JCI insight

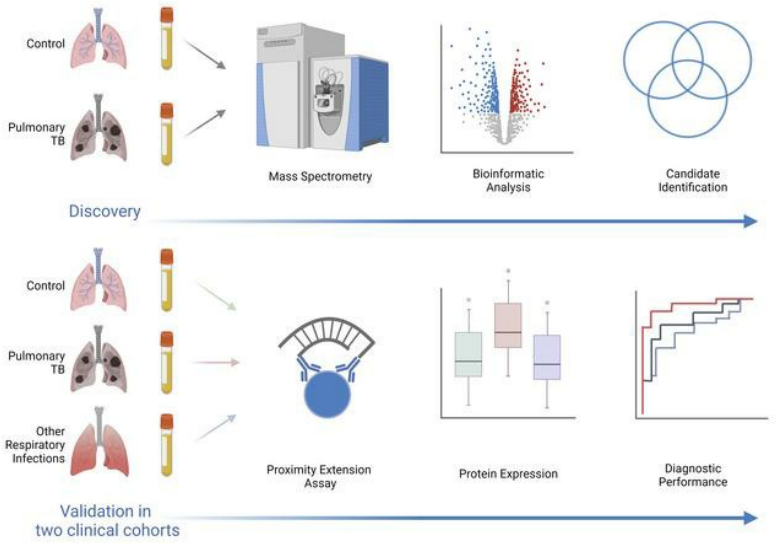
Integrated plasma proteomics identifies tuberculosis-specific diagnostic biomarkers

Hannah F. Schiff, … , Diana Garay-Baquero, Paul Elkington

JCI Insight. 2024. https://doi.org/10.1172/jci.insight.173273.

Clinical Medicine | In-Press Preview | Infectious disease | Pulmonology

Graphical abstract



Integrated plasma proteomics identifies tuberculosis-specific diagnostic biomarkers

Hannah F. Schiff[1,9], Naomi F. Walker[2], Cesar Ugarte-Gil[3,4], Marc Tebruegge[5,6,7], Antigoni Manousopoulou[8], Spiros D. Garbis[1,8], Salah Mansour[1,9], Pak Ho Wong (Michael Wong)[10], Gabrielle Rockett[10], Paolo Piazza[10], Mahesan Niranjan[9,11], Andres F. Vallejo[1], Christopher H. Woelk[12], Robert J. Wilkinson[13,14,15,16], Liku B. Tezera[1,9], Diana Garay-Baquero[1,9,*], Paul Elkington[1,9,*]

[1]NIHR Biomedical Research Centre, Clinical and Experimental Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton, UK. [2]Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK. [3]Instituto de Medicina Tropical Alexander von Humboldt, Universidad Peruana Cayetano Heredia, Lima, Peru. [4]Department of Epidemiology, School of Public and Population Health, University of Texas Medical Branch, Galveston, Texas, USA. [5]Department of Infection, Immunity & Inflammation, Great Ormond Street Institute of Child Health, University College London, London, UK. [6]Department of Paediatrics, Klinik Ottakring, Wiener Gesundheitsverbund, Vienna, Austria. [7]Department of Paediatrics, The University of Melbourne, Parkville, Australia. [8]Proteas Bioanalytics, The Lundquist Institute for Biomedical Innovation, Harbor-UCLA Medical Center, Torrance, CA, USA. [9] Institute for Life Sciences, Southampton, UK. [10]Centre for Human Genetics, University of Oxford, UK. [11] Electronics and Computer Sciences, Faculty of Engineering and Physical Sciences, University of Southampton, Southampton, UK. [12]Verge Genomics, 2 Tower Place, South San Francisco, USA [13]Centre for Infectious Diseases Research in Africa, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory, 7925, Republic of South Africa. [14]Department of Medicine, University of Cape Town, Observatory, 7925, Republic of South Africa. [15]Department of Infectious Diseases, Imperial College London, London, W12 0NN, UK. [16]The Francis Crick Institute, Midland Road, London, NW1 1AT, UK.

* These authors contributed equally

Correspondence:

Paul Elkington

Clinical and Experimental Sciences

University of Southampton

Southampton SO16 1YD, UK

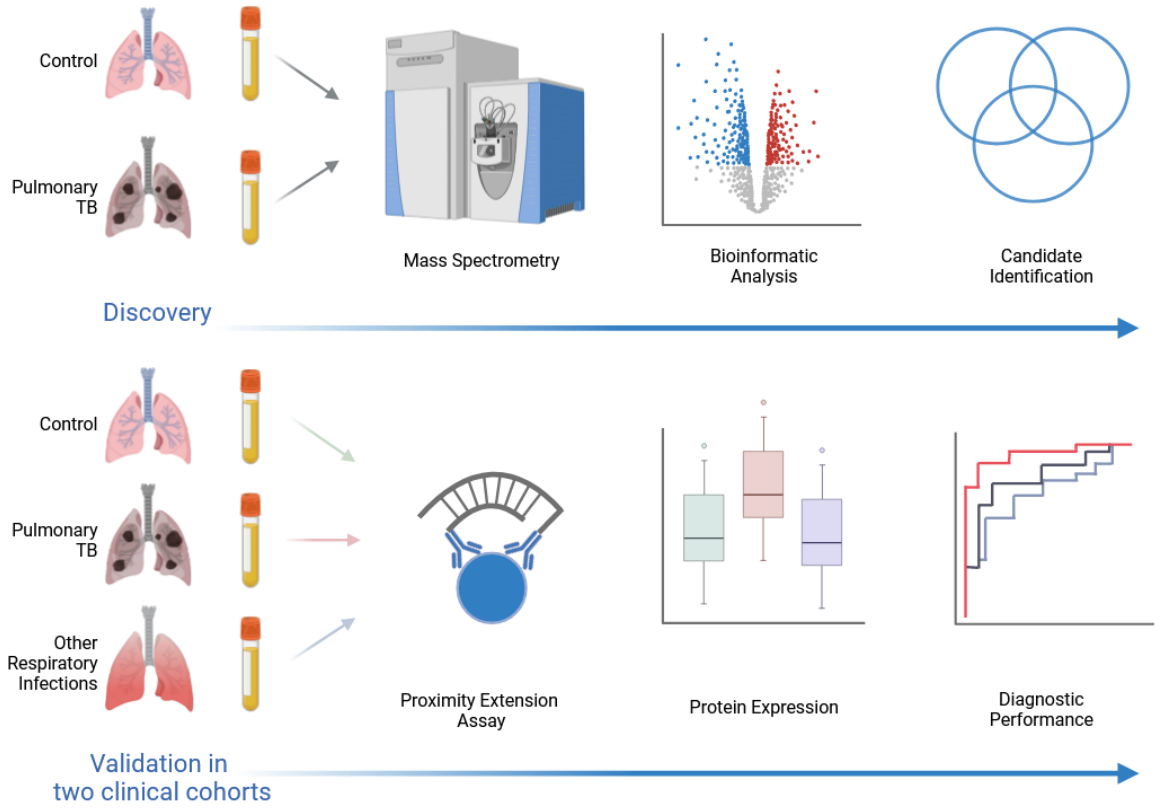
Tel: 00 44 23 8079 6671 E-mail: p.elkington@soton.ac.uk

Abstract

Background Novel biomarkers to identify infectious patients transmitting *Mycobacterium tuberculosis* are urgently needed to control the global tuberculosis (TB) pandemic. We hypothesized that proteins released into the plasma in active pulmonary TB are clinically useful biomarkers to distinguish TB cases from healthy individuals and patients with other respiratory infections. **Methods** We applied a highly sensitive non-depletion tandem mass spectrometry discovery approach to investigate plasma protein expression in pulmonary TB cases compared to healthy controls in South African and Peruvian cohorts. Bioinformatic analysis using linear modelling and network correlation analyses identified 118 differentially expressed proteins, significant through three complementary analytical pipelines. Candidate biomarkers were subsequently analysed in two validation cohorts of differing ethnicity using antibody-based proximity extension assays. **Results** TB-specific host biomarkers were confirmed. A six-protein diagnostic panel, comprising FETUB, FCGR3B, LRG1, SELL, CD14 and ADA2, differentiated patients with pulmonary TB from healthy controls and patients with other respiratory infections with high sensitivity and specificity in both cohorts. **Conclusion** This biomarker panel exceeds the World Health Organisation Target Product Profile specificity criteria for a triage test for TB. The new biomarkers have potential for further development as near-patient TB screening assays, thereby helping to close the case-detection gap that fuels the global pandemic.

199 words

Graphical Abstract



Introduction

Tuberculosis (TB) remains a disease of global significance, causing 1.6 million deaths and 10.6 million cases of active disease in worldwide each year (1). Unfortunately, global control efforts have recently faltered due to the COVID-19 pandemic (2). The World Health Organization (WHO) has identified a global case detection gap of 4 million patients between the estimated incident cases and confirmed diagnoses, with undiagnosed cases predominantly occurring in high TB burden countries. Diagnostic delays in low and middle-income settings are often many months (3), and associate with increased risk of cavitary disease and sputum smear positivity, reflecting high infectiousness (4). Most TB cases result from recently transmitted *Mycobacterium tuberculosis* (Mtb) infection, and therefore the missed diagnoses increase Mtb transmission, TB disease and mortality and fuel the ongoing pandemic (5).

TB control strategies are limited by the currently available diagnostics, which demonstrably are not meeting the needs for global control, requiring specific infrastructure and skilled operators, and do not meet the requirements of the WHO Target Product Profile (TPP) (6). Diagnostic biomarkers capable of identifying people with infectious TB in high burden settings, ideally at the point of care and not requiring sputum expectoration, are urgently needed. A new screening test would not only benefit individuals by enabling prompt and effective treatment but would also be a fundamental tool for potential TB elimination, which remains a key goal for the WHO (7).

Proteins are excellent candidates for diagnostic biomarkers, being stable and utilisable for near-patient diagnostic tests. Several studies have explored potential host plasma protein biomarkers of TB (8-16), and although numerous candidate proteins have been detected, biomarkers or combinatorial biomarker signatures have not yet been found that can reliably differentiate TB from other respiratory diseases, or predict progression (17). Most discovery mass spectrometry-based proteomic studies to date have depleted highly abundant protein components from plasma (10-12). This reduction in plasma protein complexity simplifies the analysis but will also concurrently deplete

proteins of biological interest (18-20). Candidate host proteins identified to date as biomarkers of TB disease are frequently highly sensitive but poorly specific (13-15) .

We hypothesised that analysis of plasma from individuals with pulmonary TB and healthy controls using a non-depletion untargeted proteomics method previously optimised to provide a uniquely high proteome coverage would identify novel markers that achieve both high sensitivity and specificity for TB disease. Here, we report the most detailed plasma proteome of TB to date and perform validation of upregulated proteins by a complementary antibody capture technique in two separate clinical cohorts, including patients with other respiratory infections. We demonstrate the diagnostic potential of an optimised panel incorporating the newly identified biomarkers alongside established analytes that has potential to be developed into a near-patient screening test.

Results

Discovery proteomic analysis of non-depleted plasma

The overall study design is presented in Figure 1. Plasma samples were analysed from 11 untreated male patients with active pulmonary TB and 10 male healthy control samples, from South African and Peruvian cohorts, using a protocol that involved no depletion steps (21). Each plasma sample was initially separated into four segments by size exclusion chromatography, and each segment was processed individually. Analyses of plasma segments were performed in twelve iTRAQ (isobaric tags for relative and absolute quantification) 8-plex experimental sets in a block randomised design comprising three experimental sets. Each iTRAQ experiment contained a bridging master-pool plasma sample run in every experiment. Healthy controls were matched to TB samples by age, ethnicity, and smoking status within each iTRAQ set (Supplemental Tables 1 & 2). Protein abundances from the plasma segments and multi-consensus reports were combined and adjusted for experimental batch effects (Supplemental Figure 1 and 2). Protein abundances from one TB sample failed normalisation leading to exclusion from downstream analysis. An additional TB sample clustered with controls. On review of the clinical data, the patient had minimal chest X-ray infiltration and a normal CRP, and so did not fulfil study inclusion criteria, and was also excluded from downstream analysis. Protein abundances from the remaining combined plasma segment proteomes between experimental sets and the combined multi-consensus proteomes were analysed by complementary bioinformatic approaches to identify candidate diagnostic protein biomarkers (Figure 2). In total, 4,696 protein identifications were made across all iTRAQ experiments, at 5% FDR (false discovery rate). This comprised 2,332 unique host-derived proteins and 22 Mtb-derived proteins (Supplemental Table 3). Of these, 594 host proteins had a quantification result for every sample analysed and therefore comprised the complete quantified proteome. Whilst Mtb proteins were identified across all plasma segments, they were identified in both control and disease samples with low confidence and were not analysed further after review of individual mass spectra.

Plasma proteomes cluster by clinical condition and geographical cohort

Initial exploratory data analysis of the complete quantified proteome by unsupervised hierarchical clustering demonstrated clear separation of the clinical groups (Figure 3A). Furthermore, the South African (label A_) and Peruvian cohorts (label P_) separated within clinical groups. This distinction was most marked within the healthy control plasma samples, with complete segregation depending on geographical location, whereas greater admixture occurred within the TB samples. Similarly, principal component analysis (PCA) confirmed clear separation between clinical groups, manifest by PC1 and comprising 24% of the variation within the dataset (Figure 3B). Again, sample clustering by geographical cohort within clinical groups occurred, manifest through PC2, which contained 16% of the variation within the dataset (Figure 3B).

Complementary bioinformatic analysis identifies 118 differentially expressed proteins in pulmonary TB

High confidence protein identifications, extracted at 1% FDR, were taken forward for differential expression analysis. Protein abundances from individual iTRAQ 8-plex experiments were combined following adjustment for experimental batch (22). FDR-corrected linear modelling (23) identified 195 differentially expressed proteins from analysis of each plasma segment (Supplemental Table 4). A similar *limma* approach analysing the complete multi-consensus proteome yielded 148 differentially expressed proteins (Supplemental Table 5). In parallel, examining the dataset by network correlation methodology, WGCNA (24), demonstrated hierarchical clustering by clinical group, but not experimental set, and by cohort in the healthy controls (Figure 4A). Dendrogram analysis identified a large module of 195 proteins that correlated very closely with disease status (correlation score 0.94, *p* value $2e^{-09}$) (Figure 4B, Supplemental Table 6). Protein module significance scores within the turquoise module closely correlated to protein significance for pulmonary TB (Figure 4C, $p = 6e^{-134}$).

Combined analysis of all three bioinformatic analysis approaches identified one hundred and eighteen proteins that were significant through all statistical approaches (Figure 5A and Supplemental Table 7). Consequently, this group was taken forward as robust candidate diagnostic protein biomarkers. Analysis of protein fold change by *limma* and correlation score by WGCNA demonstrates 56 proteins were significantly upregulated and 62 were significantly downregulated (Figure 5B).

Differentially expressed proteins reflect physiological changes in pulmonary TB

Chord plot analysis was performed to demonstrate key proteins, their magnitude and directionality of fold change relative to key biological processes from gene ontology analysis (Figure 6, Supplemental Table 8). The predominant pathways were consistent with the known biology of M*tb* infection, such as inflammatory response, response to bacterium and regulated exocytosis. However, the most represented process was proteolysis, and proteins regulating extracellular matrix organisation were also frequent. The final processes were negative regulation of cellular metabolic process, lipid metabolic process and platelet degranulation. Key proteins relating to proteolysis included MMP2, TIMP2, FETUB, SERPINA3, SERPINA4, SERPINA5, SERPIND1 and SERPINA10. MMP2 and TIMP2 are also key proteins relating to extracellular matrix organisation, along with the collagen subunit COL15A1, vWF and ADAMTS13. Proteins relating to exocytosis included SELL, CLEC3B and LTA4H. CRP, LBP, S100A8 and S100A9, expectedly linked to the acute inflammatory response. LRG1 and CD14 were key proteins in the response to bacterium. Network plot analysis further confirmed the importance of proteolysis, inflammation and exocytosis-related terms and their relationship to the differentially expressed proteins (Figure 7). Gene ontology analysis of all differentially expressed proteins by cellular compartment showed that proteins were associated with six main locations: endoplasmic reticulum lumen, the extracellular matrix, lipoprotein particles, insulin-like growth factor ternary complexes, secretory vesicles, and platelet granules (Supplemental Table 9). Analysis of enriched molecular function terms indicates significant peptidase and endopeptidase activity, supporting a key role for proteolysis in pulmonary TB (Supplemental Table 10).

Gene ontology analysis of upregulated proteins by cellular component revealed significant enrichment for blood microparticles and fibrinogen complexes (Supplemental Figure 3A) with terms denoting binding to lipid mediators of inflammation and lipopeptides being the dominant molecular functions (Supplemental Figure 3B). Analysis by biological process showed significant enrichment for the acute phase response and acute inflammatory response (Supplemental Figure 3C & 4). The complement and coagulation pathway was the only enriched KEGG pathway by this analysis approach. (Supplemental Figures 3D & 4). Gene ontology analysis of downregulated proteins was strikingly dominated by lipid-related terms across all analyses (Supplemental Figures 5 & 6).

Proteins forming the matrisome, a group of approximately 1000 genes encoding structural and regulatory proteins of the extracellular matrix (25), were over-represented within significantly differentiated proteins. Forty-five of the 118 (38%) divergently regulated proteins were from the matrisome, compared to the matrisome representing 5% of the human proteome (26) reflective of increased ECM turnover in TB (27) (Supplemental Figure 7).

Proximity Extension Analysis validates differential protein expression in the plasma of individuals with pulmonary TB in an independent patient cohort

We performed analysis by an antibody-capture based protein-identification approach in an entirely different cohort, studying serum to validate the potential of the mass spectrometry identified plasma biomarkers for a new diagnostic panel (Figure 8A). Circulating levels of 55 of the 118 (47%) differentially expressed proteins were tested in an independent patient cohort of mixed ethnicity and gender using antibody-based proximity extension assay (Olink™ Explore), using cardiometabolic and inflammatory panels, which gave the largest overlap with the 118 differentially expressed proteins. PEA plates take a maximum of 88 samples, and so to maintain power, 3 groups were analysed: HC, TB and ORI. Serum samples were selected from the UK-based MIMIC cohort (Supplemental Table 11) and included individuals with pulmonary TB (TB, n=32), healthy controls (HC, n=30) without risk factors for TB infection in whom latent TB infection had been ruled-out by a

negative interferon-gamma release assay (IGRA) and patients with symptoms suggestive of TB but with microbiologically confirmed other respiratory infections (ORI, n=26, Supplemental Table 12). Thirty proteins (30/55, 55%) had confirmed differential expression between healthy controls and pulmonary TB, of which 25 were upregulated and 5 downregulated (Supplemental Table 13). Fourteen proteins (14/55, 25%) showed differential expression between pulmonary TB and ORI. Four proteins, FCGR3B, FETUB, GGH and SERPIND1 were present at significantly higher levels in the plasma of pulmonary TB patients than both healthy controls and ORI cases, thereby exhibiting a high degree of specificity for TB (Figure 8B). Significantly reduced circulating levels of RBP4 were demonstrated using Luminex™ methodology, confirming the findings observed by mass spectrometry (Supplemental Figure 8).

A five protein panel differentiates pulmonary TB from healthy controls

Diagnostic performance of individual markers was evaluated using receiver operating characteristic (ROC) curves. ADA2 and CD14 were the best performing individual markers distinguishing TB from HC with an Area under the Curve (AUC) of 0.904 and 0.885 respectively (Figure 9A). Biomarker combinations were then evaluated using CombiROC analysis, to identify panels with a minimum diagnostic sensitivity of 90% and specificity of 70%, thereby meeting WHO Target Product Profile characteristics of a triage test for TB. ROC curves were generated following binary logistic regression of biomarker combinations to classify TB from HC samples. A five-protein panel comprising ADA2, CD14, LRG1, TNFSF13B and vWF gave an AUC of 0.943 (95% CI: 0.889 – 1.000, Figure 9A). Analysis of each analyte individually showed that they were highly significant compared to healthy controls, but were also significantly increased in ORI cases, suggesting they are not TB-specific and are best suited for a rule-out test (Figure 9B). This panel accurately classified patients in 88.7% of cases with a sensitivity of 84.4% (95% CI: 67.3 – 94.3) and specificity of 93.3% (95% CI: 75.8 – 98.8, Figure 9C) at a probability cut off ≥ 0.5.

A six protein panel differentiates pulmonary TB from other respiratory infections

CombiROC analysis of the 14 significantly differentially expressed proteins between TB and ORI was performed to identify the best performing panel (Figure 10A). The combination above the defined threshold comprised FCGR3B, FETUB, GSN, IGFBP3, SELL and CLEC3B (Figure 10B). This combination had an AUC of 0.906 (95% CI: 0.8333 – 0.908), correctly classifying 79.3% of cases with a sensitivity of 81.3% (95% CI: 63.0 – 92.1) and a specificity of 76.9% (95% CI: 56.0 – 90.2, Figure 10C) at a probability cut off ≥ 0.5. Analysis of individual analytes demonstrated that they were significantly different between TB and ORI (Figure 10D), but only FCGR3B and FETUB were also significantly different from healthy controls (Figure 8B).

Integration of top performing analytes into a single panel provides differentiation of TB from both healthy controls and patients with ORI

A universal biomarker panel capable of differentiating individuals with TB from both healthy individuals and individuals with ORI would be more widely applicable to different population testing scenarios. Therefore, biomarker panel combinations were explored using proteins from each of the differentiating panels to identify the best performing universal biomarker panel for both group comparisons. A six-protein marker combination of FCGR3B, FETUB, LRG1, ADA2, CD14 and SELL performed very well for both group comparisons; TB vs. HC with an AUC of 0.972 (95% CI: 0.937 – 1.000), sensitivity 90.6% (95% CI: 73.8 – 97.5) specificity 90.0% (95% CI: 72.3 – 97.4, Figure 11A) and TB vs. ORI with an AUC of 0.930 (95% CI: 0.867 – 0.993), sensitivity 90.6% (95% CI: 66.5 – 96.7), specificity 80.8% (95% CI: 68.2 – 94.5, Figure 11B) at probability cut offs of ≥ 0.5. Performance of this final six protein panel was also evaluated by gender, as the discovery set had been exclusively male. This analysis confirmed the diagnostic performance of markers in male patients, and notably exceeded this in female patients (Supplemental Figure 9).

The six protein panel discriminates TB from healthy controls and patients with ORI in a second independent patient cohort

Antibody-based proximity extension assay was then used to test the diagnostic performance of the final six protein combination in a further independent cohort of plasma samples collected in South Africa ((28), Supplemental Table 14). Samples were selected from HIV-negative individuals with microbiologically confirmed pulmonary TB (TB, n=29), healthy controls (HC, n=30) and individuals presenting with symptoms of pulmonary TB but were negative for Mtb on subsequent microbiological testing (ORI, n= 19) as outlined in Supplemental Table 13. Alternative diagnoses were not microbiologically confirmed in the ORI group due to the resource-limited healthcare setting, but symptoms were consistent with TB. Significantly elevated circulating levels of all six proteins in the panel were confirmed (Figure 12A, Supplemental Table 15). Analysis of the diagnostic performance of the six protein combined panel demonstrated diagnostic specificity for differentiation of TB from both healthy controls (AUC 0.883 (95% CI: 0.796 – 0.968), sensitivity 75.0 (95% CI: 54.8 – 88.6) and specificity 83.3 (64.5 – 93.7, Figure 12B&D) and ORI (AUC 0.876 (95% CI: 0.765- 0.987), sensitivity 92.9 (95% CI: 75.0 – 98.8), specificity 78.9 (95% CI: 53.9 – 93.0, Figure 12C&E) at probability cut offs of ≥ 0.5. Diagnostic performance of the final six protein panel was also tested in both patient cohorts against a combined group of both healthy controls and other respiratory infections, which confirmed preserved specificity of performance (Supplemental Figure 10).

Discussion

TB remains a global catastrophe, and a fundamental issue in controlling the pandemic is the limitations of the diagnostic process, resulting in an estimated 4.2 million missed cases in 2022 (3). This diagnostic gap leads to ongoing transmission, morbidity and mortality, and long-term strain on healthcare systems (6, 29). A novel diagnostic assay with high levels of accuracy would be transformative, permitting population screening to find the missing millions and thereby break the cycle of transmission (30). Indeed, mass screening is being increasingly advocated as a central pillar to control the TB pandemic (3, 7, 31-34). However, this requires new tools that are fit for purpose utilising non-sputum based approaches, but the incomplete understanding of potential plasma biomarkers has considerably limited progress (3).

Here, we utilised a non-depletion quantitative proteomics approach to generate what we believe is the most detailed description of the plasma proteome of TB to date. Complementary bioinformatic analysis using linear modelling and correlation network analysis identified 118 differentially expressed proteins compared to healthy controls. A large subset of biomarkers were successfully validated in a separate clinical cohort by an antibody capture approach, demonstrating analytes can progress to different platforms and overcome this hurdle that may limit translation of proteomics-discovered biomarkers. Four TB-specific biomarkers, FETUB, FCGR3B, GGH and SERPIND1, were raised in TB patients compared to both healthy controls and sick controls with ORI. Combinatorial analysis using a CombiROC approach identified a six-protein biomarker panel that could distinguish active pulmonary TB from healthy controls and patients with ORI achieving the Target Product Profile of the WHO (6). Further validation in a second independent cohort demonstrated statistically significant elevation of all six proteins in the plasma of TB patients and confirmation of high diagnostic performance of the combination panel, distinguishing active pulmonary TB from healthy controls and other respiratory infections. Our discovery proteomic protocol did not involve depletion

steps, in contrast to many previous mass spectrometry-based plasma proteomic studies in TB (10-12, 35, 36). Plasma depletion can co-remove proteins of potential biological interest by non-covalent interactions (18-20). We employed complementary bioinformatic methodologies to identify candidate biomarkers, with *limma* employing Bayesian statistics (23), whilst WGCNA circumvented limitations of multiple comparisons by using unsupervised analysis methods to generate modules of co-expressed proteins that correlate with clinical traits (24). The 118 proteins identified by all three complementary approaches were considered the strongest biomarker candidates.

We identified numerous previously described biomarkers of TB such as C-reactive protein (CRP), lipopolysaccharide-binding protein (LBP), serum amyloid A1 (SAA1), alpha-1-acid glycoprotein 1 (ORM1) and retinol-binding protein 4 (RBP4) alongside S100A8 and S100A9, the protein components of calprotectin. In addition, we identified several biomarkers that we believe have not previously been described, such as lymphocyte cytosolic protein 1 (LCP1), gamma-glutamyl hydrolase (GGH), marginal zone B- and B1-cell-specific protein (MZB1) and fetuin-B (FETUB), including proteins not known to be secreted into the extracellular compartment, such as transcription termination factor 1 (TTF1). LCP1 is a leukocyte specific actin-binding protein that is required for podosome formation and function in macrophages (37). LCP1 has been identified in the phagosomes of BCG-infected macrophages (38). GGH is a protease typically located in lysosomes, and serum GGH has been proposed to be a marker of oxidative stress (39). MZB1 aids peripheral B cell function and promotes secretions of IgM antibodies (40, 41). TTF1 is a multi-functional protein that usually localises to the nucleolus (42) and regulates transcription of surfactant protein B (SFTPB) in type 2 alveolar cells (43, 44). SFTPB is also upregulated in our dataset.

Lung matrix destruction and cavitation is a hallmark of pulmonary TB, which leads to morbidity, mortality, and increased disease transmission (45, 46). Our findings further highlight matrix

305 turnover as a central process in TB. Gene ontology analysis of differentially expressed proteins
306 showed that the extracellular matrix was the most significantly enriched cellular compartment; the
307 most significantly enriched molecular functions were endopeptidase and peptidase inhibitor and
308 regulator activity; and the highest proportion of significantly enriched biological processes related to
309 proteolysis. The SERPINs are a large family of serine protease inhibitors (47) and eight SERPINs were
310 differentially regulated, with elevated SERPIND1 levels shown to have the highest specificity for TB.
311 Fetuin-B (FETUB), a cysteine protease inhibitor, emerged as a key biomarker for pulmonary TB, but
312 little is known about its pathological role. FETUB was part of a 9-protein prognostic risk score in lung
313 adenocarcinoma (48) and plasma levels correlate with worsening lung function in COPD (49),
314 suggesting plasma FETUB levels may relate to destructive lung pathology.

315

316 Pulmonary TB is characterised by excessive inflammation (50), and we identified numerous
317 inflammation-related proteins such as CRP, S100A8 and S100A9. ADA2, CD14 and LRG1, part of the
318 final six-marker panel, have all been implicated in inflammatory responses. ADA2 induces the
319 differentiation of monocytes to macrophages and stimulates macrophage and helper T cell
320 proliferation (51); CD14 serves as a receptor for Mtb cell wall lipoarabinomannan (52, 53); while
321 LRG1 is a marker for neutrophilic granulocyte differentiation, which we have previously shown to be
322 elevated in the serum of patients with pulmonary TB (21). FCGR3A and FCGR3B, low affinity
323 immunoglobulin receptors, were also upregulated. These only differ by one amino acid, with
324 FCGR3A expressed on NK cells and FCGR3B in monocytes and macrophages (54). FCGR3B
325 upregulation was relatively specific for TB, not being upregulated in ORI. Complement components
326 were also upregulated, including C2, C4B, C8B, CFB, C9 and CFHR5, demonstrating broad modulation
327 of this inflammatory pathway in TB disease (55).

328

Amongst the significantly downregulated proteins, lipid-metabolism featured strongly, enriched for the lipoprotein cellular compartment, lipid-binding and lipid inflammatory-mediator binding molecular functions. Lipid metabolism and systemic inflammation are inextricably intertwined (56), with eicosanoid-mediated inflammatory imbalance implicated in human TB (57). Leukotriene A4 hydrolase (LTA4H) is elevated in TB and has been implicated in the spatial organisation of lipid signalling within TB lung granulomas by a proteomics approach (58), and regulates susceptibility to infection (59). Additionally, previous hypothesis-directed approaches have shown lower levels of cholesterol, HDL-C and LDL-C levels in pulmonary TB patients compared to controls (60).

Differences in TB pathogenesis between ethnic groups has been recognised for over a century (61, 62), and ethnicity has been shown to be a powerful determinant of clinical TB phenotype, independent of Mtb strain lineage (63). We analysed plasma samples from two geographical origins, South Africa and Peru, and identified differences in the plasma proteome by region both in healthy controls and in TB patients. Such geographical differences need consideration in developing new diagnostic tests (64). Reassuringly our top candidate biomarkers were validated in an independent cohort of mixed ethnicity and gender, and the six protein biomarker panel in a further independent clinical cohort of mixed gender.

Previous studies have explored circulating biomarkers of TB disease utilising diverse approaches. Luminex-based analysis of HIV-negative individuals from sub-Saharan African countries for pre-specified analytes has identified a two-protein panel and a nine-protein panel, both including CRP, that distinguish TB from other respiratory diseases, with comparatively high sensitivity, but lower specificity (14, 15). A Simoa ultrasensitive immunoassay comprising four host proteins and an antibody against TB antigen Ag85B was also able to discriminate between patients with TB and those with other respiratory diseases, but had lower performance characteristics than our biomarker

panel, and, importantly, requires a specific reader (65) . In another study, analysis by aptamer-based SOMAscan assays identified a six-protein panel comprising SYWC (cytoplasmic tryptophan-tRNA ligase), kallistatin, C9, gelsolin, testican-2 and aldolase C (16), which could distinguish TB from non-TB samples with a similar sensitivity and specificity to our panel, though limited data were available regarding the patients that made up the non-TB group. Our unbiased discovery approach using geographically diverse populations demonstrates a robust method for the identification of protein biomarkers with higher specificity for differentiating TB disease in carefully phenotyped comparator groups of healthy controls and other respiratory infections. Evidently, the performance of our proposed biomarkers will require validation in additional cohorts, including patients with extrapulmonary TB and individuals with HIV co-infection, which present additional diagnostic challenges (66). An assay will be needed that meets the WHO ASSURED criteria for a point-of-care test for use in resource-limited settings, being affordable, sensitive, specific, user-friendly, rapid, equipment-free and deliverable to those in need (67). Recent developments in integrated microfluidic systems may allow the translation of diagnostic panels onto an immuno-assay-based lab-on-a-chip system, that would have potential for near-patient use (6).

In summary, our integrated proteomics approach has identified TB-specific circulating biomarkers of disease amongst a group of 118 divergently regulated proteins identified through a rigorous bioinformatic pipeline. A six-protein biomarker panel can discriminate individuals with active pulmonary TB from healthy individuals and from those with other bacterial or viral pulmonary infections, with potential for onward development into a point-of-care test suitable for mass population screening. The diagnostic potential of these new protein biomarkers and panels require further validation in key clinical groups, such as HIV co-infected individuals and in cohorts with high co-prevalence of common comorbidities such as diabetes and chronic obstructive pulmonary disease. Additionally, although our study focussed on separating infection from TB, in future comparison with sarcoidosis, autoimmune pneumonias or chronic fungal pneumonias in specific

settings where these are prevalent will also be warranted. Whilst future validation in different cohorts and development of a near-patient assay represent significant future hurdles, we propose that these findings provide critical knowledge to develop an initial screening assay that can be used to triage patients to pathways involving more expensive confirmatory testing for TB (7, 68). Such active case finding will help to close the case-detection gap that is fuelling the ongoing TB pandemic.

Methods

Study participants

Participants in the discovery experiment were recruited in two separate cohorts. The South African cohort were recruited at Ubuntu TB/HIV clinic in Cape Town from June 2012 to February 2014 and were of Black African ethnicity (28). Written informed consent was provided. The diagnosis of active TB was based on sputum smear or culture positivity, GeneXpert results where available and chest radiograph findings. For healthy controls sputum samples were smear and culture negative for acid-fast bacilli. The Peruvian cohort was recruited at clinics in Lima, Peru during 2015. The diagnosis of TB was based on TB symptoms, sputum smear and culture positivity, and chest radiograph findings. Healthy control individuals were QuantiFERON negative, excluding coincidental LTBI. Plasma samples from male HIV-negative participants were randomly selected for the discovery experiment from either cohort if they were between the ages of 18 and 50 years old and had a BMI between 16 and 26 and there was sufficient sample for analysis. Exclusion criteria included anaemia (Hb ≤ 8 g/dL), significant renal impairment (creatinine ≥ 150µm/L), significant hepatic disease (ALT ≥ 80 IU/L), known malignancy or diabetes mellitus. Patients with active TB had not yet commenced treatment at the time of plasma sampling.

Participants in validation cohort 1 were from the UK-based MIMIC cohort of mixed ethnicity. Patients were recruited between June 2014 and February 2017. All participants in the MIMIC study were UK resident at the time of sample collection and were HIV-negative. Healthy control individuals were asymptomatic, without a history of previous TB disease, TB contact or travel to a high TB prevalence area, and no evidence of LTBI in interferon-gamma release assay (IGRA) testing. Active pulmonary TB cases were symptomatic individuals with microbiologically confirmed TB by either sputum smear, sputum culture or positive PCR for *Mtb*. Individuals with other respiratory infections (ORI) were symptomatic with microbiologically confirmed respiratory tract infection caused by a pathogen other than *Mtb*, without a history of previous active TB. The causative agents in this group

comprised influenza virus A and B, respiratory syncytial virus, human metapneumovirus, *Streptococcus pneumoniae*, *Staphylococcus aureus*, and *Mycoplasma pneumoniae.* All participants in validation cohort 2 were resident in Khayelitsha, Cape Town at the time of sample collection, were of Black African ethnicity and HIV-uninfected. The diagnosis of TB was based on TB symptoms, sputum smear and culture positivity, and chest radiograph findings.

Sex as a biological variable

Sex has been carefully considered as a biological variable in this investigation. For the discovery plasma mass spectrometry only samples from male patients were used as males exhibit the most florid pulmonary TB pathology. For both validation cohorts samples from males and females were tested, and ratios are presented in Supplemental Table 11 & 14.

Sample processing

For the discovery experiment, venous blood was collected in sodium heparin vacutainer tubes and plasma prepared according to standard operating procedures at the site of recruitment and stored at -80^{o}C. Aliquots of 120µL of plasma were liquid fixed with 380µL of 7 M guanidine hydrochloride and 10% methanol and stored at -20^{o}C until size exclusion chromatography. Aliquots of 20µL of each plasma sample in the discovery experiment was combined to generate a master-pool sample to help mitigate batch effects across different proteomic experiments.

For the validation experiment in the MIMIC cohort, venous blood was collected in serum vacutainer tubes and serum prepared according to standard operating procedures at the site of recruitment and stored in 100µL aliquots at -80^{o}C. For Proximity extension analysis (PEA) serum samples were thawed, centrifuged for 10 minutes at 1500 rpm, and 40µL per sample aliquoted into a 96 well plate and re-frozen at -80^{o}C until analysis at the Oxford Genomics Centre.

Discovery proteomic analysis

HP-SEC & protein digestion

The methodology for high performance size-exclusion chromatography has been previously described (21). Total protein lyophilised extracts from each plasma segment were reconstituted with 0.5 M triethylammonium bicarbonate and 0.05% sodium dodecyl sulphate and sonicated on ice. Following centrifugation at 16,000G for 10 minutes at 4°C protein content was estimated using a Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific) at 280nm. Volume-adjusted 120µg of protein was reduced with 2 µL of 50 mM Tris-2-carboxymethyl phosphine and incubated at 60°C for 1 hour. Samples were then alkylated using 1 µL of 200 mM methylmethane thiosulphonate and incubated for 10 minutes at room temperature. Protein digestion was conducted to a ratio of 1:40 enzyme/substrate with trypsin MS grade (Pierce, Thermo Fisher Scientific) overnight at 37°C in the dark.

iTRAQ-labelling

Isopropanol was added to iTRAQ labels to ensure more than 60% organic phase during sample labelling and each tag was added to the appropriate trypsinised sample. The masterpool was labelled using tag 113, and the samples were block randomised to the remaining tags according to Supplemental Table 2. The labelling reaction was conducted for 2 hours at room temperature and the reaction stopped with 8 µL of 5% ammonium hydroxylamine. Samples were lyophilised and stored at -20°C until chromatographic separation.

Peptide fractionation

Offline peptide fractionation was performed at high pH (0.08% NH_4OH) using a C_4 column (Kromasil, 3.5 µm, 2.1 mm x 150 mm) on a Shimazdu HPLC system. iTRAQ-labelled peptides were reconstituted and pooled with 100 µL of mobile phase and centrifuged at 16,000G at room temperature for 10 minutes. Supernatant was injected and separated at a flow rate of 0.3 mL/min at 30°C. Fractions were collected by peak detected at 215 nm. Peptide fractions were dried using a speedvac

concentrator (Thermo Fisher Scientific) and stored at -20°C until LC-MS/MS analysis. Highly hydrophilic and hydrophobic fractions from the extreme regions of the chromatographic traces were pooled and further cleaned using Gracepure SPE C18-AQ 100 mg/1 mL cartridges (Thermo Fisher Scientific).

Mass spectrometry analysis

Peptide fractions were analysed using a Dionex Ultimate UHPLC system coupled to a nano-ESI-LTQ-Velos Pro Orbitrap Elite mass spectrometer (Thermo Fisher Scientific). Online chromatographic separation of each peptide fraction was conducted using a AcclaimPepMap RSLC C18 nanoViper column (Thermo Fisher Scientific 2 µm, 75 µm × 25 cm). This was retrofitted to a PicTip emitter (FS360-20-10-D-20-C7) for injection into the mass spectrometer. MS characterization of eluting peptides was conducted between 380 and 1500 m/z. The top 10 +2 and +3 precursor ions were further characterized by tandem MS (MS/MS). Higher energy collisional dissociation (HCD) and collision-induced dissociation (CID) fragmentation for each of the collected fractions was performed.

Full MS scans and MS/MS scans were acquired at a resolution of 30,000 FWHM (full width at half maximum) for Set C segments 1-3, and 60,000 FWHM for all further plasma segments. Data were acquired using Xcalibur software (Thermo Fisher Scientific). Conditions for ionization, CID and HCD fragmentation, and ion detection for this method have been previously reported (69).

MS data processing

Target decoy searching of raw mass spectra was conducted with Proteome Discoverer v2.4 (Thermo Fisher Scientific). SequestHT was used for target decoy search for tryptic peptides, allowing 2 missed cleavages, 10 ppm mass tolerance and a minimum peptide length of 6 amino acids. Dynamic modifications of oxidation (M), deamidation (N, Q) and phosphorylation (S, T, Y) and static modifications of iTRAQ 8plex (N-terminus, K) and meythylthio (C) were permitted. Fragment ion mass tolerance was 0.02 Da for HCD-generated spectra and 0.5 Da for CID-generated spectra. Percolator was set to a concatenated strategy for target decoy selection with a strict FDR target of

0.01 and relaxed FDR target of 0.05. Spectra were searched against a concatenated FASTA file comprising the UniProtKB SwissProt human proteome and the reference *M. tuberculosis* H37Rv proteome (SwissProt and TrEMBL). Unique peptide spectrum matches were taken through to consensus workflow allowing a 50% co-isolation threshold and a signal-to-noise ratio of 3. Normalization was to total peptide amount and scaling was to controls average. This scaling enabled a multi-consensus workflow to generate grouped protein abundances across all four plasma segments for each experimental set. Protein abundances were imported to R for log2 transformation, median normalisation, data visualisation and bioinformatic analysis. Data from plasma samples from TB patients labelled with iTRAQ tags 118 and 121 in experimental set C were excluded from further analysis at this stage due to failure of normalisation (tag 118) and clustering with the control group (121). Clinically the latter patient had microbiologically confirmed pulmonary TB, but minimal CXR changes and a normal CRP.

Validation proteomic analysis

Serum samples from the MIMIC cohort were thawed and centrifuged at 15,000g for 10mins at 4°C. Serum was aliquoted onto 96 well PCR plates and transported on dry ice to the Oxford Genomics Centre for analysis. Proximity Extension Assay (PEA) was performed as per the proteomic method that has been previously described (70) using Olink® Explore Cardiometabolic and Inflammation II panels. Each assay has been extensively validated for limit of detection, measurement ranges, precision, reproducibility and specificity as detailed at https://olink.com/our-platform/assay-validation/#explore.

Statistics

Discovery proteomics

Differentially expressed proteins were identified using linear modelling with the R package limma (23) including FDR correction for multiple comparisons and network correlation analysis using the R package WGCNA (24). Limma was applied on combined data from each plasma segment and on

multi-consensus analyses, following adjustment for experimental batch effects using the R package ComBat (22). WGCNA was applied to ComBat-adjusted data for combined multi-consensus analyses. WGCNA was used to determine clusters of highly correlated proteins (colour modules) and explore their correlation with phenotypic traits. Module significance was expressed as a correlation score with statistical significance. Gene ontology enrichment analysis was conducted using ShinyGO (71) with all proteins identified from the discovery experiment as a background proteome. Only gene ontology terms with an FDR-adjusted p-value less than 0.05 were considered. Graphical visualisations of the enrichment analysis were generated using the R package clusterProfiler (72) for cnet plots and GOplots for chord plots.

Validation proteomics

Differences in protein expression between groups for PEA measurements were analysed using GraphPad Prism v9. Data distributions were examined for normality and differences analysed by one-way ANOVA (analysis of variance) if Gaussian distribution was found. For non-parametrically distributed data differences between groups were analysed using Kruskal-Wallis method with Dunn's test for multiple comparisons. A p-value of less than 0.05 was considered statistically significant. Combinatorial performance of biomarkers was assessed using the R package CombiROC (73). Receiver operating characteristics curves for clinical group classification were then explored for the best performing biomarker panels following binary logistic regression using SPSS v28.0.1.0 (IBM statistics).

Study approval

All clinical studies were conducted according to the Declaration of Helsinki principles. All participants gave written informed consent. The South African cohort was recruited under University of Cape Town Research Ethics Committee approval (HREC, REF 516/2011). Enrolment of participants in the Peruvian study was approved by the Universidad Peruana Cayetano Heredia Institutional Review Board (SIDISI 65314). University of Southampton Ethics and Research Governance approval was

given for transporting samples to the United Kingdom for analysis (approval 17758). The MIMIC study was approved by the National Research Ethics Service Committee South Central (Ref 13 SC 0043).

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the Proteomics IDEntification Database partner repository (74). Selected PEA data is available in Supplemental Tables 13 & 15. Values for all data points shown in graphs are reported in the Supporting Data Values file. Further data and analysis code are available from the corresponding author on request.

Author Contributions

HS performed 6 of the 12 discovery proteomic experiments, analysed and integrated the data from all 12 discovery data sets, performed and optimised the subsequent bioinformatic analysis, managed the MIMIC sample cohort, directed the validation proteomics analysis for both cohorts, performed statistical analysis of the validation datasets, and drafted the manuscript. DJGB and PTE designed the discovery experiment. DJGB performed high performance SEC of all discovery plasma samples, 6 of the 12 discovery proteomic experiments, optimised the wet lab proteomic method and provided R scripts for protein abundance normalisation, limma, ComBat, and principal component analysis. NFW and RJW recruited the South African cohorts and provided clinical annotation. CUG recruited the Peruvian cohort and provided clinical annotation. MT designed the MIMIC study, recruited the MIMIC cohort (Southampton site), and provided clinical annotation. DJGB, AM and SDG provided expertise in the plasma proteomic protocol. AFV, CHW and MN provided expert insight into bioinformatic analysis and AFV provided the R scripts for WGCNA. LBT and SM provided expert insight into wet lab methodology and useful discussions throughout the project. PHW, GR and PP performed the PEA analysis. HS & PE secured funding for the project. PE was involved in the study

556 design, provided oversight to the project, and contributed to the manuscript writing and editing. All
557 authors reviewed the manuscript, provided intellectual input, and approved the final version.
558

Acknowledgements

HS was funded by a Clinical Research Training Fellowship from the Medical Research Council, UK (MR/R001065/1). This award is jointly funded by the UK Medical Research Council (MRC) and the UK Foreign Commonwealth and Development Office (FCDO) under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. HS was also supported by a Clinical Lectureship from the UK National Institute for Health Research (NIHR). NFW was supported by Wellcome Trust (094000), NIHR, Starter Grant for Clinical Lecturers (Academy of Medical Sciences UK), and British Infection Association. CUG received support from the Program for Advanced Research Capacities for AIDS in Peru at Universidad Peruana Cayetano Heredia (D43TW00976301) from the Fogarty International Center at the US NIH. We are grateful to the Centre for Infectious Diseases Research in Africa clinical research team and to the participants, staff, and patients of Ubuntu Clinic and the Western Cape Government Department of Health. The MIMIC study and MT were supported by a grant from the UK Technology Strategy Board / Innovate UK (grant 101556). MT was also supported by a Clinical Lectureship from the UK National Institute for Health Research (NIHR). SM and DJGB were supported by the MRC (MR/S024220/1). PE was supported by the MRC (MR/P023754/1 and MR/W025728/1). RJW is supported by Wellcome (203135), by the Francis Crick Institute which receives funding from UKRI-MRC (CC2112), Cancer research UK (CC2112) and Wellcome (CC2112). RJW is supported in part by the NIHR Biomedical Research Centre of Imperial College NHS. For the purposes of open access, the authors have applied a CC-BY public copyright to any author-accepted manuscript arising from this submission. We thank the Oxford Genomics Centre and Olink for the validation proteomics experimental data. We acknowledge the support of the Southampton NIHR Biomedical Research Centre. We thank Regina Teo and Rebecca Fulton for excellent technical support. Graphical abstract was created using BioRender.com.

Conflict-of-Interest Statement

HS, DJGB and PE are cited as co-inventors on a patent “Biomarker and Uses thereof” which lists some of the markers identified within this manuscript as potential new diagnostic markers for tuberculosis (UK 2306925.5).

References

1. WHO. Global Tuberculosis Report. 2022.
2. Pai M, Kasaeva T, and Swaminathan S. Covid-19's Devastating Effect on Tuberculosis Care - A Path to Recovery. *The New England journal of medicine.* 2022;386(16):1490-3.
3. Pai M, Dewan PK, and Swaminathan S. Transforming tuberculosis diagnosis. *Nat Microbiol.* 2023;8(5):756-9.
4. Cheng S, Chen W, Yang Y, Chu P, Liu X, Zhao M, et al. Effect of Diagnostic and Treatment Delay on the Risk of Tuberculosis Transmission in Shenzhen, China: An Observational Cohort Study, 1993-2010. *Plos One.* 2013;8(6):e67516.
5. Dale KD, Karmakar M, Snow KJ, Menzies D, Trauer JM, and Denholm JT. Quantifying the rates of late reactivation tuberculosis: a systematic review. *Lancet Infectious Diseases.* 2021;21(10):E303-E17.
6. Hong JM, Lee H, Menon NV, Lim CT, Lee LP, and Ong CWM. Point-of-care diagnostic tests for tuberculosis disease. *Sci Transl Med.* 2022;14(639):eabj4124.
7. Marks GB, Horsburgh CR, Jr., Fox GJ, and Nguyen TA. Epidemiological approach to ending tuberculosis in high-burden countries. *Lancet.* 2022.
8. 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.
9. Liu JY, Jiang TT, Wei LL, Yang XY, Wang C, Zhang X, et al. The discovery and identification of a candidate proteomic biomarker of active tuberculosis. *Bmc Infectious Diseases.* 2013;13.
10. Xu D, Li Y, Li X, Wei LL, Pan Z, Jiang TT, et al. Serum protein S100A9, SOD3, and MMP9 as new diagnostic biomarkers for pulmonary tuberculosis by iTRAQ-coupled two-dimensional LC-MS/MS. *Proteomics.* 2015;15(1):58-67.
11. 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.
12. Chen C, Yan T, Liu L, Wang J, and 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.
13. 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;71(9):785-94.
14. Morris TC, Hoggart CJ, Chegou NN, Kidd M, Oni T, Goliath R, et al. Evaluation of Host Serum Protein Biomarkers of Tuberculosis in sub-Saharan Africa. *Front Immunol.* 2021;12:639174.
15. Mutavhatsindi H, van der Spuy GD, Malherbe ST, Sutherland JS, Geluk A, Mayanja-Kizza H, et al. Validation and Optimization of Host Immunological Bio-Signatures for a Point-of-Care Test for TB Disease. *Front Immunol.* 2021;12:607827.
16. De Groote MA, Sterling DG, Hraha T, Russell T, Green LS, Wall K, et al. Discovery and Validation of a Six-Marker Serum Protein Signature for the Diagnosis of Active Pulmonary Tuberculosis. *J Clin Microbiol.* 2017.

17. Penn-Nicholson A, Hraha T, Thompson EG, Sterling D, Mbandi SK, Wall KM, et al. Discovery and validation of a prognostic proteomic signature for tuberculosis progression: A prospective cohort study. *Plos Med.* 2019;16(4):e1002781.

18. Asanov K. Multi-Site Assessment of ProteoRed Plasma Reference Sample for Benchmarking LC-MS Platform Performance. *J Biomol Tech.* 2011.

19. 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.

20. Hakimi A, Auluck J, Jones GD, Ng LL, and 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.

21. Garay-Baquero DJ, White CH, Walker NF, Tebruegge M, Schiff HF, Ugarte-Gil C, et al. Comprehensive plasma proteomic profiling reveals biomarkers for active tuberculosis. *Jci Insight.* 2020.

22. Johnson WE, Li C, and Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8(1):118-27.

23. 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.

24. Langfelder P, and Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008;9:559.

25. Naba A, Clauser KR, Ding H, Whittaker CA, Carr SA, and Hynes RO. The extracellular matrix: Tools and insights for the "omics" era. *Matrix Biol.* 2016;49:10-24.

26. Ponomarenko EA, Poverennaya EV, Ilgisonis EV, Pyatnitskiy MA, Kopylov AT, Zgoda VG, et al. The Size of the Human Proteome: The Width and Depth. *Int J Anal Chem.* 2016;2016:7436849.

27. Elkington P, Polak ME, Reichmann MT, and Leslie A. Understanding the tuberculosis granuloma: the matrix revolutions. *Trends Mol Med.* 2022;28(2):143-54.

28. 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.

29. Ivanova O, Hoffmann VS, Lange C, Hoelscher M, and Rachow A. Post-tuberculosis lung impairment: systematic review and meta-analysis of spirometry data from 14 621 people. *Eur Respir Rev.* 2023;32(168).

30. Uplekar M, Weil D, Lonnroth K, Jaramillo E, Lienhardt C, Dias HM, et al. WHO's new end TB strategy. *Lancet.* 2015;385(9979):1799-801.

31. Esmail A, Randall P, Oelofse S, Tomasicchio M, Pooran A, Meldau R, et al. Comparison of two diagnostic intervention packages for community-based active case finding for tuberculosis: an open-label randomized controlled trial. *Nat Med.* 2023;29(4):1009-16.

32. Burke RM, Nliwasa M, Dodd PJ, Feasey HR, Khundi M, Choko A, et al. Impact of community-wide tuberculosis active case finding and HIV testing on tuberculosis trends in Malawi. *Clin Infect Dis.* 2023.

33. Calligaro GL, Zijenah LS, Peter JG, Theron G, Buser V, McNerney R, et al. Effect of new tuberculosis diagnostic technologies on community-based intensified case finding: a multicentre randomised controlled trial. *Lancet Infect Dis.* 2017;17(4):441-50.

34. Corbett EL, Bandason T, Duong T, Dauya E, Makamure B, Churchyard GJ, et al. Comparison of two active case-finding strategies for community-based diagnosis of symptomatic smear-positive tuberculosis and control of infectious tuberculosis in Harare, Zimbabwe (DETECTB): a cluster-randomised trial. *Lancet.* 2010;376(9748):1244-53.

35. 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.

36. Wang C, Wei LL, Shi LY, Pan ZF, Yu XM, Li TY, et al. Screening and identification of five serum proteins as novel potential biomarkers for cured pulmonary tuberculosis. *Sci Rep.* 2015;5:15615.

37. De Clercq S, Boucherie C, Vandekerckhove J, Gettemans J, and Guillabert A. L-Plastin Nanobodies Perturb Matrix Degradation, Podosome Formation, Stability and Lifetime in THP-1 Macrophages. *Plos One.* 2013;8(11).

38. Lee BY, Jethwaney D, Schilling B, Clemens DL, Gibson BW, and Horwitz MA. The Mycobacterium bovis Bacille Calmette-Guerin Phagosome Proteome. *Mol Cell Proteomics.* 2010;9(1):32-53.

39. Lee DH, Blomhoff R, and Jacobs DR. Is serum gamma glutamyltransferase a marker of oxidative stress? *Free Radical Res.* 2004;38(6):535-9.

40. Flach H, Rosenbaum M, Duchniewicz M, Kim S, Zhang SL, Cahalan MD, et al. Mzb1 protein regulates calcium homeostasis, antibody secretion, and integrin activation in innate-like B cells. *Immunity.* 2010;33(5):723-35.

41. Rosenbaum M, Andreani V, Kapoor T, Herp S, Flach H, Duchniewicz M, et al. MZB1 is a GRP94 cochaperone that enables proper immunoglobulin heavy chain biosynthesis upon ER stress. *Gene Dev.* 2014;28(11):1165-78.

42. Platt T. Transcription termination and the regulation of gene expression. *Annu Rev Biochem.* 1986;55:339-72.

43. Yang MC, Guo Y, Liu CC, Weissler JC, and Yang YS. The TTF-1/TAP26 complex differentially modulates surfactant protein-B (SP-B) and -C (SP-C) promoters in lung cells. *Biochem Biophys Res Commun.* 2006;344(2):484-90.

44. Whitsett JA, and Glasser SW. Regulation of surfactant protein gene transcription. *Biochim Biophys Acta.* 1998;1408(2-3):303-11.

45. Elkington PT, D'Armiento JM, and Friedland JS. Tuberculosis immunopathology: the neglected role of extracellular matrix destruction. *Sci Transl Med.* 2011;3(71):71ps6.

46. Urbanowski ME, Ordonez AA, Ruiz-Bedoya CA, Jain SK, and Bishai WR. Cavitary tuberculosis: the gateway of disease transmission. *Lancet Infect Dis.* 2020;20(6):e117-e28.

47. Law RH, Zhang Q, McGowan S, Buckle AM, Silverman GA, Wong W, et al. An overview of the serpin superfamily. *Genome Biol.* 2006;7(5):216.

48. Chen SJ, Zhang J, Li Q, Xiao LY, Feng X, Niu Q, et al. A Novel Secreted Protein-Related Gene Signature Predicts Overall Survival and Is Associated With Tumor Immunity in Patients With Lung Adenocarcinoma. *Frontiers in Oncology.* 2022;12.

49. 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.

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

51. Zavialov AV, Gracia E, Glaichenhaus N, Franco R, Zavialov AV, and Lauvau G. Human adenosine deaminase 2 induces differentiation of monocytes into macrophages and stimulates proliferation of T helper cells and macrophages. *J Leukocyte Biol.* 2010;88(2):279-90.

52. Yu WM, Soprana E, Cosentino G, Volta M, Lichenstein HS, Viale G, et al. Soluble CD14(1-152) confers responsiveness to both lipoarabinomannan and lipopolysaccharide in a novel HL-60 cell bioassay. *Journal of Immunology.* 1998;161(8):4244-51.

53. Pugin J, Heumann D, Tomasz A, Kravchenko VV, Akamatsu Y, Nishijima M, et al. Cd14 Is a Pattern-Recognition Receptor. *Immunity.* 1994;1(6):509-16.

54. Ravetch J.V. PB. Alternative membrane forms of Fc gamma RIII(CD16) on human natural killer cells and neutrophils. Cell type-specific expression of two genes that differ in single nucleotide substitutions. *J Exp Med.* 1989;170(2):487-97.

55. Reis ES, Mastellos DC, Hajishengallis G, and Lambris JD. New insights into the immune functions of complement. *Nature Reviews Immunology.* 2019;19(8):503-16.

56. van Diepen JA, Berbee JFP, Havekes LM, and Rensen PCN. Interactions between inflammation and lipid metabolism: Relevance for efficacy of anti-inflammatory drugs in the treatment of atherosclerosis. *Atherosclerosis.* 2013;228(2):306-15.

57. Tobin DM, and Ramakrishnan L. TB: the Yin and Yang of lipid mediators. *Current Opinion in Pharmacology.* 2013;13(4):641-5.

58. Marakalala MJ, Raju RM, Sharma K, Zhang YJ, Eugenin EA, Prideaux B, et al. Inflammatory signaling in human tuberculosis granulomas is spatially organized. *Nat Med.* 2016;22(5):531-8.

59. Tobin DM, Vary JC, Jr., Ray JP, Walsh GS, Dunstan SJ, Bang ND, et al. The lta4h locus modulates susceptibility to mycobacterial infection in zebrafish and humans. *Cell.* 2010;140(5):717-30.

60. 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.

61. Stead WW, Senner JW, Reddick WT, and Lofgren JP. Racial differences in susceptibility to infection by Mycobacterium tuberculosis. *The New England journal of medicine.* 1990;322(7):422-7.

62. M. T. Tuberculosis among American Negroes: Medical Research on a Racial Disease, 1830–1950. *Journal of the History of Medicine and Allied Sciences.* 1977;XXXII(3):252-79.

63. 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.

64. Guo J, Zhang X, Chen X, and Cai Y. Proteomics in Biomarker Discovery for Tuberculosis: Current Status and Future Perspectives. *Front Microbiol.* 2022;13:845229.

65. Ahmad R, Xie L, Pyle M, Suarez MF, Broger T, Steinberg D, et al. A rapid triage test for active pulmonary tuberculosis in adult patients with persistent cough. *Sci Transl Med.* 2019;11(515).

66. Nogueira BMF, Krishnan S, Barreto-Duarte B, Araujo-Pereira M, Queiroz ATL, Ellner JJ, et al. Diagnostic biomarkers for active tuberculosis: progress and challenges. *Embo Mol Med.* 2022:e14088.

67. Mabey D, Peeling RW, Ustianowski A, and Perkins MD. Diagnostics for the developing world. *Nature reviews Microbiology.* 2004;2(3):231-40.

68. Yoon C, Dowdy DW, Esmail H, MacPherson P, and Schumacher SG. Screening for tuberculosis: time to move beyond symptoms. *Lancet Respir Med.* 2019;7(3):202-4.

69. 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. *Journal of proteome research.* 2014;13(11):5094-105.

70. Wik L, Nordberg N, Broberg J, Bjorkesten J, Assarsson E, Henriksson S, et al. Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. *Mol Cell Proteomics.* 2021;20:100168.

71. Ge SX, Jung DM, and Yao RA. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics.* 2020;36(8):2628-9.

72. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation.* 2021;2(3):100141.

73. Bombaci M, and Rossi RL. Computation and Selection of Optimal Biomarker Combinations by Integrative ROC Analysis Using CombiROC. *Methods Mol Biol.* 2019;1959:247-59.

74. Perez-Riverol Y, Bai J, Bandla C, Garcia-Seisdedos D, Hewapathirana S, Kamatchinathan S, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* 2022;50(D1):D543-D52.

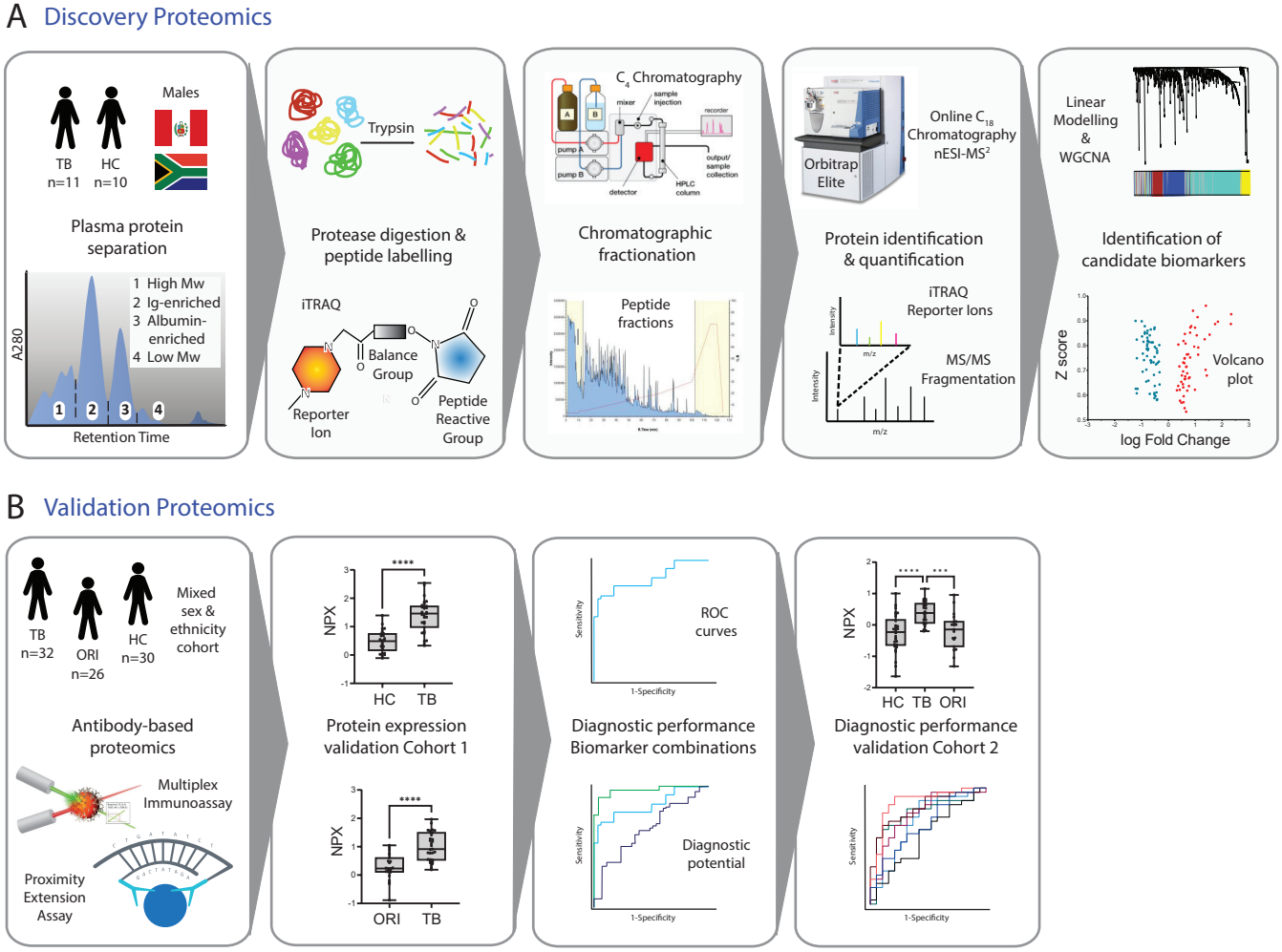


Figure 1: Integrated proteomic study design for TB biomarker identification and validation.
(A) Discovery stage comprising sequential orthogonal fractionation of non-depleted plasma at both protein and peptide level, iTRAQ peptide labelling and tandem mass spectrometry for protein identification and relative quantification. Complementary bioinformatic analysis approaches (linear modelling, using limma, and WGCNA) were then used to identify and prioritise diagnostic biomarkers by combining outputs of these pipelines. (B) Candidate protein biomarkers were then validated by multiplex antibody-based techniques (Luminex and proximity extension assay, PEA) in serum samples from a separate patient cohort in healthy control, pulmonary TB and other respiratory infections of mixed gender and ethnicity. High-performing combinatorial panels were identified for key clinical comparisons and diagnostic performance assessed in two separate patient cohorts using binary logistic regression and receiver operating characteristic curves.
iTRAQ: isobaric tags for relative and absolute quantification; nESI-MS2: nano-electrospray ionisation tandem mass spectrometry; limma: linear modelling for microarray data; WGCNA: whole gene correlation network analysis; PEA: proximity extension assay; NPX: normalised protein expression; TB: tuberculosis; HC: healthy control; ORI: other respiratory infections. ROC receiver operating characteristic

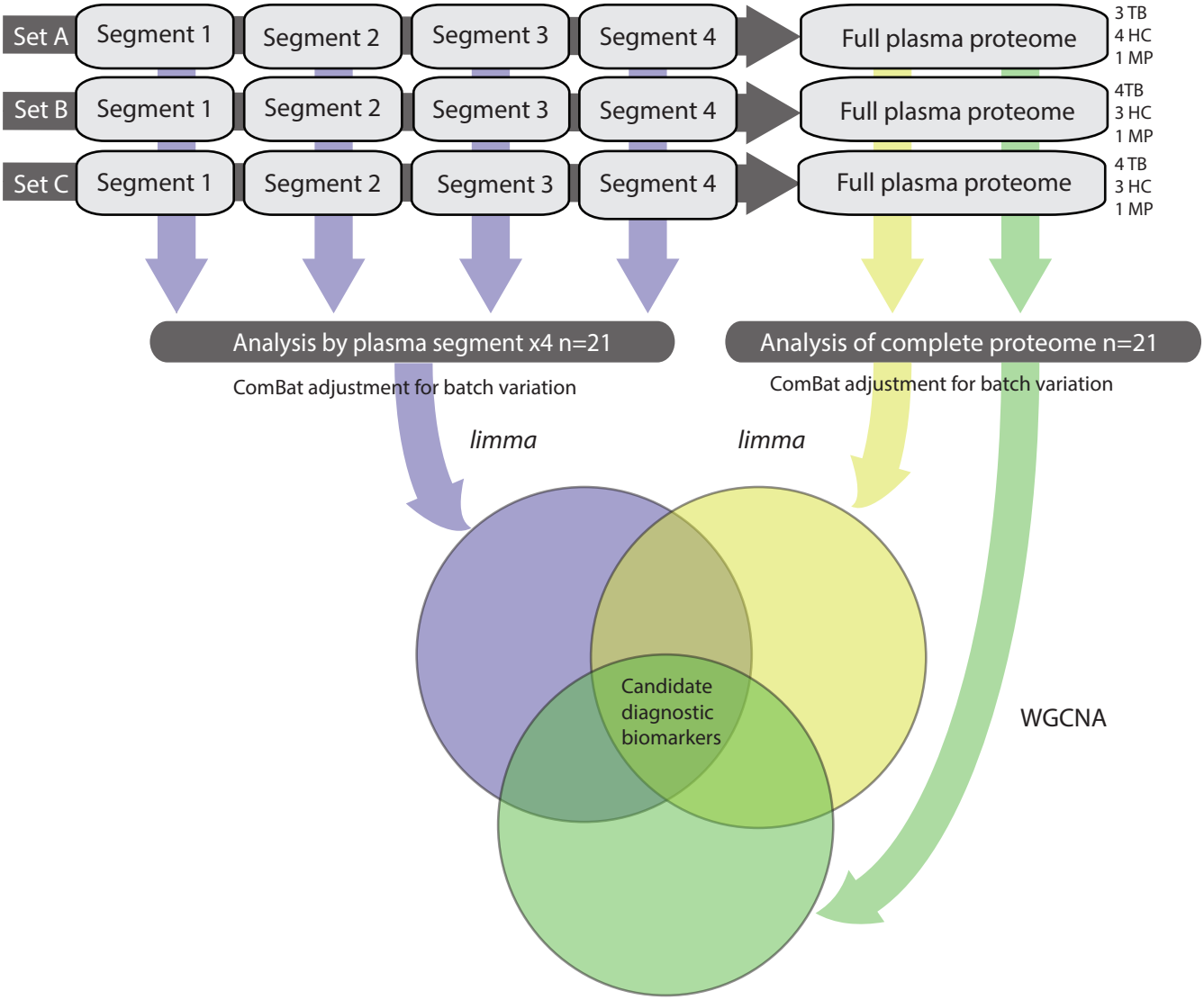


Figure 2: Bioinformatic analysis pipeline

Discovery proteomics experiments were conducted in 12 separate iTRAQ-labelled 8-plex experiments with block randomization of HC and TB samples into 3 experimental sets. Each plasma segment 8-plex experiment included one aliquot of a plasma masterpool. Grouped protein abundances were calculated across plasma segments for each experimental set to permit analysis over the whole plasma proteome. Protein abundances were then combined by plasma segment and by experimental set and adjusted for experimental batch variation using ComBat. Differential protein expression was analysed by *limma*. In parallel, the complete proteome was analysed by WGCNA to identify protein networks most strongly correlated with TB. Proteins identified as significant by all three bioinformatic approaches were then prioritized for validation.

iTRAQ: isobaric tags for relative and absolute quantification; Combat: adjustment for batch effects using an empirical Bayes framework (R package); WGCNA: whole gene network correlation analysis; *limma* linear modelling for microarray data (R package)

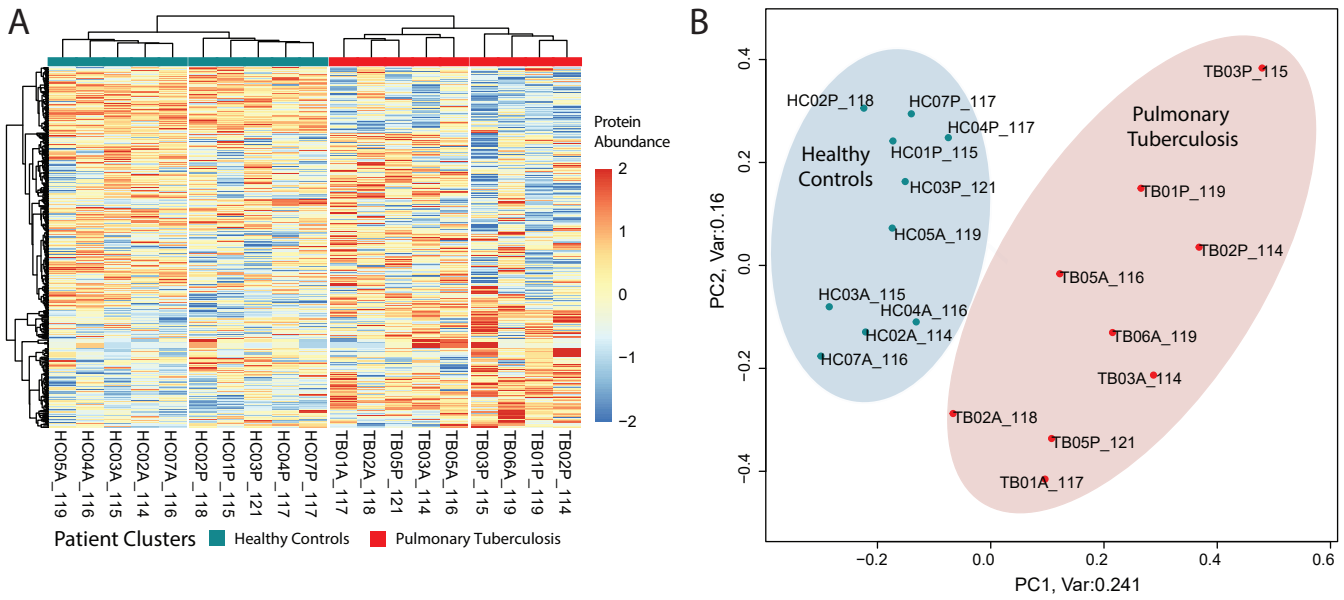


Figure 3: Summary data overview by unsupervised analysis

(A) Clustered heatmap for log_2-transformed fully quantified protein abundances (n=594) shows clear separation of protein abundances between the healthy control and pulmonary TB groups. iTRAQ tags and clinical groups are indicated. Within healthy controls distinct clustering was observed for discovery cohorts of different ethnicity (sample identification: A = South African, P = Peru). This was also observed within the TB group although some overlap occurred. **(B)** Principal component analysis (PCA) of log_2-transformed protein abundances demonstrates clear separation by clinical group, responsible for 24% of the variance within the dataset.

iTRAQ: isobaric tags for relative and absolute quantification; TB: tuberculosis; PCA: principal component analysis

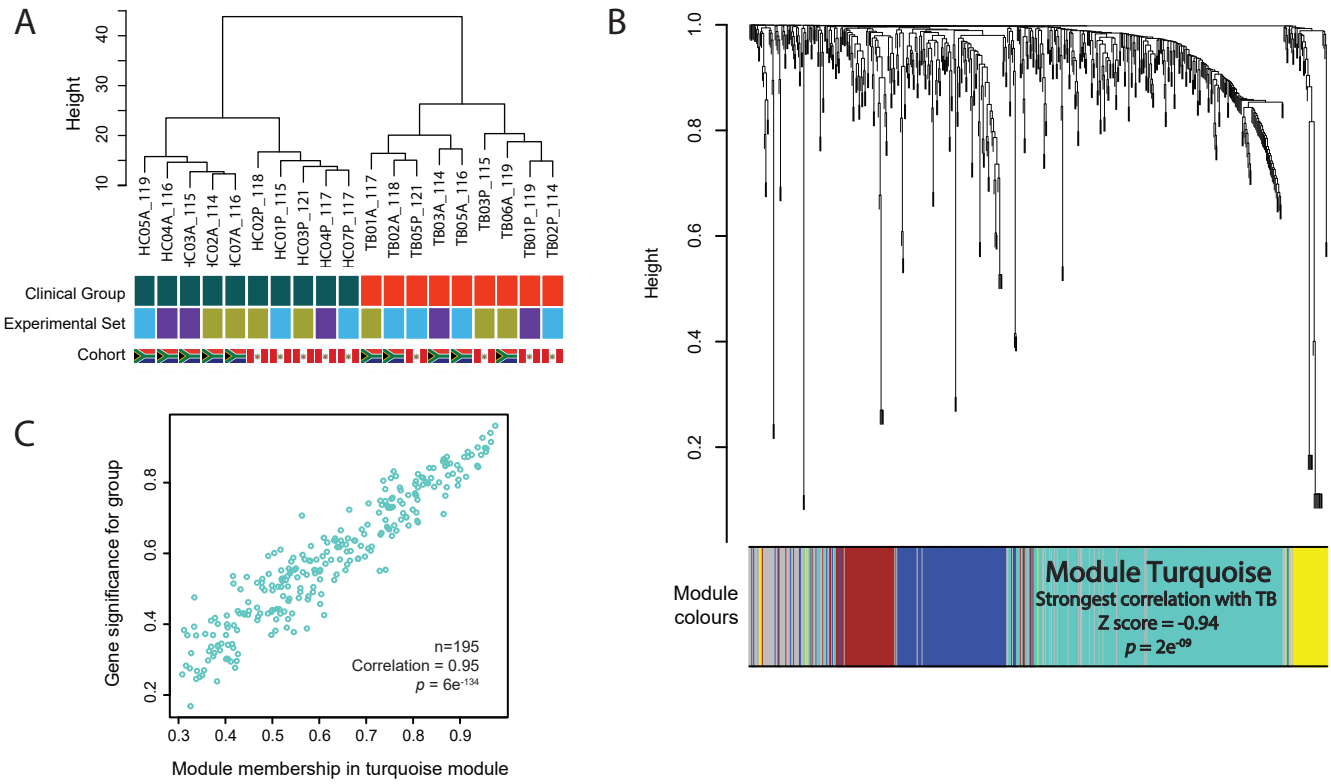


Figure 4: Whole genome correlation network analysis (WGCNA)
(A) Hierarchical clustering of samples showing discrete clusters by clinical group and absence of clustering by experimental batch. Discrete clustering by cohort ethnicity is again observed in the healthy control group, but not in TB patients. **(B)** Protein dendrogram and module colours. Module turquoise, containing 195 proteins, had the strongest correlation with TB (correlation (Z) score -0.94, $p=2^{e09}$). (C) A scatterplot of protein significance by clinical group confirming very high correlation of module turquoise with clinical group (0.95, $p=6^{e-134}$).
WGCNA: whole genome network correlation analysis; TB: tuberculosis

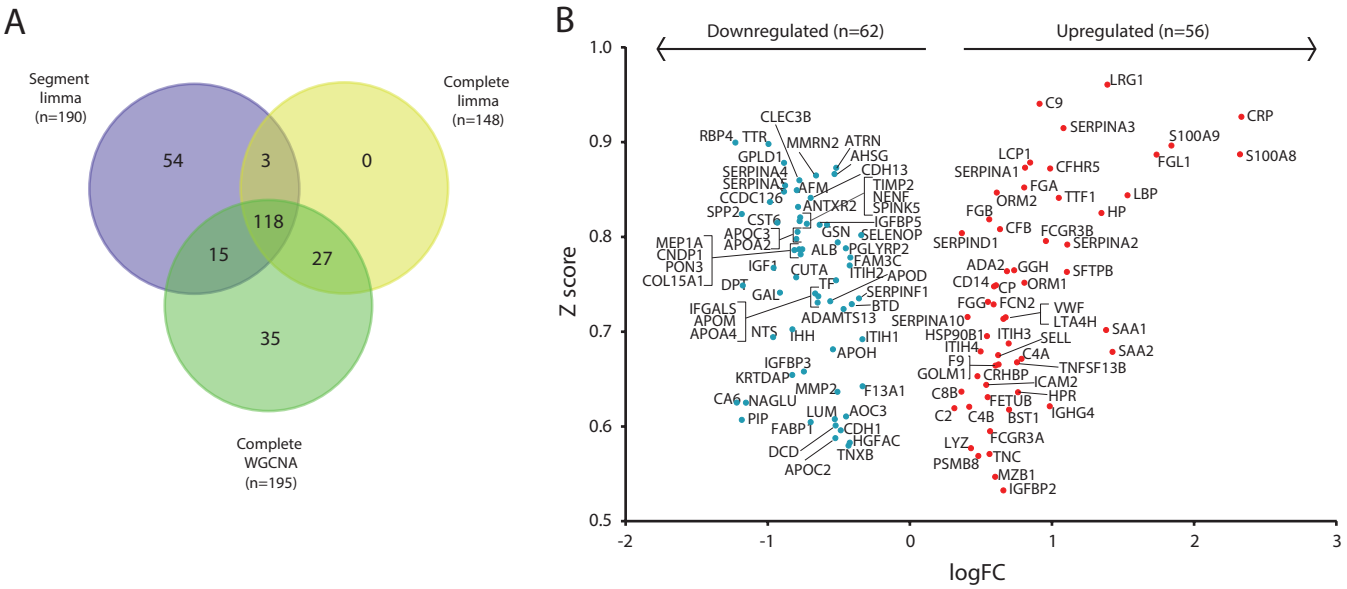


Figure 5: Complementary bioinformatic analyses identify 118 significantly differentially expressed plasma proteins in TB

(A) Proteins identified by each bioinformatic approach: 190 from *limma* analysis of segmental plasma proteomes, 148 by *limma* analysis of complete plasma proteomes and 195 proteins within WGCNA module turquoise. 118 proteins were common to significant via all three analytical approaches. **(B)** Volcano plot of all 118 significantly differentially expressed proteins by $\log_2$fold change by *limma* and correlation (Z) score from WGCNA. Markers in the upper outer quadrants have the highest fold changes and strongest correlation to TB. All markers have a *p* value <0.05 with adjustment for multiple testing within *limma*.

limma: linear modelling for microarray data (R package); WGCNA: whole genome correlation network analysis

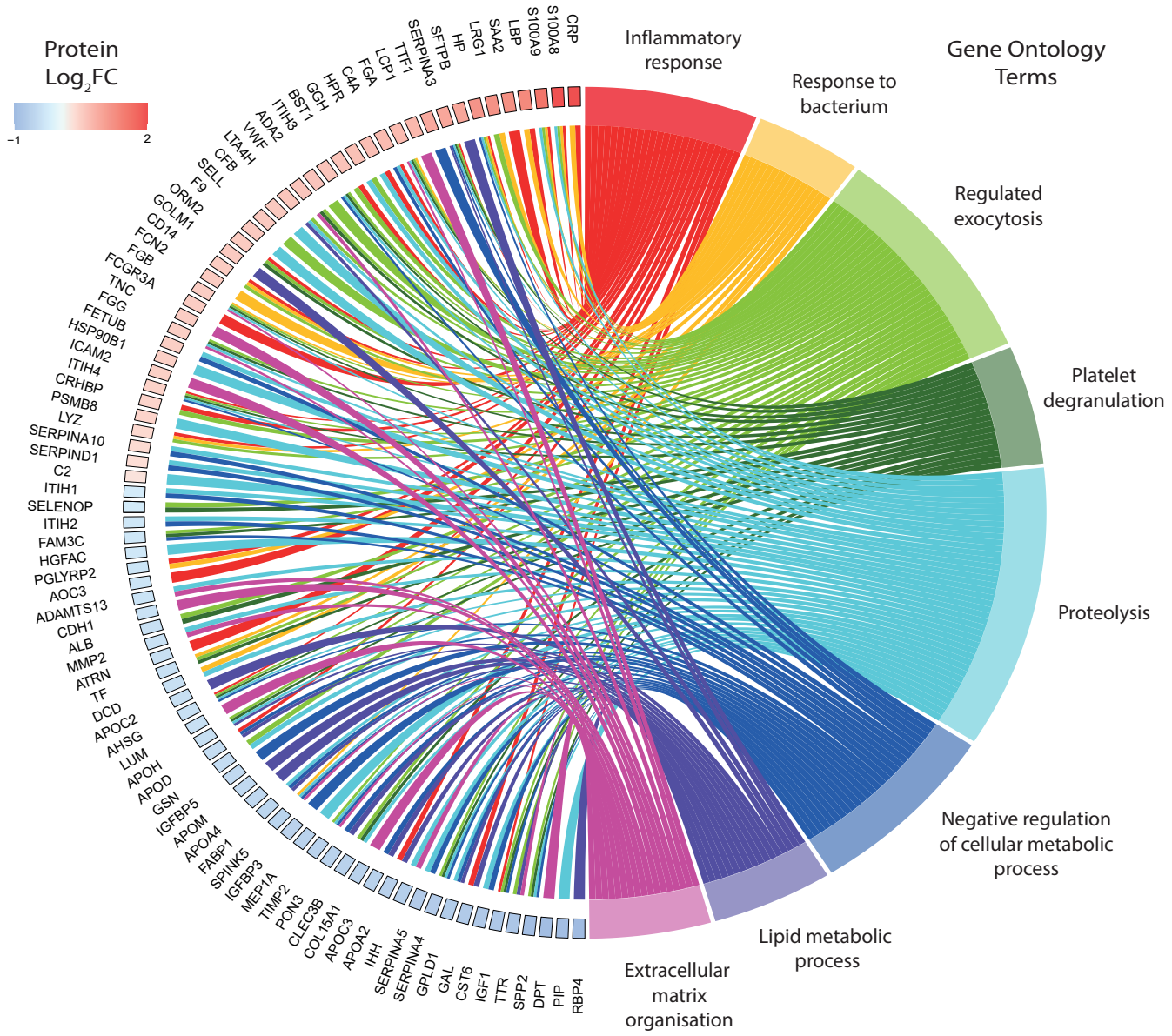


Figure 6: Divergently regulated proteins link with key biological processes in pulmonary TB
A chord plot depicting proteins with a $\log_2$ fold change greater than +/- 0.5 and their links to significantly enriched biological processes in TB. Gene ontology enrichment for biological process was performed using ShinyGO and only significant terms (FDR $q \leq 0.05$) are shown. Plot generated with the R package GOplots.

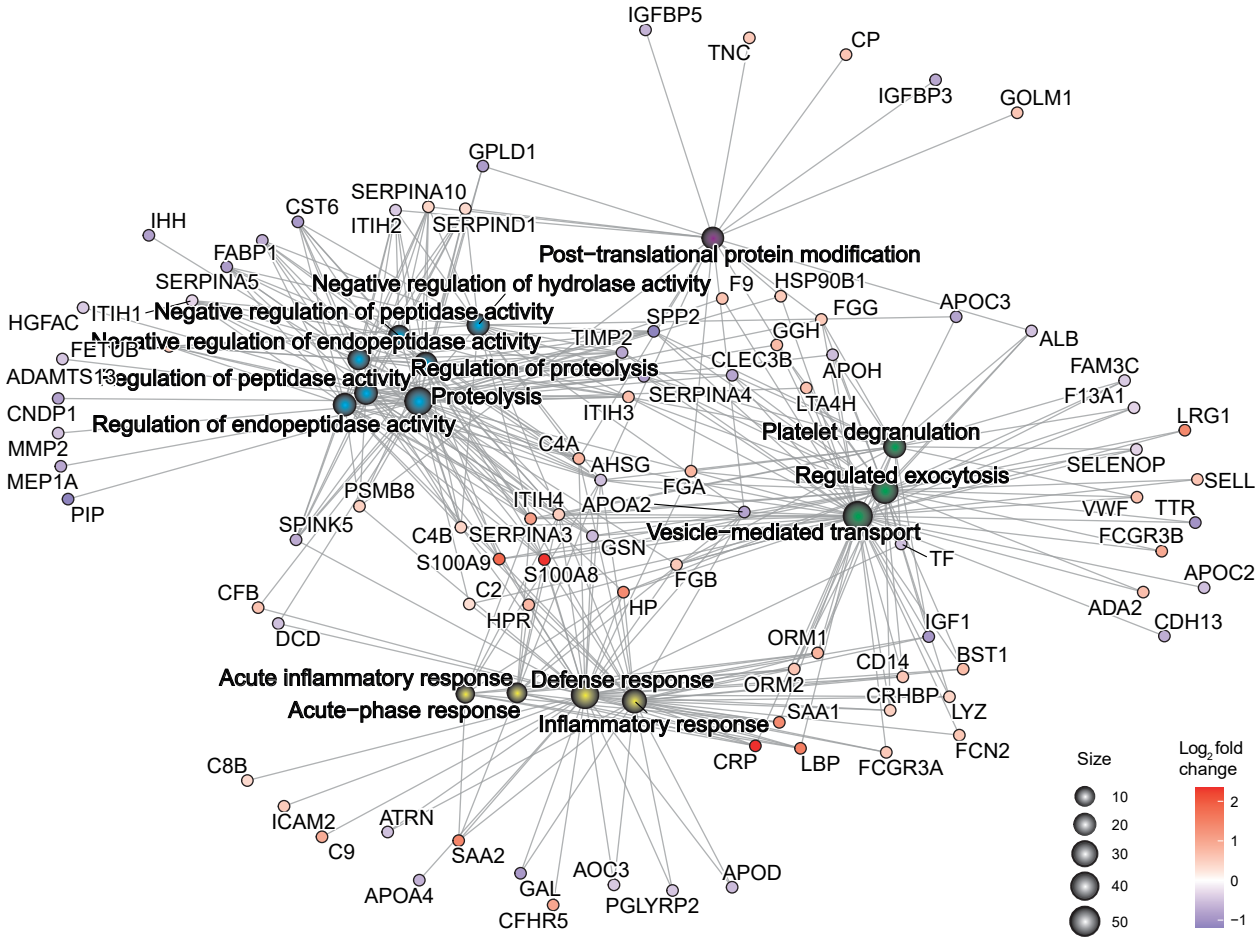


Figure 7: Physiological changes in TB are reflected in the plasma proteome
Functional enrichment analysis by biological process was performed on the 118 differentially expressed plasma proteins in TB. The gene concept network plot depicts the top 15 most enriched biological processes and their linkages to divergently regulated proteins. Gene ontology enrichment was performed using ShinyGO and the plot was generated using the cnetplot function in the R package GOplots.

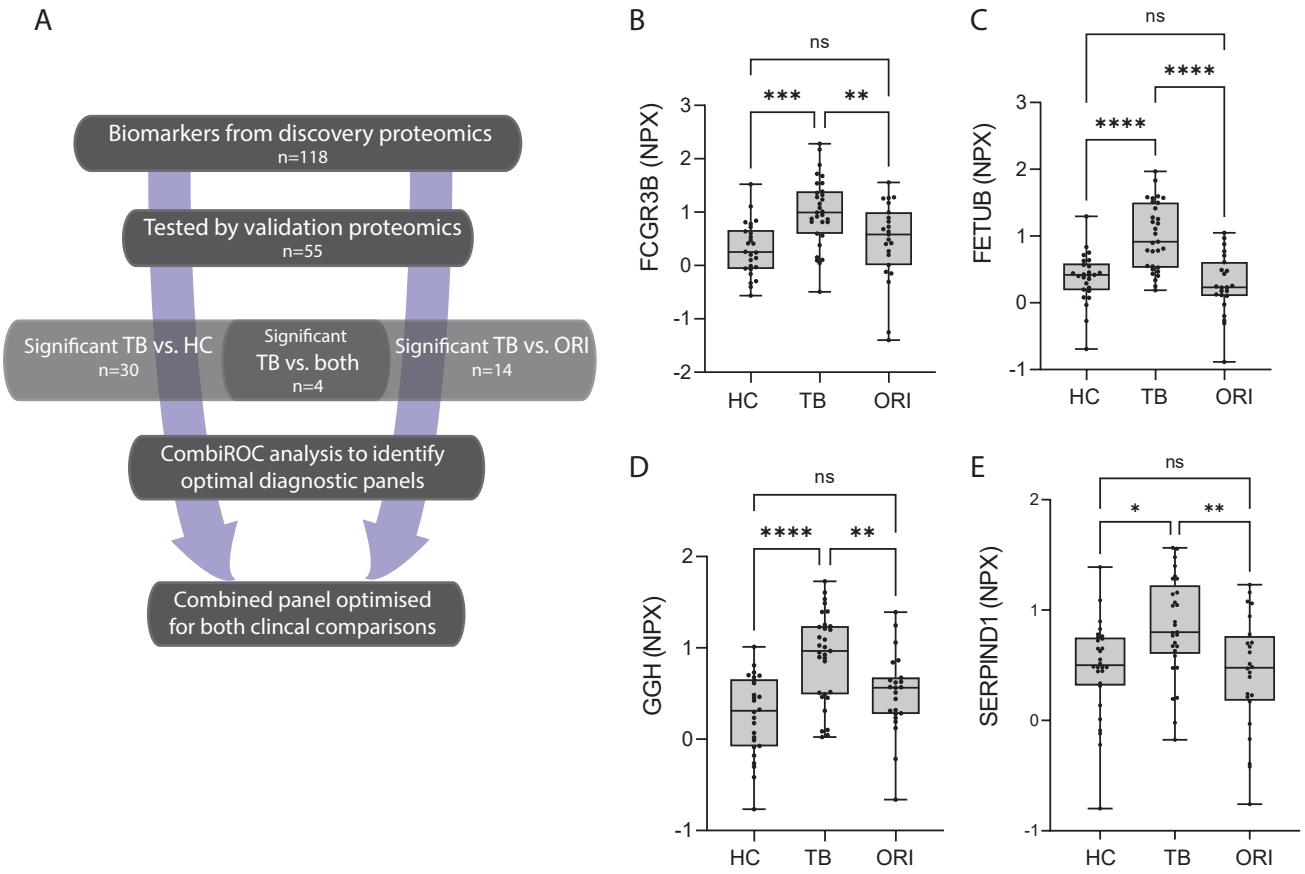
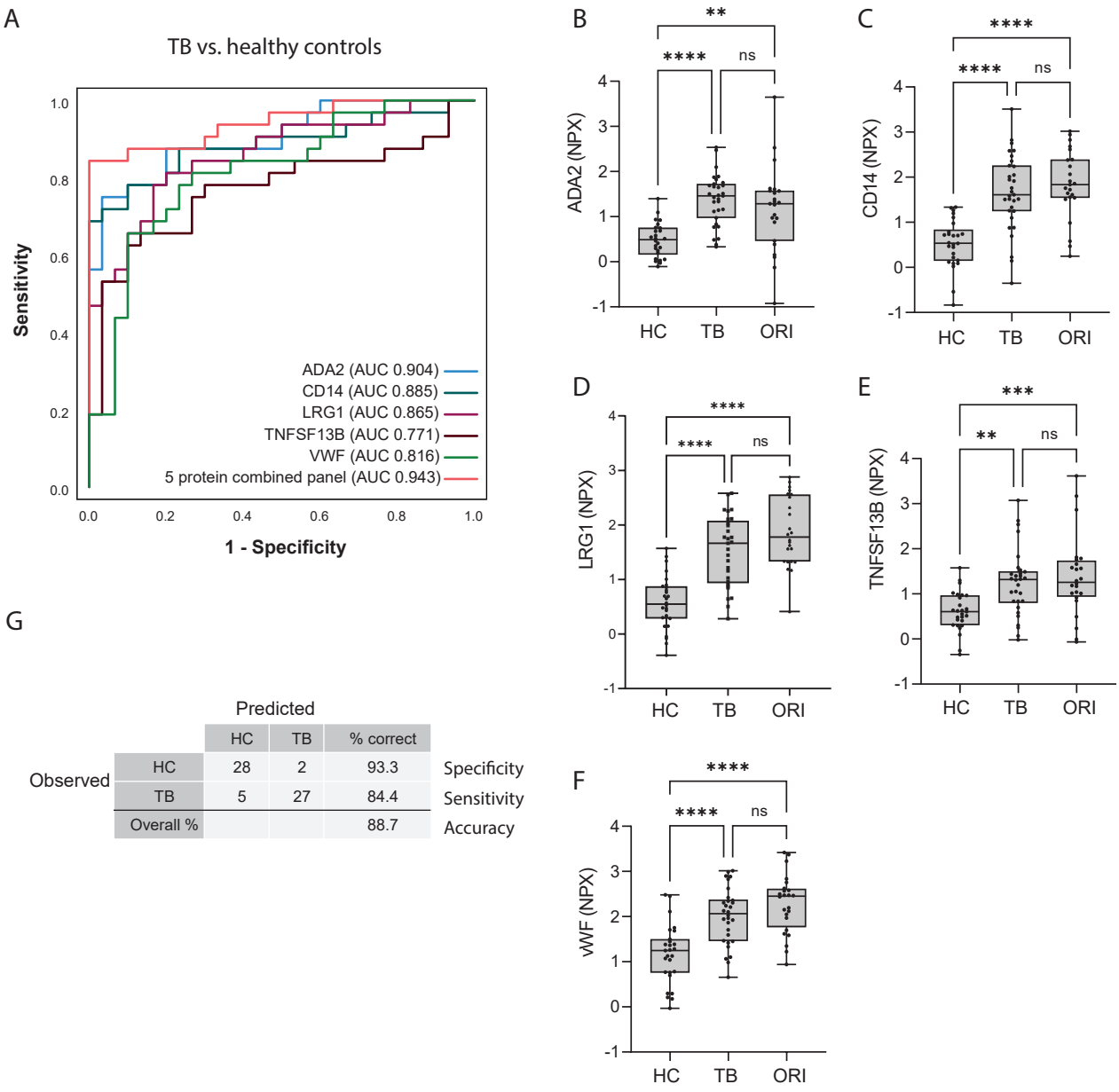


Figure 8: Discovery biomarker candidates validated by proximity extension analysis identify TB-specific biomarkers. (A) Flow chart outlining the analysis approach to identify significant biomarkers and the best performing biomarker combinations from our integrated proteomics approach. (B-E) Box and whisker plots of four protein biomarkers significantly differentially expressed in TB compared with both healthy controls and other respiratory infections by proximity extension assay. Boxes show median values and interquartile ranges, whiskers show minimum to maximum values. Statistical differences were calculated using one-way ANOVA with Tukey's multiple comparisons test for data with a Gaussian distribution and Kruskal-Willis test with Dunn's multiple comparisons test for non-parametrically distributed data.

ANOVA; analysis of variance; NPX: normalised protein expression (log_2scale); AUC: area under the curve; HC: healthy control (n = 30); TB: tuberculosis; (n= 32); ORI: other respiratory infections (n = 26); FCGR3B: low-affinity immunoglobulin receptor 3B; FETUB: fetuin-B; GGH gamma-glutamyl hydrolase ; SERPIND1 serpin D1, also known as heparin cofactor 2. ns meaning $p > 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$, *** $p \leq 0.001$; **** $p \leq 0.0001$



G

		Predicted			
		HC	TB	% correct	
Observed	HC	28	2	93.3	Specificity
	TB	5	27	84.4	Sensitivity
	Overall %			88.7	Accuracy

Figure 9: A five protein biomarker panel distinguishes pulmonary TB from healthy controls
(A) Receiver operating curve (ROC) characteristics of the best performing five biomarker combination distinguishing pulmonary TB from healthy controls, demonstrating an AUC of 0.943 (95% CI: 0.889 - 1.000) **(B-F)** Box and whisker plots of the five constituent proteins significantly differentially expressed in TB compared with healthy controls by proximity extension assay. Boxes show median values and interquartile ranges, whiskers show minimum to maximum values. Statistical differences were calculated using one-way ANOVA with Tukey's multiple comparisons test for data with a Gaussian distribution and Kruskal-Willis test with Dunn's multiple comparisons test for nonparametrically distributed data. **(G)** Classification grid illustrating diagnostic performance of the five protein biomarker panel in the validation cohort demonstrating a sensitivity of 84.4% (95% CI 67.3 - 94.3), specificity of 93.3% (95% CI: 75.8 - 98.8) and correct classification in 88.7% of cases.
ANOVA; analysis of variance; NPX: normalised protein expression (log_2scale); AUC: area under the curve; HC: healthy control (n = 30); TB: tuberculosis; (n= 32); ORI: other respiratory infection (n = 26); ADA2: adenosine deaminase 2; CD14: monocyte differentiation antigen CD14; LRG1: leucine-rich alpha-2-glycoprotein; TNFSF13B: tumour necrosis factor ligand superfamily member 13B; vWF: von Willebrand factor. ns meaning $p > 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$, *** $p \leq 0.001$; **** $p \leq 0.0001$

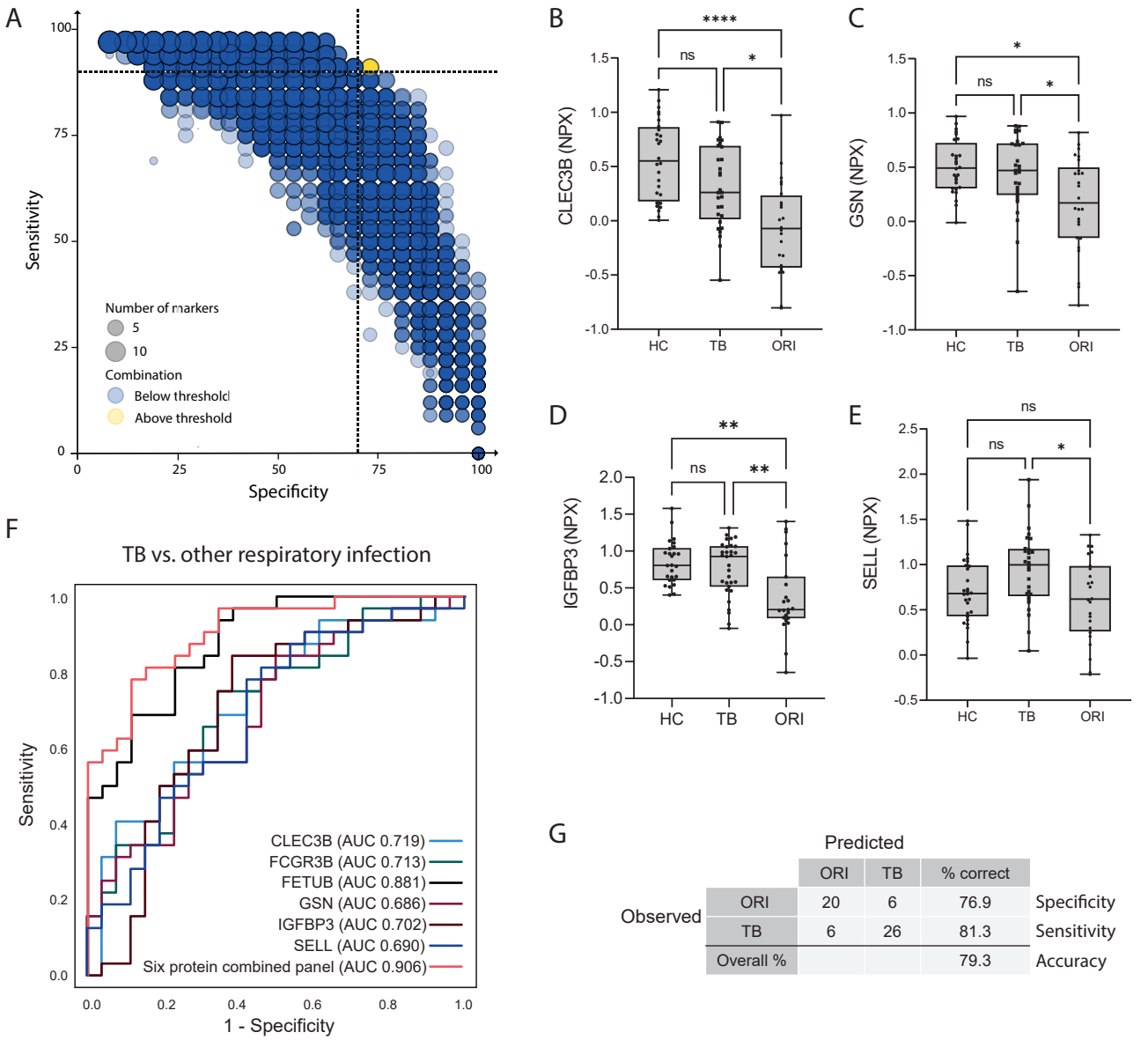
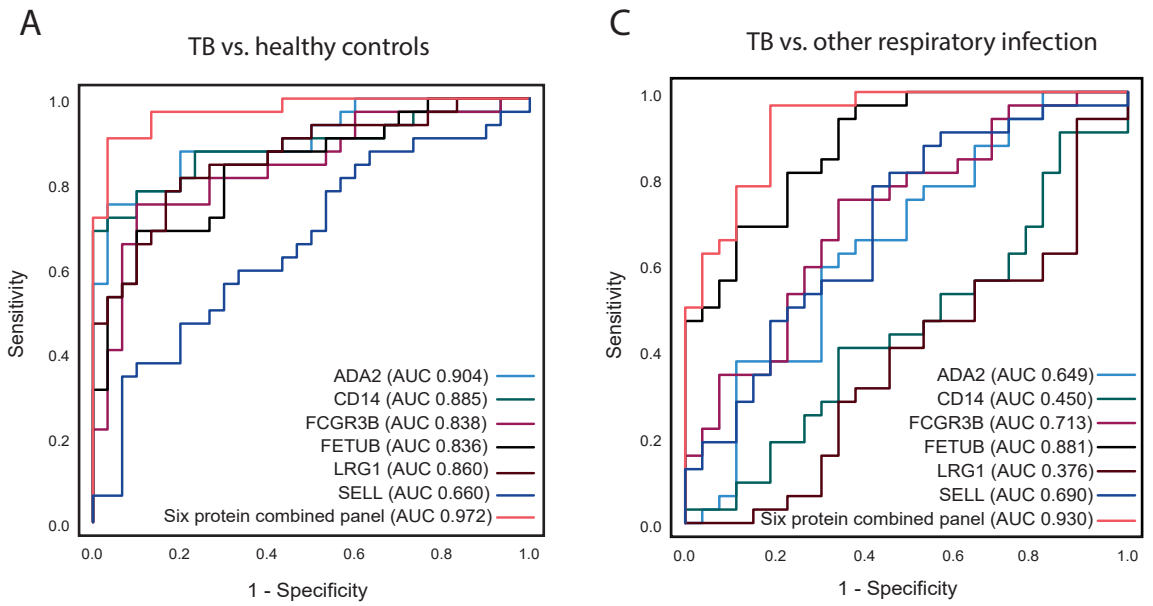


Figure 10: A six protein biomarker panel distinguishes pulmonary TB from other respiratory infections
(A) Bubble plot of possible protein combinations within the 14 proteins showing significant differential expression between TB and ORI groups, generated using CombiROC R package. Dotted lines at 90% sensitivity and 70% specificity corresponding to the WHO Target Product Profile for a triage test for active TB. (B-E) Box and whisker plots of protein biomarkers significantly differentially expressed in TB compared with other respiratory infections by proximity extension assay. Box and whisker plots of FCGR3B and FETUB are shown in Figure 8. Boxes show mean values and interquartile ranges, whiskers from minimum to maximum values. (F) Receiver operating curve (ROC) characteristics of best performing biomarker combination and constituent proteins. The six protein combined panel AUC 0.906 (95% CI: 0.833 - 0.908) (G) Classification grid illustrating diagnostic performance of the six protein biomarker panel in the validation cohort demonstrating a sensitivity of 81.3% (95% CI: 63.0 - 92.1), specificity of 76.9% (95% CI: 56.0 - 90.2) and correct classification in 79.3% of cases.

NPX: normalised protein expression (log_2scale); AUC: area under the curve; HC: healthy control; TB: tuberculosis; ORI: other respiratory infections; CLEC3B: tetranectin; GSN: gelsolin; IGFBP3: insulin-like binding protein 3; SELL: L-selectin; FCGR3B: low affinity immunoglobulin receptor 3B; FETUB: fetuin-B. ns meaning $p > 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$, *** $p \leq 0.001$; **** $p \leq 0.0001$



B

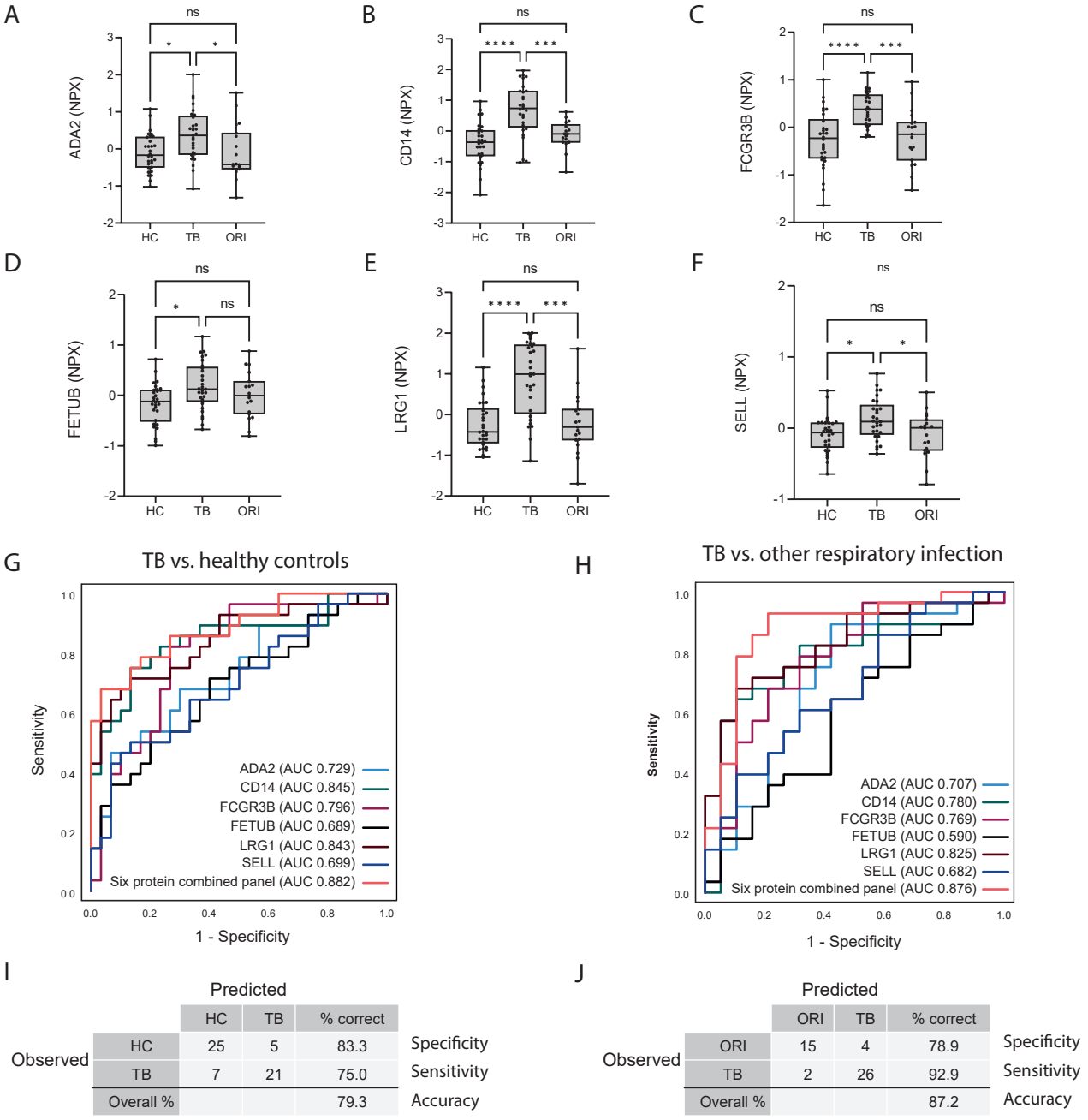
	Predicted			
Observed	HC	TB	% correct	
HC	27	3	90.0	Specificity
TB	3	29	90.6	Sensitivity
Overall %			90.3	Accuracy

D

	Predicted			
Observed	ORI	TB	% correct	
ORI	21	5	80.8	Specificity
TB	3	29	90.6	Sensitivity
Overall %			86.2	Accuracy

Figure 11: A final combined six protein panel discriminates patients with TB from both healthy controls and other respiratory infections

(A) ROC curve and (B) classification grid of the final six protein panel comprising FCGR3B, FETUB, LRG1, ADA2, CD14 and SELL, demonstrating discrimination of patients with TB from healthy controls (AUC 0.972 (95% CI: 0.937 - 1.000), sensitivity 90.6% (95% CI: 73.8 - 97.5), specificity 90.0% (95% CI: 72.3 - 97.4)).
(C) ROC curve and (D) classification grid of the final six protein panel discriminating patients with TB from patients with other respiratory infections (AUC 0.930 (95% CI: 0.867 - 0.993), sensitivity 90.6% (95% CI: 66.5 - 96.7), specificity 80.8% (95% CI: 68.2 - 94.5)).
All ROC curves and classification grids were generated using SPSS v28.0.1.0 after binary logistic regression for combined proteins. AUC was calculated under non-parametric assumption. TB was set as the positive test outcome and the test direction such that a larger test result indicates a more positive test.
ADA2: adenosine deaminase 2; CD14: monocyte differentiation antigen; FCGR3B: low-affinity immunoglobulin receptor 3B; FETUB: fetuin-B; LRG1: leucine-rich alpha-2-glycoprotein; SELL: L-selectin. TB: tuberculosis; HC: healthy control; ORI: other respiratory infection



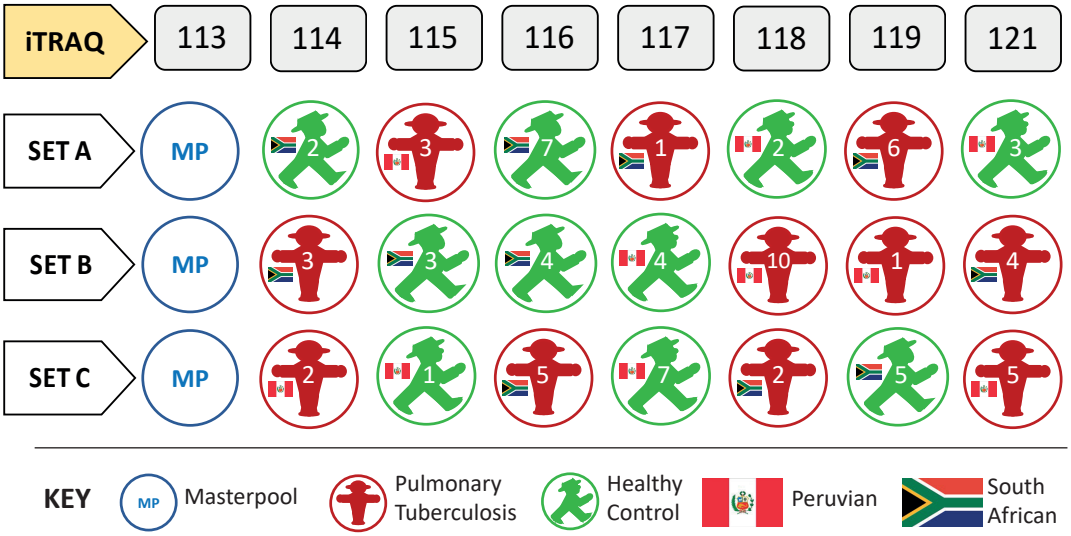
I

		Predicted HC	Predicted TB	% correct	
Observed	HC	25	5	83.3	Specificity
	TB	7	21	75.0	Sensitivity
	Overall %			79.3	Accuracy

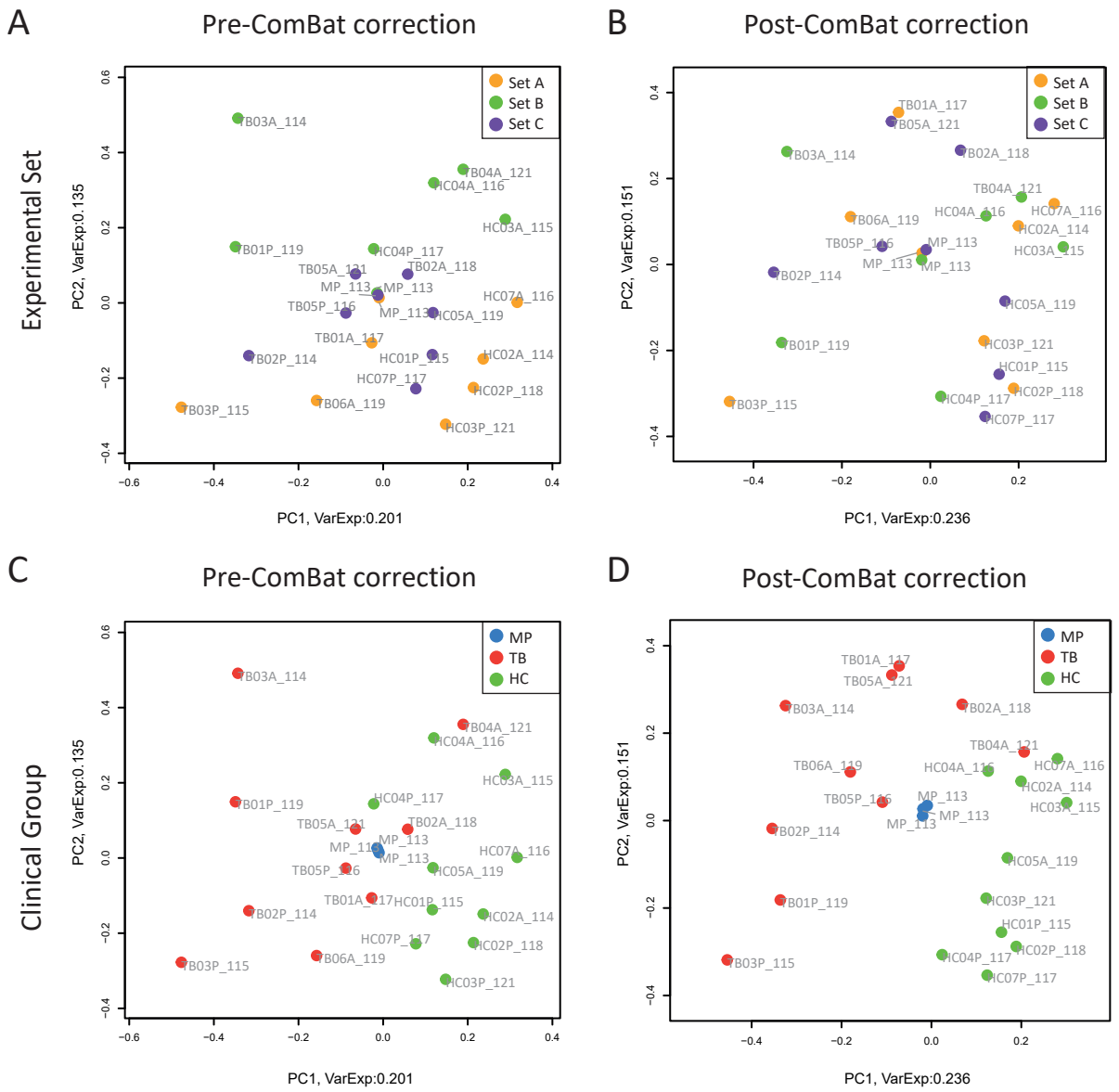
J

		Predicted ORI	Predicted TB	% correct	
Observed	ORI	15	4	78.9	Specificity
	TB	2	26	92.9	Sensitivity
	Overall %			87.2	Accuracy

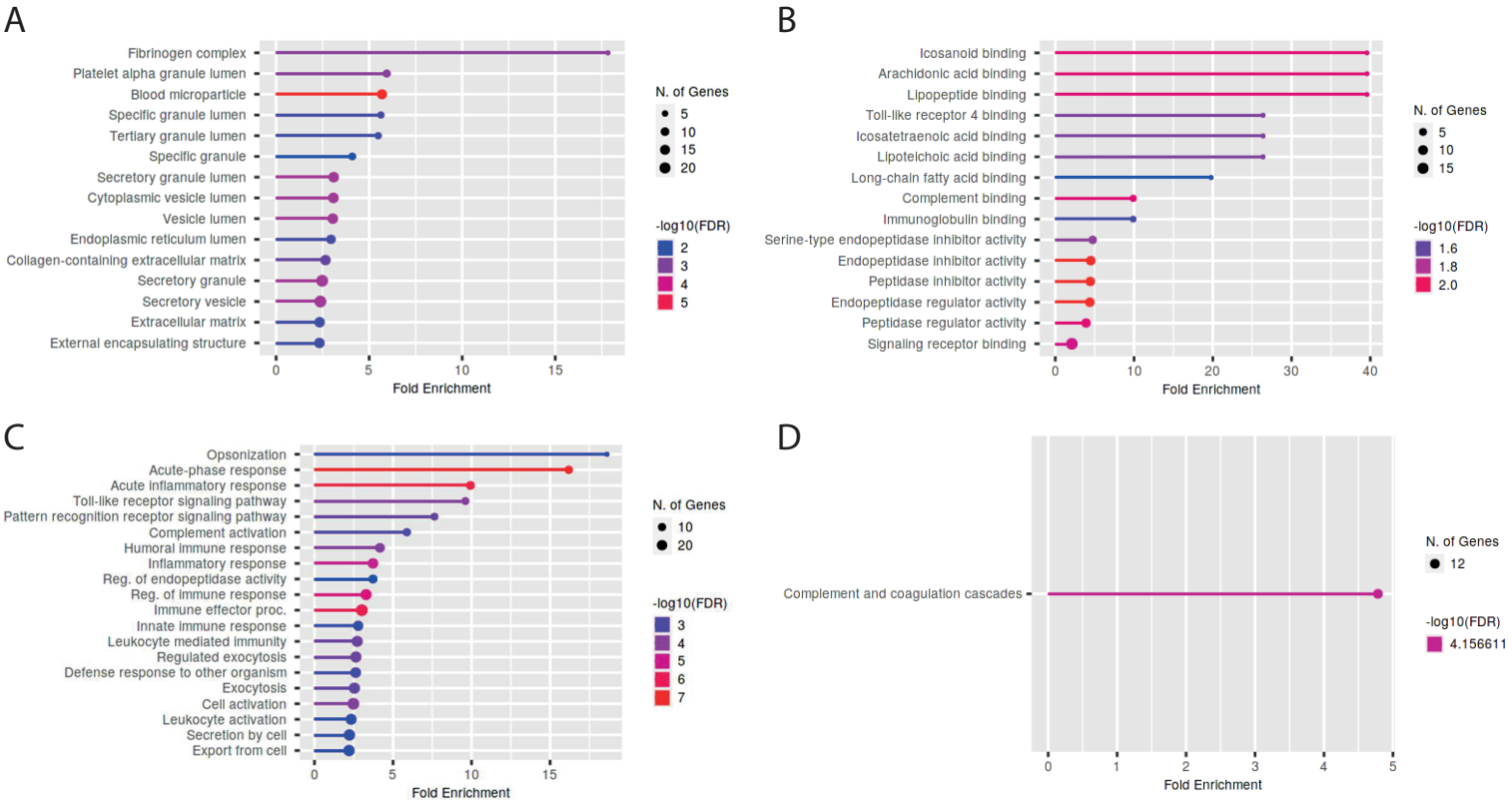
Figure 12: The final six protein panel differentiates TB from both HC and ORI in a separate clinical cohort
(A-F) Box and whisker plots of the six proteins in the panel in pulmonary TB compared with HC and ORI by proximity extension assay. Boxes show median values and interquartile ranges, whiskers show minimum to maximum values. Statistical differences were calculated using one-way ANOVA with Tukey's multiple comparisons test for data with a Gaussian distribution and Kruskal-Willis test with Dunn's multiple comparisons test for nonparametrically distributed data. (G) Receiver operating curve (ROC) characteristics of the six protein panel distinguishing pulmonary TB from healthy controls. The six protein combined panel AUC 0.882 (95% CI: 0.796 - 0.968). (H) Receiver operating curve (ROC) characteristics of the six protein panel distinguishing pulmonary TB from other respiratory infection, AUC 0.876 (95% CI: 0.765 - 0.987). (I) Classification grid illustrating diagnostic performance of the six protein panel distinguishing pulmonary TB from healthy controls demonstrating a sensitivity of 75.0% (95% CI: 54.8 - 88.6), specificity of 83.3% (95% CI: 64.5 - 93.7) and correct classification in 79.3% of cases in this cohort. (J) Classification grid illustrating diagnostic performance of the six protein panel distinguising pulmonary TB from other respiratory infection demonstrating a sensitivity of 92.9% (95% CI: 75.0 - 98.8), specificity of 78.9% (95% CI: 53.9 - 93.0) and correct classification in 87.2% of cases in this cohort.
All ROC curves and classification grids were generated using SPSS v28.0.1.0 after binary logistic regression for combined proteins. AUC was calculated under non-parametric assumption. TB was set as the positive test outcome and the test direction such that a larger test result indicates a more positive test.
ANOVA; analysis of variance; NPX: normalised protein expression (log_2scale); AUC: area under the curve; HC: healthy control (n = 30); TB: tuberculosis; (n= 29); ORI: other respiratory infection (n = 19); ADA2: adenosine deaminase 2; CD14: monocyte differentiation antigen CD14; LRG1: leucine-rich alpha-2-glycoprotein; TNFSF13B: tumour necrosis factor ligand superfamily member 13B; vWF: von Willebrand factor. ns meaning $p > 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$, *** $p \leq 0.001$; **** $p \leq 0.0001$



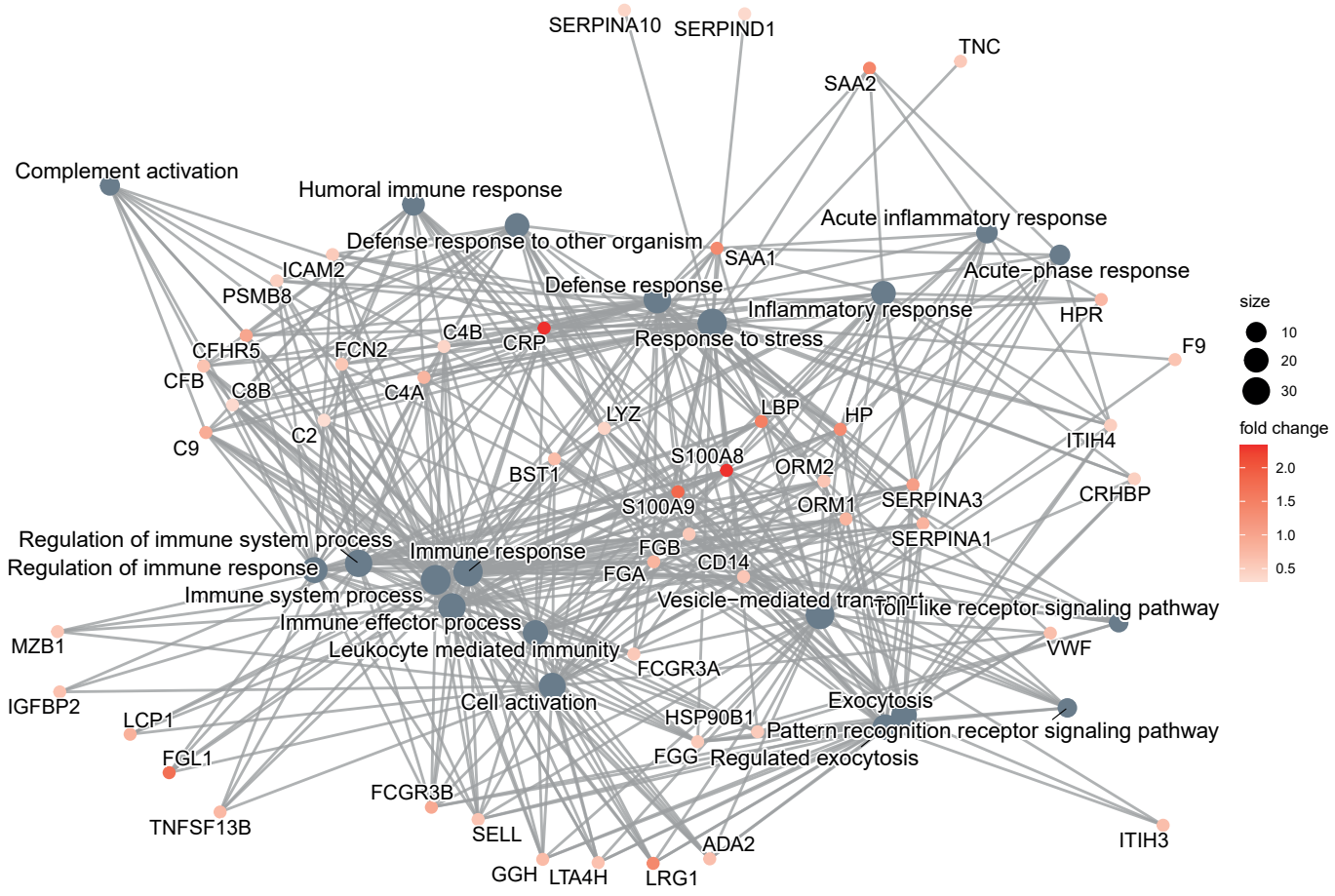
S1: Block randomised design of discovery proteomics experiment.
The design comprised three experimental sets: A, B & C. Peptides from each sample were iTRAQ-labelled following trypsin digestion according to this block randomised design. Each experimental set contained a bridging masterpool plasma sample which was labelled with iTRAQ tag 113 and either 3 or 4 plasma samples from healthy controls and individuals with pulmonary TB.



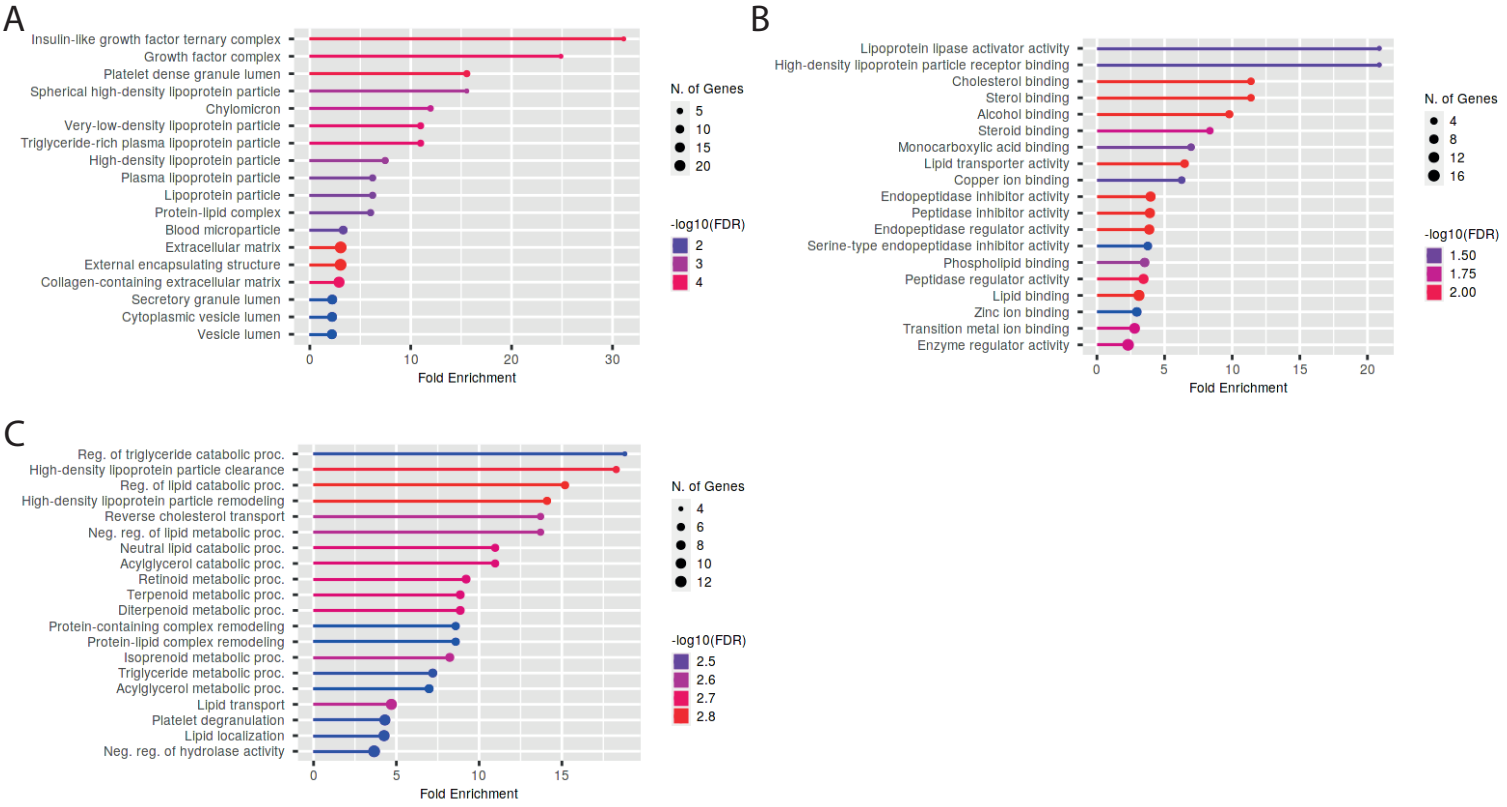
S2: Adjustment for batch effects between experimental sets.
Principal component analysis (PCA) depicting batch effect correction using the R package ComBat. PCA of protein abundances by experimental set before (A) and after (B) ComBat correction. PCA of protein abundances by clinical group before (C) and after (D) ComBat correction.
PC1 principal component one; PC2 principal component 2; VarExp explained variance; MP masterpool; TB pulmonary tuberculosis; HC healthy control.



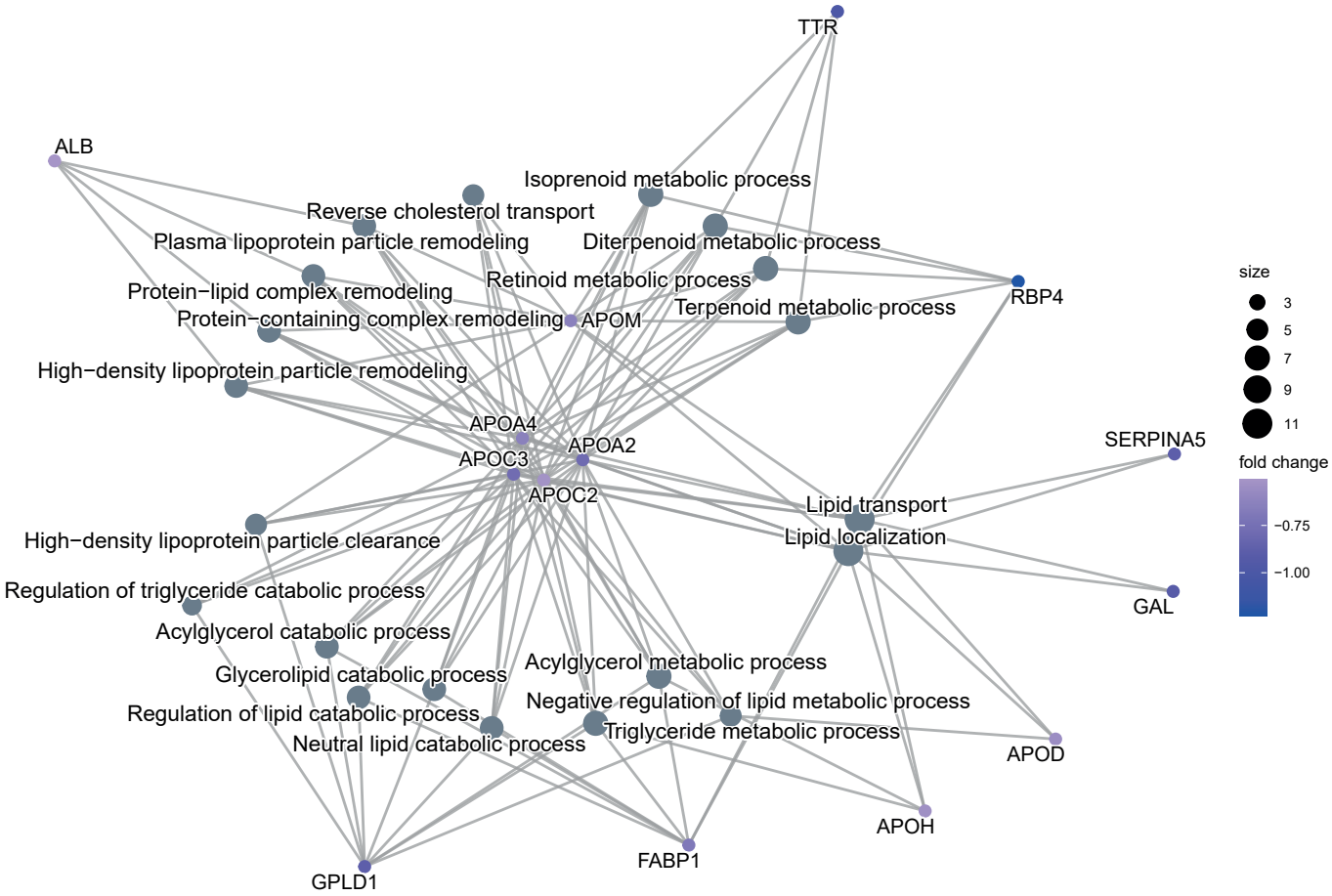
S3: Gene ontology analysis of significantly upregulated proteins.
Lollipop plots displaying fold enrichment and significance as false discovery rate (FDR) of ontology terms for (A) cellular compartment (B) molecular function (C) biological process and (D) KEGG pathways of upregulated proteins. The length of the lollipop is the fold enrichment of the pathway, the size of lollipop head indicates the number of proteins in the input dataset that are found within the pathway and the colour indicates the statistical significance of the enrichment. Gene ontology enrichment performed using ShinyGO against the background of the entire plasma proteome identified from discovery mass spectrometry proteomics.



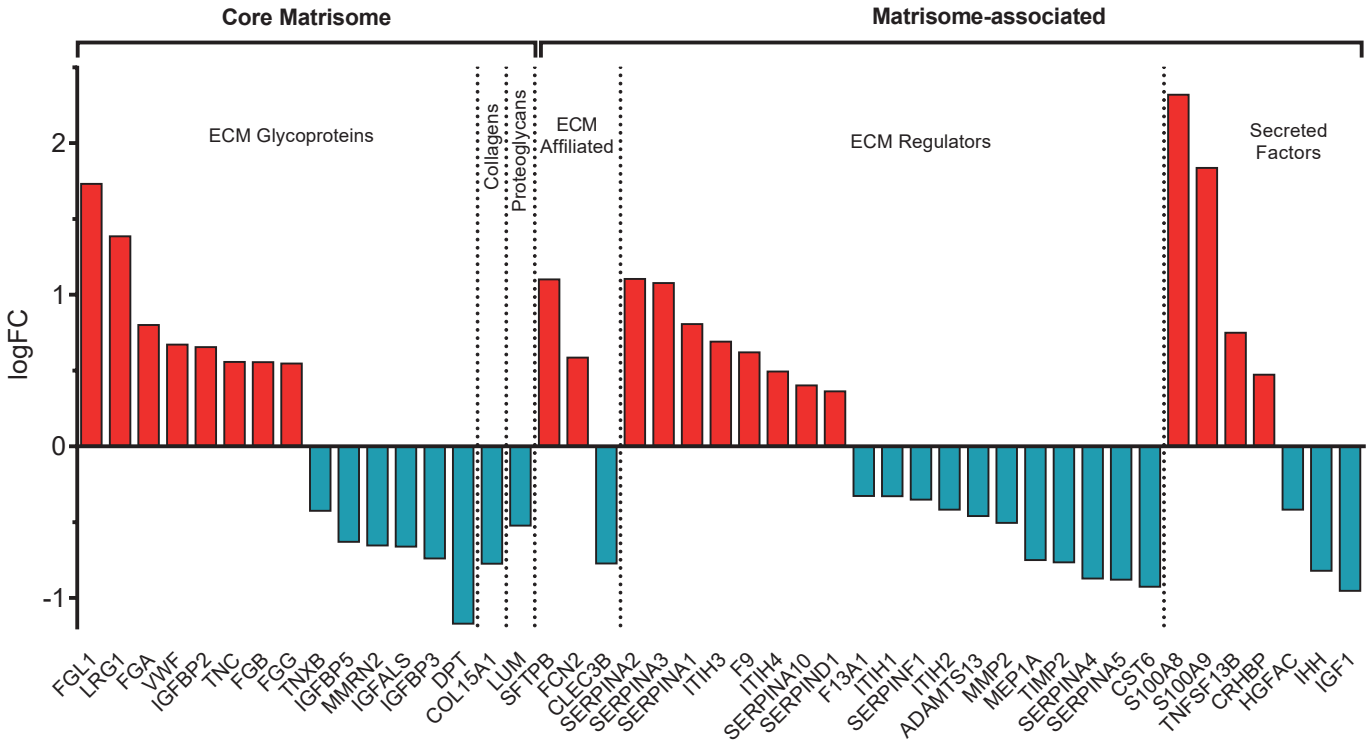
S4: Concept network plot of significantly upregulated proteins and their enriched biological processes.
Plot generated from ShinyGO enrichment by biological process of upregulated proteins using the cnetplot function in the R package GOplots. The top 20 most enriched pathways are displayed linked to their relevant differentially expressed proteins.



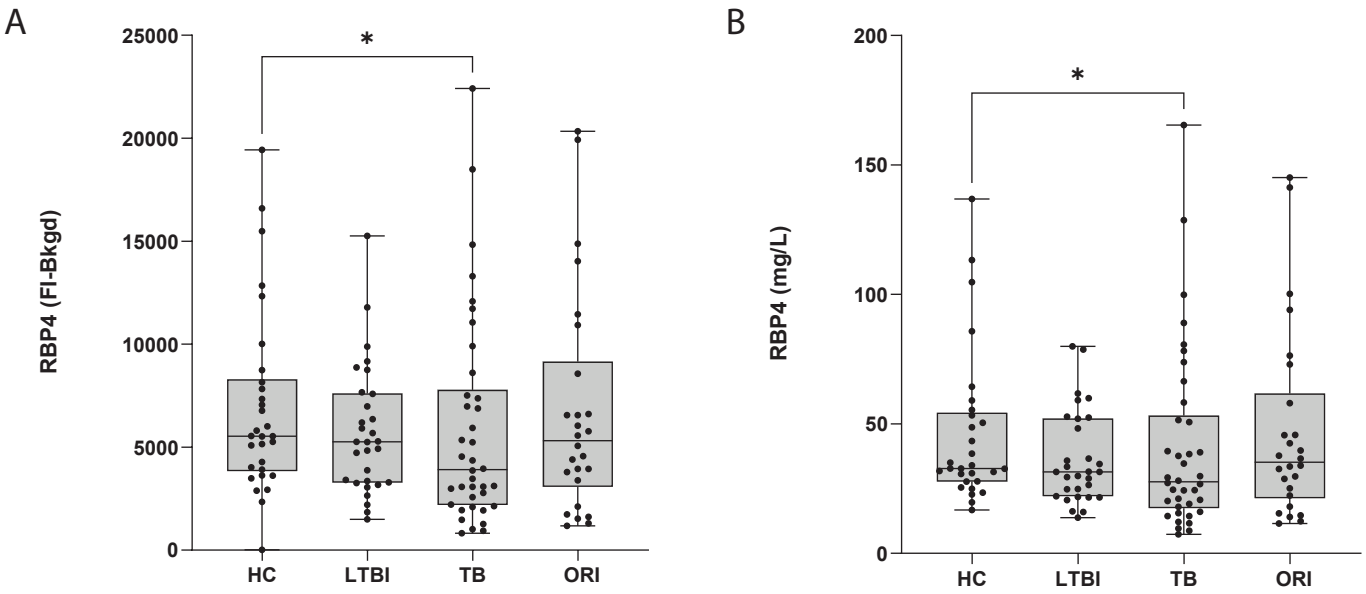
S5: Gene ontology analysis of significantly downregulated proteins.
Lollipop plots displaying fold enrichment and significance as false discovery rate (FDR) of ontology terms for (A) cellular compartment (B) molecular function (C) biological processes of downregulated proteins. The length of the lollipop is the fold enrichment of the pathway, the size of lollipop head indicates the number of proteins in the input dataset that are found within the pathway and the colour indicates the statistical significance of the enrichment. Gene ontology enrichment performed using ShinyGO against the background of the entire plasma proteome identified from discovery mass spectrometry proteomics.



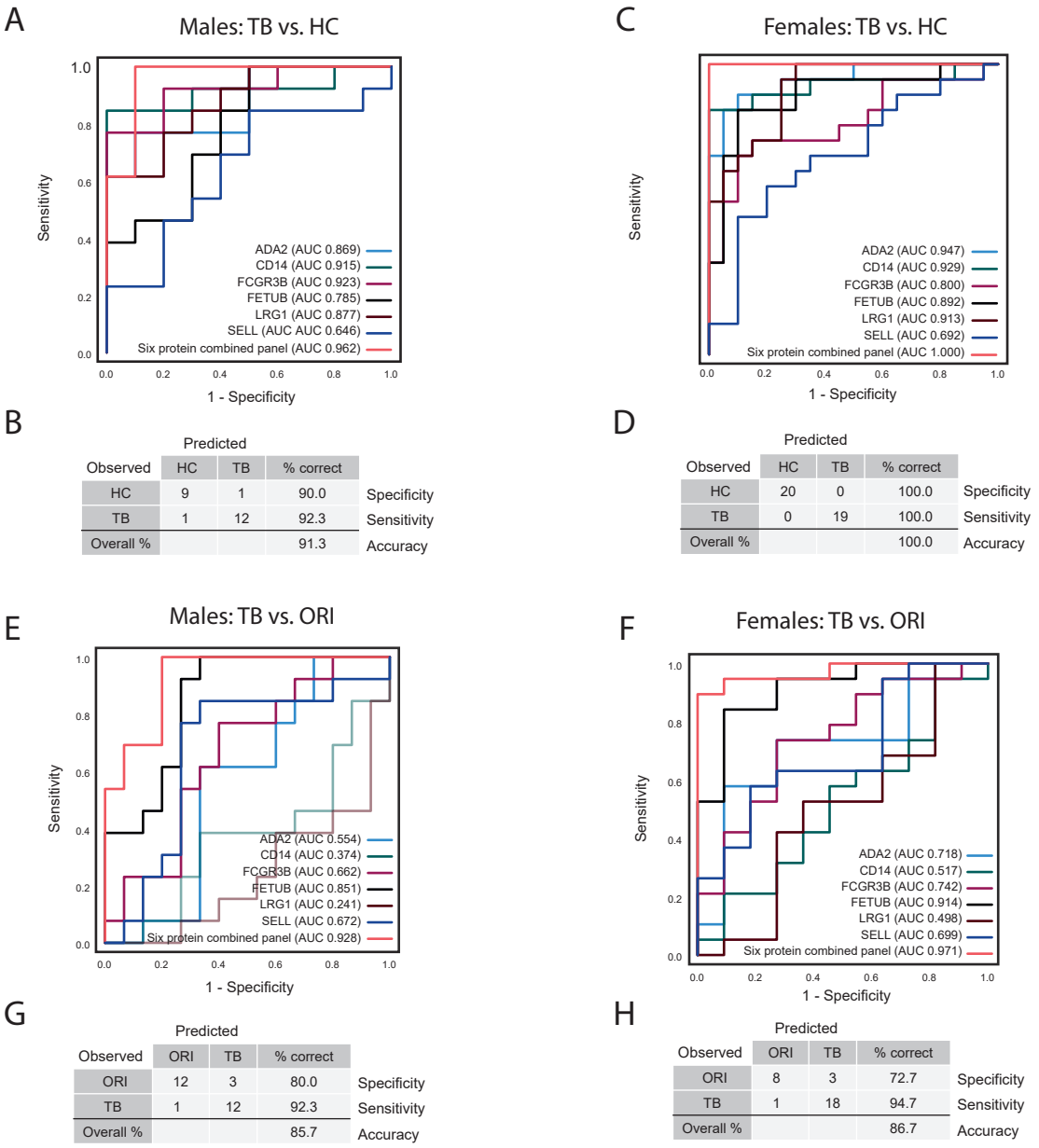
S6: Concept network plot of significantly downregulated proteins and their enriched biological processes.
Plot generated from ShinyGO enrichment by biological process of downregulated proteins using the cnetplot function in the R package GOplots. The top 20 most enriched pathways are displayed linked to their relevant differentially expressed proteins.



S7: Differential expression of 'matrisome'-associated proteins in the plasma of active pulmonary TB patients. 45 of 118 (38%) of differentially expressed plasma proteins in active pulmonary TB are contained within the 'matrisome', an ensemble of ~300 genes which encode the core extracellular matrix (ECM) and a further ~700 genes which encode ECM associated and regulatory proteins. Matrisome data accessed from http://matrisomeproject.mit.edu/other-resources/human-matrisome/ in September 2022.



S8: RBP4 is significantly downregulated in the plasma of patients with active pulmonary TB
(A) Box and whisker plot of fluorescence intensity values minus background levels in contrasting clinical groups of the MIMIC cohort. (B) Box and whisker plot of RBP4 serum concentration showing significant downregulation of RBP4 in active pulmonary TB. Values measured by Luminex assay.
HC healthy control; LTBI latent TB infection; TB active pulmonary TB; ORI other respiratory infections; * $p \leq 0.05$



B

Observed	Predicted HC	TB	% correct	
HC	9	1	90.0	Specificity
TB	1	12	92.3	Sensitivity
Overall %			91.3	Accuracy

D

Observed	Predicted HC	TB	% correct	
HC	20	0	100.0	Specificity
TB	0	19	100.0	Sensitivity
Overall %			100.0	Accuracy

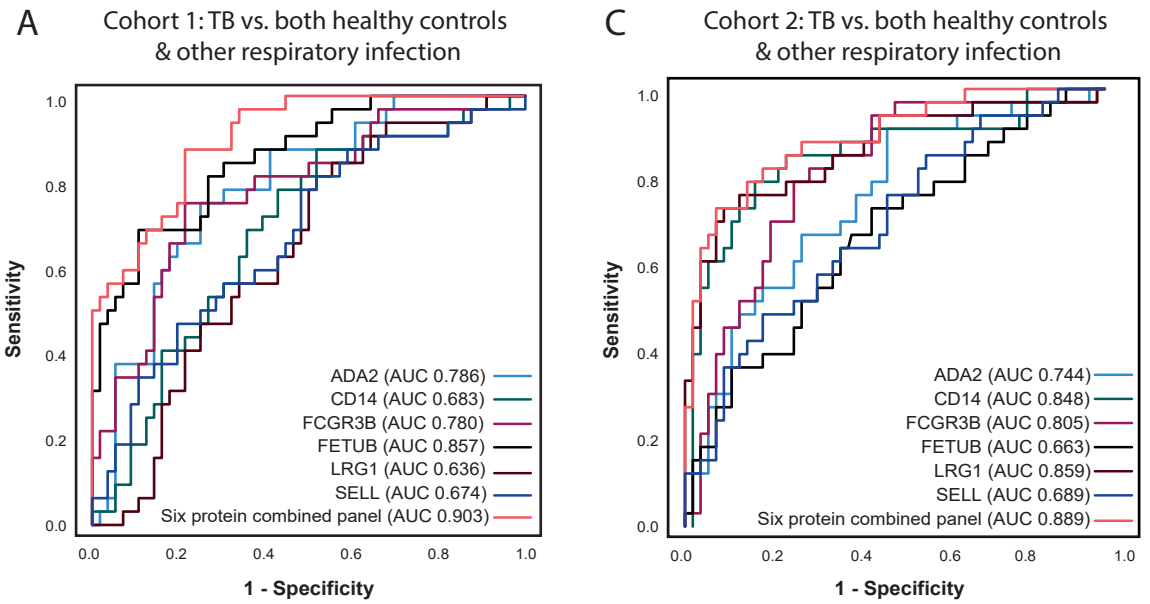
G

Observed	Predicted ORI	TB	% correct	
ORI	12	3	80.0	Specificity
TB	1	12	92.3	Sensitivity
Overall %			85.7	Accuracy

H

Observed	Predicted ORI	TB	% correct	
ORI	8	3	72.7	Specificity
TB	1	18	94.7	Sensitivity
Overall %			86.7	Accuracy

S9: Diagnostic performance of the final six protein panel in the UK MIMIC Cohort disaggregated by sex
(A) ROC curve and (B) classification grid of the final six protein panel demonstrating discrimination of male patients with TB from male healthy controls (AUC 0.962, sensitivity 92.3%, specificity 90.9%)
(C) ROC curve and (D) classification grid of the final six protein panel demonstrating discrimination of female patients with TB from female healthy controls (AUC 1.000, sensitivity 100%, specificity 100%)
(E) ROC curve and (F) classification grid of the final six protein panel demonstrating discrimination of male patients with TB from male ORI (AUC 0.928, sensitivity 92.3%, specificity 80.0%)
(G) ROC curve and (H) classification grid of the final six protein panel demonstrating discrimination of female patients with TB from female ORI (AUC 0.971, sensitivity 94.7%, specificity 72.7%)
All ROC curves and classification grids were generated using SPSS v28.0.1.0 after binary logistic regression for combined proteins. AUC were calculated under non-parametric assumption. TB was set as the positive test outcome and the test direction such that a larger test result indicates a more positive test.
ADA2: adenosine deaminase 2; CD14: monocyte differentiation antigen; FCGR3B: low-affinity immunoglobulin receptor 3B; FETUB: fetuin-B; LRG1: leucine-rich alpha-2-glycoprotein; SELL: L-selectin. TB: tuberculosis; HC: healthy control; ORI: other respiratory infection



B

	Predicted			
Observed	HC & ORI	TB	% correct	
HC & ORI	50	6	89.3	Specificity
TB	12	20	62.5	Sensitivity
Overall %			79.5	Accuracy

D

	Predicted			
Observed	HC & ORI	TB	% correct	
HC & ORI	49	5	90.7	Specificity
TB	9	24	72.7	Sensitivity
Overall %			83.9	Accuracy

S10: The final combined six protein panel discriminates patients with TB from a combined group of both healthy controls and other respiratory infections with high specificity in both patient cohorts
(A) ROC curve and (B) classification grid of the final six protein panel comprising FCGR3B, FETUB, LRG1, ADA2, CD14 and SELL, demonstrating discrimination of patients with TB from both healthy controls and other respiratory infection as a combined group in Cohort 1 (AUC 0.903, sensitivity 62.5%, specificity 89.3%)
(C) ROC curve and (D) classification grid of the final six protein panel comprising FCGR3B, FETUB, LRG1, ADA2, CD14 and SELL, demonstrating discrimination of patients with TB from both healthy controls and other respiratory infection as a combined group in Cohort 2 (AUC 0.889, sensitivity 72.7%, specificity 90.7%).
ROC curves and classification grids were generated using SPSS v28.0.1.0 after binary logistic regression for combined proteins. AUC was calculated under non-parametric assumption. TB was set as the positive test outcome and the test direction such that a larger test result indicates a more positive test.
ADA2: adenosine deaminase 2; CD14: monocyte differentiation antigen; FCGR3B: low-affinity immunoglobulin receptor 3B; FETUB: fetuin-B; LRG1: leucine-rich alpha-2-glycoprotein; SELL: L-selectin. TB: tuberculosis; HC: healthy control; ORI: other respiratory infection