 1989;246(4926):64-71.
178. Winger BE, Light-Wahl KJ, Ogorzalek Loo RR, Udseth HR, Smith RD. Observation and implications of high mass-to-charge ratio ions from electrospray ionization mass spectrometry. J Am Soc Mass Spectrom. 1993;4(7):536-45.
179. Kalli A, Smith GT, Sweredoski MJ, Hess S. Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: focus on LTQ-Orbitrap mass analyzers. J Proteome Res. 2013;12(7):3071-86.
180. Papayannopoulos IA. The interpretation of collision-induced dissociation tandem mass spectra of peptides. Mass Spectrometry Reviews. 1995;14(1):49-73.
181. Jedrychowski MP, Huttlin EL, Haas W, Sowa ME, Rad R, Gygi SP. Evaluation of HCD- and CID-type Fragmentation Within Their Respective Detection Platforms For Murine Phosphoproteomics. Molecular & Cellular Proteomics. 2011;10(12).
182. Orbitrap P. Orbitrap Elite Hybrid Ion Trap Mass Spectrometer 2016 [Available from: http://planetorbitrap.com/orbitrap-elite#tab:schematic.
183. Pachl F, Ruprecht B, Lemeer S, Kuster B. Characterization of a high field Orbitrap mass spectrometer for proteome analysis. Proteomics. 2013;13(17):2552-62.
184. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem. 2007;389(4):1017-31.
185. Wasinger VC, Zeng M, Yau Y. Current status and advances in quantitative proteomic mass spectrometry. Int J Proteomics. 2013;2013:180605.
186. Rauniyar N, Yates JR, 3rd. Isobaric labeling-based relative quantification in shotgun proteomics. J Proteome Res. 2014;13(12):5293-309.

187. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002;1(5):376-86.
188. Looso M, Borchardt T, Kruger M, Braun T. Advanced identification of proteins in uncharacterized proteomes by pulsed in vivo stable isotope labeling-based mass spectrometry. Mol Cell Proteomics. 2010;9(6):1157-66.
189. Cambridge SB, Gnad F, Nguyen C, Bermejo JL, Kruger M, Mann M. Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. J Proteome Res. 2011;10(12):5275-84.
190. Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. Super-SILAC mix for quantitative proteomics of human tumor tissue. Nat Methods. 2010;7(5):383-5.
191. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem. 2003;75(8):1895-904.
192. Evans C, Noirel J, Ow SY, Salim M, Pereira-Medrano AG, Couto N, et al. An insight into iTRAQ: where do we stand now? Anal Bioanal Chem. 2012;404(4):1011-27.
193. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 2004;3(12):1154-69.
194. Li Z, Adams RM, Chourey K, Hurst GB, Hettich RL, Pan C. Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos. J Proteome Res. 2012;11(3):1582-90.
195. Shadforth IP, Dunkley TPJ, Lilley KS, Bessant C. i-Tracker: For quantitative proteomics using iTRAQ (TM). Bmc Genomics. 2005;6.
196. Karp NA, Huber W, Sadowski PG, Charles PD, Hester SV, Lilley KS. Addressing accuracy and precision issues in iTRAQ quantitation. Mol Cell Proteomics. 2010;9(9):1885-97.
197. Aggarwal S, Yadav AK. Dissecting the iTRAQ Data Analysis. In: Jung K, editor. Statistical Analysis in Proteomics. New York, NY: Springer New York; 2016. p. 277-91.
198. Zhang Y, Askenazi M, Jiang JR, Luckey CJ, Griffin JD, Marto JA. A Robust Error Model for iTRAQ Quantification Reveals Divergent Signaling between Oncogenic FLT3 Mutants in Acute Myeloid Leukemia. Molecular & Cellular Proteomics. 2010;9(5):780-90.
199. Wenger CD, Lee MV, Hebert AS, McAlister GC, Phanstiel DH, Westphall MS, et al. Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. Nat Methods. 2011;8(11):933-5.
200. McAlister GC, Nusinow DP, Jedrychowski MP, Wuhr M, Huttlin EL, Erickson BK, et al. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal Chem. 2014;86(14):7150-8.
201. Ting L, Rad R, Gygi SP, Haas W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. Nat Methods. 2011;8(11):937-40.
202. Mahoney DW, Therneau TM, Heppelmann CJ, Higgins L, Benson LM, Zenka RM, et al. Relative quantification: characterization of bias, variability and fold changes in mass spectrometry data from iTRAQ-labeled peptides. J Proteome Res. 2011;10(9):4325-33.
203. Levin Y. The role of statistical power analysis in quantitative proteomics. Proteomics. 2011;11(12):2565-7.
204. Duncan MW, Aebersold R, Caprioli RM. The pros and cons of peptide-centric proteomics. Nat Biotechnol. 2010;28(7):659-64.
205. Li YF, Radivojac P. Computational approaches to protein inference in shotgun proteomics. BMC Bioinformatics. 2012;13 Suppl 16:S4.

206. Choi H, Nesvizhskii AI. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. J Proteome Res. 2008;7(1):47-50.

207. Kall L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res. 2008;7(1):40-4.

208. Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. Mol Cell Proteomics. 2015;14(9):2394-404.

209. Cohen Freue GV, Meredith A, Smith D, Bergman A, Sasaki M, Lam KK, et al. Computational biomarker pipeline from discovery to clinical implementation: plasma proteomic biomarkers for cardiac transplantation. PLoS Comput Biol. 2013;9(4):e1002963.

210. Tuck MK, Chan DW, Chia D, Godwin AK, Grizzle WE, Krueger KE, et al. Standard operating procedures for serum and plasma collection: early detection research network consensus statement standard operating procedure integration working group. J Proteome Res. 2009;8(1):113-7.

211. Walker NF, Wilkinson KA, Meintjes G, Tezera LB, Goliath R, Peyper JM, et al. Matrix Degradation in Human Immunodeficiency Virus Type 1-Associated Tuberculosis and Tuberculosis Immune Reconstitution Inflammatory Syndrome: A Prospective Observational Study. Clin Infect Dis. 2017;65(1):121-32.

212. Papachristou EK, Roumeliotis TI, Chrysagi A, Trigoni C, Charvalos E, Townsend PA, et al. The shotgun proteomic study of the human ThinPrep cervical smear using iTRAQ mass-tagging and 2D LC-FT-Orbitrap-MS: the detection of the human papillomavirus at the protein level. J Proteome Res. 2013;12(5):2078-89.

213. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37(Web Server issue):W305-11.

214. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25(8):1091-3.

215. Theocharidis A, van Dongen S, Enright AJ, Freeman TC. Network visualization and analysis of gene expression data using BioLayout Express(3D). Nat Protoc. 2009;4(10):1535-50.

216. Luo R, Zhao H. Protein quantitation using iTRAQ: Review on the sources of variations and analysis of nonrandom missingness. Stat Interface. 2012;5(1):99-107.

217. Chen X, Ge Y. Ultrahigh pressure fast size exclusion chromatography for top-down proteomics. Proteomics. 2013;13(17):2563-6.

218. Hoshino A, Costa-Silva B, Shen TL, Rodrigues G, Hashimoto A, Tesic Mark M, et al. Tumour exosome integrins determine organotropic metastasis. Nature. 2015;527(7578):329-35.

219. Polyansky AA, Zagrovic B. Protein Electrostatic Properties Predefining the Level of Surface Hydrophobicity Change upon Phosphorylation. J Phys Chem Lett. 2012;3(8):973-6.

220. Maes E, Hadiwikarta WW, Mertens I, Baggerman G, Hooyberghs J, Valkenborg D. CONSTANd : a normalization method for isobaric labeled spectra by constrained optimization. Mol Cell Proteomics. 2016.

221. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

222. Pathan M, Keerthikumar S, Ang CS, Gangoda L, Quek CY, Williamson NA, et al. FunRich: An open access standalone functional enrichment and interaction network analysis tool. Proteomics. 2015;15(15):2597-601.

223. Wang C, Liu CM, Wei LL, Shi LY, Pan ZF, Mao LG, et al. A Group of Novel Serum Diagnostic Biomarkers for Multidrug-Resistant Tuberculosis by iTRAQ-2D LC-MS/MS and Solexa Sequencing. Int J Biol Sci. 2016;12(2):246-56.

224. Bouchal P, Roumeliotis T, Hrstka R, Nenutil R, Vojtesek B, Garbis SD. Biomarker discovery in low-grade breast cancer using isobaric stable isotope tags and two-dimensional liquid chromatography-tandem mass spectrometry (iTRAQ-2DLC-MS/MS) based quantitative proteomic analysis. J Proteome Res. 2009;8(1):362-73.
225. Kammers K, Cole RN, Tiengwe C, Ruczinski I. Detecting Significant Changes in Protein Abundance. EuPA Open Proteom. 2015;7:11-9.
226. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.
227. Massanella M, Singhania A, Beliakova-Bethell N, Pier R, Lada SM, White CH, et al. Differential gene expression in HIV-infected individuals following ART. Antiviral Res. 2013;100(2):420-8.
228. Chawla DS. 'One-size-fits-all’ threshold for P values under fire: Nature; 2017 [updated 19 September Available from: https://www.nature.com/news/one-size-fits-all-threshold-for-p-values-under-fire-1.22625.
229. Baker M. Statisticians issue warning over misuse of P values: Nature; 2016 [updated 07 March. Available from: https://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503.
230. Nuzzo R. Scientific method: Statistical errors: Natute; 2014 [updated 12 February. Available from: https://www.nature.com/news/scientific-method-statistical-errors-1.14700.
231. Chang SW, Pan WS, Lozano Beltran D, Oleyda Baldelomar L, Solano MA, Tuero I, et al. Gut hormones, appetite suppression and cachexia in patients with pulmonary TB. PLoS One. 2013;8(1):e54564.
232. Schaible UE, Kaufmann SH. Malnutrition and infection: complex mechanisms and global impacts. PLoS Med. 2007;4(5):e115.
233. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.
234. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
235. Walker NF, Clark SO, Oni T, Andreu N, Tezera L, Singh S, et al. Doxycycline and HIV infection suppress tuberculosis-induced matrix metalloproteinases. Am J Respir Crit Care Med. 2012;185(9):989-97.
236. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-7.
237. Goletti D, Lee MR, Wang JY, Walter N, Ottenhoff THM. Update on tuberculosis biomarkers: From correlates of risk, to correlates of active disease and of cure from disease. Respirology. 2018.
238. Garcia-Basteiro AL, DiNardo A, Saavedra B, Silva DR, Palmero D, Gegia M, et al. Point of care diagnostics for tuberculosis. Pulmonology. 2018;24(2):73-85.
239. Wejse C. Point-of-care diagnostics for tuberculosis elimination? Lancet. 2014;383(9915):388-90.
240. Johnston HE, Carter MJ, Cox KL, Dunscombe M, Manousopoulou A, Townsend PA, et al. Integrated Cellular and Plasma Proteomics of Contrasting B-cell Cancers Reveals Common, Unique and Systemic Signatures. Mol Cell Proteomics. 2017.
241. Larkin SE, Johnston HE, Jackson TR, Jamieson DG, Roumeliotis TI, Mockridge CI, et al. Detection of candidate biomarkers of prostate cancer progression in serum: a depletion-free 3D LC/MS quantitative proteomics pilot study. Br J Cancer. 2016;115(9):1078-86.
242. Zeidan B, Manousopoulou A, Garay-Baquero DJ, White CH, Larkin SET, Potter KN, et al. Increased circulating resistin levels in early-onset breast cancer patients of normal body mass index

correlate with lymph node negative involvement and longer disease free survival: a multi-center POSH cohort serum proteomics study. Breast Cancer Res. 2018;20(1):19.
243. Rost HL, Malmstrom L, Aebersold R. Reproducible quantitative proteotype data matrices for systems biology. Mol Biol Cell. 2015;26(22):3926-31.
244. Schubert OT, Rost HL, Collins BC, Rosenberger G, Aebersold R. Quantitative proteomics: challenges and opportunities in basic and applied research. Nat Protoc. 2017;12(7):1289-94.
245. Chen LS, Wang J, Wang X, Wang P. A Mixed-Effects Model for Incomplete Data from Labeling-Based Quantitative Proteomics Experiments. Ann Appl Stat. 2017;11(1):114-38.
246. Yoon C, Semitala FC, Atuhumuza E, Katende J, Mwebe S, Asege L, et al. Point-of-care C-reactive protein-based tuberculosis screening for people living with HIV: a diagnostic accuracy study. Lancet Infect Dis. 2017.
247. Niu WY, Wan YG, Li MY, Wu ZX, Zhang LG, Wang JX. The diagnostic value of serum procalcitonin, IL-10 and C-reactive protein in community acquired pneumonia and tuberculosis. Eur Rev Med Pharmacol Sci. 2013;17(24):3329-33.
248. Wilson D, Badri M, Maartens G. Performance of serum C-reactive protein as a screening test for smear-negative tuberculosis in an ambulatory high HIV prevalence population. PLoS One. 2011;6(1):e15248.
249. Chegou NN, Sutherland JS, Malherbe S, Crampin AC, Corstjens PL, Geluk A, et al. Diagnostic performance of a seven-marker serum protein biosignature for the diagnosis of active TB disease in African primary healthcare clinic attendees with signs and symptoms suggestive of TB. Thorax. 2016.
250. Fuzery AK, Levin J, Chan MM, Chan DW. Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges. Clinical Proteomics. 2013;10.
251. Bein K, Di Giuseppe M, Mischler SE, Ortiz LA, Leikauf GD. LPS-treated macrophage cytokines repress surfactant protein-B in lung epithelial cells. Am J Respir Cell Mol Biol. 2013;49(2):306-15.
252. Gargiulo P, Banfi C, Ghilardi S, Magri D, Giovannardi M, Bonomi A, et al. Surfactant-Derived Proteins as Markers of Alveolar Membrane Damage in Heart Failure. Plos One. 2014;9(12).
253. Agostoni P, Banfi C, Brioschi M, Magri D, Sciomer S, Berna G, et al. Surfactant protein B and RAGE increases in the plasma during cardiopulmonary bypass: a pilot study. Eur Respir J. 2011;37(4):841-7.
254. Gopal R, Monin L, Torres D, Slight S, Mehra S, McKenna KC, et al. S100A8/A9 proteins mediate neutrophilic inflammation and lung pathology during tuberculosis. Am J Respir Crit Care Med. 2013;188(9):1137-46.
255. Wang S, Song R, Wang Z, Jing Z, Wang S, Ma J. S100A8/A9 in Inflammation. Front Immunol. 2018;9:1298.
256. Chegou NN, Sutherland JS, Namuganga AR, Corstjens PL, Geluk A, Gebremichael G, et al. Africa-wide evaluation of host biomarkers in QuantiFERON supernatants for the diagnosis of pulmonary tuberculosis. Sci Rep. 2018;8(1):2675.
257. World Health Organization. High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting. Geneva2014.
258. Liu C, Zhao Z, Fan J, Lyon CJ, Wu HJ, Nedelkov D, et al. Quantification of circulating Mycobacterium tuberculosis antigen peptides allows rapid diagnosis of active disease and treatment monitoring. Proc Natl Acad Sci U S A. 2017;114(15):3969-74.
259. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet. 2018;50(5):764.

260. Pareek M, Evans J, Innes J, Smith G, Hingley-Wilson S, Lougheed KE, et al. Ethnicity and mycobacterial lineage as determinants of tuberculosis disease phenotype. Thorax. 2013;68(3):221-9.

261. Pepys MB, Hirschfield GM. C-reactive protein: a critical update. J Clin Invest. 2003;111(12):1805-12.

262. Brown J, Clark K, Smith C, Hopwood J, Lynard O, Toolan M, et al. Variation in C - reactive protein response according to host and mycobacterial characteristics in active tuberculosis. BMC Infect Dis. 2016;16:265.

263. Schroder NW, Schumann RR. Non-LPS targets and actions of LPS binding protein (LBP). J Endotoxin Res. 2005;11(4):237-42.

264. Jacobs R, Maasdorp E, Malherbe S, Loxton AG, Stanley K, van der Spuy G, et al. Diagnostic Potential of Novel Salivary Host Biomarkers as Candidates for the Immunological Diagnosis of Tuberculosis Disease and Monitoring of Tuberculosis Treatment Response. Plos One. 2016;11(8).

265. Jozsi M, Tortajada A, Uzonyi B, Goicoechea de Jorge E, Rodriguez de Cordoba S. Factor H-related proteins determine complement-activating surfaces. Trends Immunol. 2015;36(6):374-84.

266. Gebremicael G, Amare Y, Challa F, Gebreegziabxier A, Medhin G, Wolde M, et al. Lipid Profile in Tuberculosis Patients with and without Human Immunodeficiency Virus Infection. Int J Chronic Dis. 2017;2017:3843291.

267. Deniz O, Gumus S, Yaman H, Ciftci F, Ors F, Cakir E, et al. Serum total cholesterol, HDL-C and LDL-C concentrations significantly correlate with the radiological extent of disease and the degree of smear positivity in patients with pulmonary tuberculosis. Clinical Biochemistry. 2007;40(3-4):162-6.

268. Zhong LJ, Li Y, Tian HF, Guo LJ, Qin SB, Shen J. Data-independent acquisition strategy for the serum proteomics of tuberculosis. International Journal of Clinical and Experimental Pathology. 2017;10(2):1172-85.

269. UNICEF. Tuberculosis now a global threat 2000 [updated March 23. Available from: http://www.unicef.org/newsline/00pr24.htm.

270. Ow SY, Salim M, Noirel J, Evans C, Wright PC. Minimising iTRAQ ratio compression through understanding LC-MS elution dependence and high-resolution HILIC fractionation. Proteomics. 2011;11(11):2341-6.

271. Manousopoulou A, Hayden A, Mellone M, Garay-Baquero DJ, White CH, Noble F, et al. Quantitative proteomic profiling of primary cancer-associated fibroblasts in oesophageal adenocarcinoma. Br J Cancer. 2018;118(9):1200-7.

272. UNITAID. TB Diagnostics Market in Select High-Burden Countries: Current Market and Future Opportunities for Novel Diagnostics. Switzerland; 2015.

273. Johansen FE, Braathen R, Brandtzaeg P. Role of J chain in secretory immunoglobulin formation. Scand J Immunol. 2000;52(3):240-8.

274. Castro CD, Flajnik MF. Putting J chain back on the map: how might its expression define plasma cell development? J Immunol. 2014;193(7):3248-55.

275. Palankar R, Kohler TP, Krauel K, Wesche J, Hammerschmidt S, Greinacher A. Platelets kill bacteria by bridging innate and adaptive immunity via platelet factor 4 and FcgammaRIIA. J Thromb Haemost. 2018;16(6):1187-97.

276. Diao WQ, Shen N, Du YP, Liu BB, Sun XY, Xu M, et al. Fetuin-B (FETUB): a Plasma Biomarker Candidate Related to the Severity of Lung Function in COPD. Sci Rep. 2016;6:30045.

277. Sathyamoorthy T, Tezera LB, Walker NF, Brilha S, Saraiva L, Mauri FA, et al. Membrane Type 1 Matrix Metalloproteinase Regulates Monocyte Migration and Collagen Destruction in Tuberculosis. J Immunol. 2015;195(3):882-91.

278. Cheng L, Han Y, Zhao X, Xu X, Wang J. Identifying pathway modules of tuberculosis in children by analyzing multiple different networks. Exp Ther Med. 2018;15(1):755-60.