

Comprehensive plasma proteomic profiling reveals biomarkers for active tuberculosis

Diana J. Garay-Baquero, ... , Spiros D. Garbis, Paul Elkington

JCI Insight. 2020. <https://doi.org/10.1172/jci.insight.137427>.

Clinical Medicine In-Press Preview Infectious disease

Background

Tuberculosis (TB) kills more people than any other infection and new diagnostic tests to identify active cases are urgently required. We aimed to discover and verify novel markers for TB in non-depleted plasma.

Methods

We applied an optimised quantitative proteomics discovery methodology based on multidimensional and orthogonal liquid chromatographic separation hyphenated with high-resolution mass spectrometry (q3D LC-MS) to study non-depleted plasma of 11 patients with active TB compared to 10 healthy control donors. Prioritised candidates were verified in an independent UK-based (n=118) and a South African cohorts (n=203).

Results

We generated the most comprehensive TB plasma proteome to date, profiling 5022 proteins spanning 11 orders-of-magnitude concentration range with diverse biochemical and molecular properties. We further analysed the predominantly low molecular weight sub-proteome; identifying 46 proteins with significantly increased and 90 with decreased abundance (peptide FDR $\leq 1\%$, q-value ≤ 0.05). Biological network analysis showed regulation of new pathways involving lipid and organophosphate ester transport. Verification was performed for novel candidate biomarkers (CFHR5, ILF2) in two independent cohorts. These proteins were elevated in both TB and other respiratory diseases (ORD). Receiver-operating-characteristics analyses using a 5-protein panel (CFHR5, LRG1, CRP, LBP and SAA1) [...]

Find the latest version:

<https://jci.me/137427/pdf>



1 **Comprehensive plasma proteomic profiling reveals biomarkers for active tuberculosis**

2

3 Diana J. Garay-Baquero^{1,2,3}, Cory H. White^{1,†}, Naomi F. Walker^{4,5,6,7}, Marc Tebruegge^{8,9,10}, Hannah
4 F. Schiff¹, Cesar Ugarte Gil^{7,11}, Stephen Morris-Jones^{12,13}, Ben G Marshall^{1,14}, Antigoni
5 Manousopoulou^{2,#}, John Adamson¹⁵, Andres F. Vallejo¹, Magdalena K. Bielecka¹, Robert J.
6 Wilkinson^{4,6,16,17}, Liku B. Tezera^{1,2}, Christopher H. Woelk^{1,†}, Spiros D. Garbis^{2,3,18}, Paul Elkington^{1,}
7 2,14

8

9 ¹ School of Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton,
10 UK.

11 ² Institute for Life Sciences, University of Southampton, UK.

12 ³ Proteome Exploration Laboratory, Beckman Institute, California Institute of Technology, Pasadena,
13 CA, USA.

14 ⁴ Wellcome Centre for Infectious Diseases Research in Africa, Institute of Infectious Disease and
15 Molecular Medicine, University of Cape Town, Observatory 7925, Republic of South Africa.

16 ⁵ Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK.

17 ⁶ Department of Medicine, University of Cape Town, Observatory 7925, Republic of South Africa.

18 ⁷ TB Centre and Department of Clinical Research, London School of Hygiene and Tropical Medicine,
19 Keppel St, London, UK.

20 ⁸ Department of Paediatric Infectious Diseases & Immunology, Evelina London Children's Hospital,
21 Guy's and St. Thomas' NHS Foundation Trust, London, UK.

22 ⁹ Department of Infection, Immunity, and Inflammation, UCL Great Ormond Street Institute of Child
23 Health, University College London, London, UK.

24 ¹⁰ Department of Paediatrics, University of Melbourne, Parkville, Australia.

25 ¹¹ Instituto de Medicina Tropical Alexander von Humboldt, School of Medicine, Universidad Peruana
26 Cayetano Heredia, Lima, Peru.

27 ¹² Department of Microbiology, University College London Hospitals NHS Trust, London, UK.

28 ¹³ Division of Infection and Immunity, University College London, London, UK.

29 ¹⁴ NIHR Biomedical Research Centre, University Hospital NHS Foundation Trust, Southampton, UK.

30 ¹⁵ Pharmacology Core, Africa Health Research Institute (AHRI), Durban, South Africa.

31 ¹⁶ The Francis Crick Institute, London, UK.

32 ¹⁷ Department of Infectious Diseases, Imperial College, London, UK.

33 ¹⁸ Cancer Sciences Division, Faculty of Medicine, University of Southampton, UK

34 [†] Current address: Exploratory Science Center, Merck & Co., Inc., Cambridge, Massachusetts, USA.

35 [#] Current address: Department of Immuno-Oncology, Beckman Research Institute, City of Hope
36 National Medical Center, Duarte, CA, USA.

37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57

Address for correspondence:

Professor Paul T 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

Dr. Spiros D Garbis
Proteome Exploration Laboratory
Beckman Institute
Division of Biology and Biological Engineering
California Institute of Technology
Pasadena, California, 91125, USA
Tel: 00 1 626 395 2339
E-mail: sgarbis@caltech.edu

Running title: Comprehensive TB plasma proteome

Keywords: plasma, proteomics, tuberculosis, diagnosis, biomarker, bioinformatic analysis

The funding body requires a creative Commons CC-BY license for this work
No authors declare a conflict of interest

58 **Abstract**

59

60 **Background**

61 Tuberculosis (TB) kills more people than any other infection and new diagnostic tests to identify
62 active cases are urgently required. We aimed to discover and verify novel markers for TB in non-
63 depleted plasma.

64

65 **Methods**

66 We applied an optimised quantitative proteomics discovery methodology based on multidimensional
67 and orthogonal liquid chromatographic separation hyphenated with high-resolution mass spectrometry
68 (q3D LC-MS) to study non-depleted plasma of 11 patients with active TB compared to 10 healthy
69 control donors. Prioritised candidates were verified in an independent UK-based (n=118) and a South
70 African cohorts (n=203).

71

72 **Results**

73 We generated the most comprehensive TB plasma proteome to date, profiling 5022 proteins spanning
74 11 orders-of-magnitude concentration range with diverse biochemical and molecular properties. We
75 further analysed the predominantly low molecular weight sub-proteome; identifying 46 proteins with
76 significantly increased and 90 with decreased abundance (peptide false discovery rate, FDR $\leq 1\%$, q -
77 value ≤ 0.05). Biological network analysis showed regulation of new pathways involving lipid and
78 organophosphate ester transport. Verification was performed for novel candidate biomarkers (CFHR5,
79 ILF2) in two independent cohorts. These proteins were elevated in both TB and other respiratory
80 diseases (ORD). Receiver-operating-characteristics analyses using a 5-protein panel (CFHR5, LRG1,
81 CRP, LBP and SAA1) exhibited discriminatory power in distinguishing between TB and ORD (AUC
82 =0.81).

83

84 **Conclusions**

85 We report the most comprehensive TB plasma proteome to date, identifying numerous novel markers
86 with verification in two independent cohorts, which led to a 5-protein biosignature with potential to
87 improve TB diagnosis. With further development, these biomarkers have potential as a diagnostic
88 triage test.

89

90 **Funding**

91 Colombia: Colciencias. UK: Medical Research Council, Innovate UK, National Institute for Health
92 Research, Academy of Medical Sciences. Peru: Program for Advanced Research Capacities for AIDS.
93 South Africa: Wellcome Centre for Infectious Diseases Research.

94 **Introduction**

95 The tuberculosis (TB) pandemic continues relentlessly, killing more humans than any other infectious
96 disease, and progress is lagging behind other major diseases such as HIV and malaria (1). A
97 fundamental issue with controlling the global pandemic is the inadequacy of current diagnostic tests
98 for TB, which have multiple limitations such as insufficient sensitivity, high cost and reliance on
99 laboratory infrastructure (2, 3). The World Health Organisation has defined the characteristics of an
100 optimal TB diagnostic, including low-cost, use of a non-sputum sample, high sensitivity and
101 specificity, as well as stability at extremes of temperature and humidity, and may include both rule-in
102 and rule-out tests (4). However, development of a point-of-care test suitable for resource-limited
103 settings faces multiple challenges in the pathway from discovery to validation and implementation,
104 such as translation between platforms, application across different populations and the disease
105 heterogeneity of TB.

106 Proteins have been proposed as viable diagnostic candidates given their phenotypic relevance and
107 stability under specified conditions. Blood plasma contains a wide spectrum of proteins that may
108 serve as biological signatures of physiological status during homeostasis or its perturbation (5). For
109 example, the plasma matrix encompasses tissue leakage proteins, thus providing systemic and
110 organotypic insight about specific immunopathologic features such as lung tissue destruction relevant
111 to active TB (6, 7). Furthermore, plasma protein signatures are highly amenable for translation to
112 rapid test devices and this technology is rapidly evolving, including colorimetric gold nanoparticles
113 on paper-based devices, label-free biosensors and nanofluidic disposable chips (8, 9). Extensive
114 proteomic discovery research has been conducted in TB. Although this has identified novel diagnostic
115 markers for the active disease (10-16) and progression from latent disease (17), an optimal diagnostic
116 panel has yet to be defined (18). Other analytes, such as matrix degradation products, have been found
117 by a hypothesis-driven approach (7, 19), but conversely have not been identified by mass
118 spectrometry based strategies. This implies that improved discovery strategies are required to
119 increase the plasma proteome coverage, thus improving the prospect in capturing novel protein
120 markers with potential clinical utility.

121 Current limitations to mainstream serum or plasma proteomics pipelines partly stems from the
122 predominance in protein mass (>95%) of the top 20 most abundant proteins. These high abundant
123 proteins either mask the presence of or are non-covalently bound to lower abundant proteins with
124 potential clinical relevance. In an effort to overcome this limitation, an initial serum/plasma depletion
125 step to remove such high abundance proteins is typically employed prior to the mass spectrometry-
126 based analysis. This plasma proteome analysis strategy has been used in samples from patients with
127 TB (11, 13, 20-24). However, this approach will result in the inadvertent loss of a wide spectrum of
128 physiologically important proteins including those typically encountered in lipid microvesicles such

129 as exosomes, proteases and their cleavage products, and native peptides such as hormones (25, 26).
130 Consequently, an alternative methodological approach has been optimised, wherein the entire
131 repertoire of secreted and exosome-enriched proteins, including the high-abundant carrier and
132 immunoglobulin proteins, and their derivative proteotypic peptides are subjected to multi-dimensional
133 or orthogonal liquid chromatographic separation combined with high-definition mass spectrometry
134 analysis (Figure 1) (27-29). The present study optimised critical aspects of this methodology to
135 generate a highly comprehensive plasma proteome coverage to capture novel biomarkers in active
136 TB.

137

138 **Results**

139 **Proteomic analysis of non-depleted plasma identifies numerous modulated proteins in TB**

140 For each sample, four protein segments were generated from plasma by high-performance size
141 exclusion chromatography (HP-SEC) partitioning under highly chaotropic mobile phase conditions.
142 Then, each HP-SEC segment was subjected to downstream 2D LC-MS analysis to achieve a
143 comprehensive profile of the non-depleted plasma proteome (Supplementary figure 1). The HP-SEC
144 fractionation traces were highly reproducible (Supplementary figure 2). All four segments from one
145 set of seven plasma samples, comprising four samples from active TB patients, three from healthy
146 donors and one master pool (Set A, Supplementary table 1) were profiled to generate an exploratory
147 in-depth plasma proteome in TB (Figure 2). Samples included in this first stage were obtained from
148 donors from South Africa and Peru. The samples from Peru were collected prospectively to match
149 BMI and age of the donors from South Africa (Supplementary table 1). A total of 5022 non-
150 redundant proteins (peptide FDR $\leq 5\%$) were identified from which 3577 were quantified across all 8
151 samples. Only quantified proteins profiled at a strict 1% FDR were subjected to further bioinformatic
152 and statistical analysis. Proteins profiled in the sub-proteome contained in segment four presented the
153 widest distribution of molecular weight ranging from 5KDa to 630KDa (Figure 2A). A total of 53% of
154 the quantified proteins had reported circulating levels in the literature or the human plasma dataset
155 (integrated) from the reference PaxDb^{4.1} protein abundance database (30, 31). Based on these reported
156 circulating levels, the plasma proteomic profile covered abundance levels of 11 orders of magnitude
157 (Figure 2B) representing classical, tissue-leakage and signalling proteins (32). Furthermore, 905
158 profiled proteins were annotated as exosome-, microvesicle- or microparticle-derived proteins (33).
159 The actual abundance dynamic range is expected to be larger, as the LC-MS signal intensity observed
160 for many proteins with unknown native concentration levels is below that of proteins with the lowest
161 previously reported concentrations.

162 Principal component analysis (PCA) demonstrated that this plasma proteome could distinguish
163 between controls and active TB patients (Figure 2C). Overall, 62% of the variance was explained by

164 PC1 and PC2. The master pool was a combination of plasma from healthy control and TB patients,
165 and clustered in the centre of control and diseased groups. One TB patient profile (reporter ion at m/z
166 121) clustered with the control group, and review of the clinical data showed that although the
167 *Mycobacterium tuberculosis* (*Mtb*) sputum culture was positive, the plasma CRP level was normal
168 and the chest X-ray showed no consolidation, suggesting very early disease, in contrast to all other
169 patients who had lung inflammation. This demonstrates that proteomic profiling reflects disease
170 heterogeneity that is consistent with clinical features.

171

172 Similar to the PCA, Spearman correlation showed clustering between TB and controls, but with
173 reporter ion at m/z 121 clustering with controls (Figure 2D). Defined patterns of protein expression
174 associated to the disease status were observed in two clusters. Cluster blue includes proteins with
175 reduced abundance in the TB group while cluster magenta contains proteins with increased abundance
176 in the TB group. Gene ontology enrichment analysis indicated regulation of immune response to
177 external stimulus mainly through the innate response, including the complement pathway and
178 phagocytosis.

179

180 Recently, analytical models such as Linear Models for Microarray Data (LIMMA) have been
181 translated to proteomic datasets from large-scale gene expression data (34). Empirical Bayes
182 approaches have been proven to be particularly powerful with small sample numbers by using the full
183 datasets to reduce observed sample variances towards an estimate while allowing for variance
184 distribution (35-37). This statistical approach results in a more realistic distribution of biological
185 variances compared to other methods. Furthermore, LIMMA offered the best statistical properties
186 when compared to generalised linear model (GLM) and mixed models in the context of multiplexed
187 isobaric quantitative proteomics (34). Statistical assessment of differential expression showed 119
188 proteins significantly modulated (nominal $p\text{-value} \leq 0.05$) (Supplementary table 2). However, after
189 FDR correction for multiple comparisons, no significant differences were retained. Therefore, we
190 increased the sample size to identify TB biomarker proteins with high confidence.

191

192 **In-depth analysis of segment 4 identifies multiple new TB biomarkers**

193 Robust statistics are crucial at the discovery stage of biomarker identification to increase chances of
194 later validation. Considering that HP-SEC segment 4 captured the most diverse range of protein
195 molecular weight (Figure 2A), we interrogated this segment further to increase statistical power.
196 Reported simulations for statistical power in proteomic studies, including power curves estimated for
197 iTRAQ relative ratios (37), predict that a minimum of 9 biological or clinical replicate samples per
198 group are needed to achieve a statistical power of 0.9 when an effect size of 1.5 is considered (37, 38).
199 Therefore, 10 healthy control and 11 active TB plasma samples were analysed. These samples were

200 randomly allocated into three iTRAQ experiments (Supplementary figure 2A) and analysed as three
201 independent mass spectrometry (MS) experiments. A maximum of 1,248 proteins were quantified at
202 1% FDR and 426 proteins were common to the three MS runs (Supplementary figure 2B). The overall
203 relative protein expression variation was evaluated using the common proteins profiled across the
204 three independent iTRAQ experiments. The relative standard deviation (RSD) was >25 , which
205 accounts for the combined technical and biological variation (Supplementary figure 2C). Using an
206 alternative approach to estimate the mean-variance relationship in the data, the locally weighted
207 regression (LOWESS) trend was calculated using the function *voom* (39) from the limma R package,
208 analysing the same group of proteins (Supplementary figure 2D). The square-root-standard-deviation,
209 $\sqrt{\text{SD}}$, was >1.4 and the LOWESS *voom* trend indicates a degree of heteroscedasticity in the data,
210 where greater \log_2 relative expression values were related to higher variation. The range of RSD and
211 $\sqrt{\text{SD}}$ estimated across these three multidimensional experiments indicates a good overall method
212 performance.

213 The datasets generated were inspected to evaluate batch effects and data distribution. Sixty percent of
214 the variance was explained by the batch (Supplementary figure 4A). The group effect was then
215 distinguishable when considering dimensions PC2 and PC3 ($\sim 17\%$ variance, Supplementary figure
216 4B). Batch effect correction was performed using normalisation to the master pool or by ComBat (40)
217 (Supplementary figure 4C and 4D, respectively), with ComBat providing the best reduction of batch
218 effects. Statistical assessment of significant differential protein expression using LIMMA revealed
219 136 proteins significantly modulated ($q\text{-value} \leq 0.05$; Supplementary table 3). Proteins with
220 significantly increased and reduced abundance were identified in patients with active TB infection
221 (Figure 3A). In addition to the identification of proteins known to be regulated during the course of
222 the active TB immunopathology, such as CRP, SAA, S100A8, RBP4, MMP14 and diverse
223 apolipoproteins, completely novel proteins were found, such as DLG4, SFTPB, CFHR5 and SPP2.

224 Further data mining of the output from segment 4 was performed to interpret biologically relevant
225 patterns in pulmonary TB. Weighted gene co-expression network analysis (WGCNA) (41) was used
226 to explore relationships between clusters of highly correlated proteins (colour modules) and specific
227 sample traits. Technical and biological variables of batch, smoking history and ethnicity were
228 evaluated as possible confounders in the data using hierarchical clustering. The resulting dendrogram
229 demonstrated that disease status was the primary determinant of sample clustering (Supplementary
230 figure 5A). In order to select highly interconnected proteins exhibiting the strongest correlation with
231 the disease status, detection of modules was performed (Supplementary figure 5B). The dendrogram
232 of the topological overlap matrix (TOM) representing clusters of highly interconnected proteins with
233 assigned colour modules and association to particular traits, demonstrated that the protein module
234 turquoise was strongly associated to disease status (Figure 3B, Z score = -0.87 ; $p\text{-value} = 2 \times 10^{-07}$).
235 One hundred and eighty nine proteins were contained in the turquoise module (Supplementary table

236 4) of which 129 (65.8%) were common to the differentially expressed proteins defined with LIMMA
237 (7 protein unique to LIMMA and 60 unique to WGCNA). Gene ontology enrichment was performed
238 using the package clusterProfiler (42) on the turquoise module and demonstrated that proteins profiled
239 were mainly associated to a variety of intracellular and secretory vesicles, extracellular matrix, blood
240 microparticles and lipoprotein particles (Figure 3C). Analysis revealed four main hubs for the top 20
241 biological processes: inflammatory/acute-phase response, exocytosis/vesicle-mediated transport, lipid
242 transport and proteolysis (Figure 4).

243

244 To generate the most robust list of candidates for validation, we identified proteins in common
245 between the module turquoise derived from WGCNA and significant by empirical Bayes moderated t-
246 statistics in LIMMA, thereby combining co-expression analytical approaches and t-statistics.

247 Combining the approaches, we identified 26 common proteins with increased and 20 proteins with
248 reduced abundance, with a high predicted significance (full list, Supplementary table 5; \log_2 Fold
249 change $\geq |0.5|$; WGCNA: Z score $\geq |0.65|$ and p -value ≤ 0.05 ; LIMMA: q -value ≤ 0.05). This highly
250 stringent approach is likely to omit numerous other differentially regulated proteins, but maximises
251 the chance of subsequent validation for diagnostic use. Proteins in this list are associated to a wide
252 range of biological processes, including acute inflammatory response, defence response to bacterium,
253 lipid localisation, cell adhesion and regulation of peptidase activity (Figure 5).

254

255 **Host plasma proteins exhibit increased abundance in TB and other respiratory diseases**

256 Circulating levels of five proteins amongst the top 15 proteins with increased expression levels (Table
257 1) were subjected to independent verification with ELISA or luminex array. C-reactive protein (CRP)
258 and serum amyloid A1 (SAA1) were included in the verification panel as these are considered
259 established major acute-phase effectors and are expected to increase in individuals with pulmonary
260 TB. Lipopolysaccharide binding protein (LBP) and leucine rich alpha-2-glycoprotein 1 (LRG1) have
261 been described in other proteomic TB profiles (11, 43, 44); therefore, the expression of these proteins
262 on specific cohorts may add valuable information for the design of a multi-marker panel. Newly
263 identified proteins from our analysis such as complement factor H related 5 (CFHR5) were
264 additionally selected for verification. Proteins closely biologically associated to the selected proteins
265 were excluded for further verification such as serum amyloid A2 (SAA2), since independency is
266 recognised to benefit performance of multi-marker panels. In addition to these selected candidates, the
267 seven most consistently divergently regulated proteins, analysed by fold change, derived from the
268 profile of HP-SEC segments 1 to 3 were included, RPGRIP1L, FGL1, COMP, KCNN2, TNFSF11,
269 LTN and ILF2, to compare verification efficiency between the smaller and larger discovery groups.

270 First, we studied a UK-recruited independent cohort of mixed ethnicity from the Multifunctional
271 Integrated Microsystem for rapid point-of-care TB Identification (MIMIC) study, for verification of

272 selected candidates. CFHR5, LRG1, LBP, SAA1 and CRP showed significantly increased levels of
273 expression in active TB patients when compared to healthy controls or latently infected individuals
274 (Figure 6A-E). Evaluation of the markers selected from the initial discovery experiment on seven
275 samples showed that RPGRIP1L, FGL1, COMP, KCNN2 and TNFSF11 failed verification
276 (Supplementary figure 6). LTN (Supplementary figure 7, *p-value* = 0.04) abundance was significantly
277 higher in TB patients. Additionally, ILF2, identified from segment 3 analysis, showed elevated
278 abundance in latent TB and active TB patients compared to healthy donors (Figure 6F, *p-value* =
279 0.0005). Consequently, two out of seven proteins successfully verified from the smaller discovery
280 group, whereas all were verified from the larger discovery group. In addition to being elevated in
281 TB, patients with other respiratory diseases (ORD) also exhibited elevated abundance in all verified
282 markers (Figure 6A-F).

283 Diagnostic performance of individual and combined verified markers was evaluated using Receiver
284 Operator Characteristic (ROC) curves. ROC curves were generated based on two different
285 comparisons; circulating level of markers in active TB patients vs. healthy controls (Figure 7A) and
286 active TB patients vs. ORD patients (Figure 7B). In both cases, the best performance was achieved by
287 combining the 5 markers (CFHR5, LRG1, LBP, SAA1 and CRP). The area under the curve (AUC)
288 was 0.93 (95% confidence interval: 0.89-1.00, *p-value* ≤ 0.001) for TB vs. healthy controls and 0.81
289 (95% confidence interval: 0.68-0.94, *p-value* = 0.001) for TB vs. ORD, thus demonstrating that only
290 the combination of markers allowed the discrimination of active TB from healthy controls and ORD.
291 Although ILF2 abundance was significantly upregulated in the active TB and ORD patients from this
292 cohort (Figure 6F), it did not contribute towards a better diagnostic performance of the panel.

293 We then further verified the biomarkers in a South African cohort, which included HIV-uninfected
294 and HIV-infected active TB and ORD patients. Again, the novel diagnostic marker CFHR5 exhibited
295 significant increased abundance in HIV-uninfected patients. In HIV co-infected patients, CFHR5 was
296 elevated compared to healthy controls, but not significantly different to healthy HIV-infected
297 individuals, although this group had limited numbers (Figure 8A). CFHR5 showed no significantly
298 increased abundance in ORD, irrespective of HIV status. Again, interpretation may be due to limited
299 sample numbers reducing statistical power. LBP and SAA1 both showed increased abundance in the
300 active TB group regardless of HIV status. This trend was observed relative to the ORD group HIV un-
301 and co-infected (Figure 8B-C). CRP showed increased abundance in TB compared to healthy controls
302 and ORD groups, irrespective of HIV status (cohort data previously published (7)). In this cohort,
303 ILF2 and LRG1 could not be measured due to sample exhaustion and were thus excluded from the
304 panel. A summary of the analytes tested in each cohort and verification results is presented as
305 Supplementary table 6.

306 ROC curves generated by comparing circulating levels of CFHR5, LBP, SAA1 and CRP in TB
307 patients vs. ORD in the HIV uninfected group showed that the best performance was achieved by
308 combining markers (Figure 9A, AUC 0.89 [95% confidence interval: 0.80-0.98, p -value ≤ 0.001]).
309 Similarly, in the context of HIV-associated TB, the combination panel performed best, and provided a
310 surprisingly high discrimination between active TB and ORD (Figure 9B, AUC 0.98 [95% confidence
311 interval: 0.94-1.00, p -value ≤ 0.001]). By contrast, the combination of markers did not improve the
312 diagnostic performance when the active TB group was analysed against the healthy controls relative
313 to analysis of CRP alone (Supplementary figure 8). Finally, we evaluated whether our 4-protein panel
314 correlated to sputum mycobacterial load in the South African cohort. Mean Z-scores were calculated
315 from CFHR5, LBP, SAA1 and CRP levels in TB patients (HIV negative) and compared to the
316 bacterial burden in sputum. A significant positive correlation was observed (Spearman coefficient $r =$
317 0.37 , p -value = 0.03).

318

319 **Discussion**

320 We applied a unique non-depletion based quantitative proteomics method (q3D LC-MS) to generate
321 the most comprehensive TB plasma proteome to date. Statistical power was increased by studying
322 one HP-SEC segment in additional patients and combined WGCNA and LIMMA analysis
323 approaches, identified numerous novel host biomarkers with high confidence. We verified a subset of
324 biomarkers in two separate cohorts, with high success rate. Diagnostic accuracy for TB was
325 maximised by use of a multi-marker panel. These markers are frequently also increased in other
326 respiratory conditions and therefore host biomarkers are likely to be of greatest use in a rule-out panel.

327 Translation of novel biomarkers for clinical utility is challenging, involving a stepwise process where
328 most candidates fail to reach the bedside. Verification of new candidates typically relies on antibody-
329 based assays, requiring change of platform from mass spectrometry to immunoassays prior to field-
330 testing, and this is frequently a point of failure. We completed this transition for three new analytes,
331 thereby supporting the robustness of the approach. Validation will require quantification of the
332 additional 15 entirely new biomarkers in the top candidates (Figure 5, Supplementary table 5)
333 identified by the combined WGCNA and LIMMA approaches, and inter-laboratory collaboration
334 across large cohorts from multi-centre biobanks, including analysis of how biomarkers relate to
335 disease severity and change over time.

336 Plasma is a complex matrix to analyse, and high-abundant protein depletion is the most common
337 strategy to address this complexity (5, 27-29, 45, 46). However, depletion may inadvertently co-
338 remove important analytes non-covalently bound to high abundance proteins (26). In this study,
339 sample preparation was principally based on the use of orthogonal chromatographic hyper-
340 fractionation instead of depletion. Such a strategy entailed the dissolution of 120 μ L neat plasma with
341 7M guanidine/10% methanol that stabilised the protein content and was subjected to HP-SEC
342 separation as part of the hyper-fractionation pipeline. The use of multi-dimensional liquid
343 chromatographic approaches as part of the isobaric quantitative proteomics pipeline has gained
344 increasing prominence in translational research studies (47). Such approaches compensate for the
345 complexity of biological specimens in capturing and analysing very low abundant proteins of clinical
346 significance. Furthermore, they are amenable to laboratory automation and scale-up, thus improving
347 analysis throughput, accuracy and precision (47, 48). In line with this, the collective attributes of the
348 present study method facilitated the analysis of proteins encompassed in blood microparticles, such as
349 exosomes and other lipid vesicles (27, 28), along with protease derived cleavage proteins and soluble
350 proteins. The efficacy of our approach was demonstrated by the profiling of over 5,000 proteins from
351 only 120 μ L plasma per patient, compared to the identification of a maximum of 800 proteins in
352 similar TB discovery studies from larger volumes of plasma (16, 20, 49). Most importantly, however,
353 the deep proteome coverage achieved also encoded for a wide spectrum of biological and disease

354 specific pathways and networks of physiological relevance to TB. Encompassed in these pathways
355 and networks were many novel proteins of potential clinical significance.

356 Analysis of the entire proteome from HP-SEC segments 1 to 4 using seven samples was
357 underpowered for biomarker discovery, with only two out of seven candidates subsequently validating
358 on a larger cohort. Therefore, detailed profiling was focused on the sub-proteome segment 4, which is
359 primarily enriched for low-molecular weight proteins and protein degradation products, recapitulating
360 multiple biological processes (28, 29, 50-52). In-depth profiling of this segment from 10 healthy
361 controls and 11 pulmonary TB patients provided much greater statistical power, consistent with
362 mathematical estimations (38). The high-dimensional data produced from isobaric labelling-based
363 relative quantification (iTRAQ or TMT) poses bioinformatic processing challenges (34). Small
364 sample sizes, incomplete datasets and batch effects across experiments create difficulties in the
365 effective detection of protein abundance changes (35). Batch effects are particularly relevant to
366 multiplexing of iTRAQ experiments. In our study, Combat correction performed better than the most
367 common strategy of normalizing to a common reference sample (Supplementary figure 4).
368 Complementary analysis using LIMMA and WGCNA on the adjusted data resulted in a powerful
369 approach producing a set of robust markers for verification (Supplementary table 5), with three out of
370 three tested proteins successfully converting to an immunoassay platform, compared to two out of
371 seven from the smaller sample set (Set A profile). Thus, this methodology led to the identification
372 and independent verification of known and novel candidate biomarkers of TB infection.

373 WGCNA identified one co-expression module as strongly associated to the group TB (turquoise
374 module, $p\text{-value} = 2 \times 10^{-07}$) containing 189 proteins. Ninety-five percent of the differentially expressed
375 proteins identified with LIMMA were common to this module, showing excellent concordance
376 between analytical strategies. Notably, over 60% of the co-expressed proteins showed decreased
377 abundance in the active TB group, suggesting that studying these proteins may provide additional
378 insight into disease process in TB, and analysis should not purely focus on proteins of increased
379 abundance. Gene ontology enrichment of module turquoise revealed regulation of biological
380 processes associated to responses to external stimulus ($q\text{ value} = 2 \times 10^{-03}$) encompassing acute-
381 phase/inflammation ($q\text{ value} = 5.2 \times 10^{-06}$) and humoral responses ($q\text{ value} = 9.2 \times 10^{-05}$). Within this
382 module, CRP, LBP, SAA1, SAA2, S100A8, S100A9, SERPINA3 and HP are involved in the
383 activation of the acute-phase and inflammatory response, which are well described in TB (20, 53, 54).
384 This concordance supports the overall validity of our methodology.

385 Connected to the acute-phase hubs, proteolysis ($q\text{ value} = 1.1 \times 10^{-06}$) and lipid transport and localisation
386 ($q\text{ value} = 1.4 \times 10^{-05}$) were significantly enriched. Proteolysis is consistent with the extensive pulmonary
387 destruction that occurs in human TB (55). Among the proteins with increased abundance in this hub,
388 ECM1 was previously reported elevated in saliva of TB patients (56), MMP-14 is expressed in TB

389 granulomas (57) and PSMB8 may be part of the regulatory cascade of the blood transcriptome of TB
390 patients (58). Among the proteins found with decreased abundance, TIMP2 is an inhibitor of matrix
391 metalloproteinases, and so reduced levels may increase matrix degradation (55). Lipid metabolism was
392 another major signal expressed, and the role of lipids and cholesterol in TB immunopathology remains
393 poorly characterised. Cholesterol uptake and catabolism are central for maintenance of the pathogen in
394 the host and contribute to pathogenesis and virulence (59). However, the low circulating lipid profiles
395 in pulmonary TB patients may be a consequence of the disease or may have wider biological
396 implications. Apolipoproteins are associated to lipid transport and form lipoprotein particles such as
397 HDL, LDL and VLDL. Serum HDL-C concentrations negatively correlate with the radiological extent
398 of disease and smear positivity in pulmonary TB (60). Decreased circulating concentrations of
399 apolipoproteins are consistently reported in different serum/plasma proteomic profiles for pulmonary
400 TB (11-13), in agreement with our findings. Further data mining of these biological processes may
401 identify host-directed therapy targets.

402 To verify newly-identified biomarkers, well-characterised TB cohorts with complementary profiles and
403 from geographically diverse populations are required (4). We studied two different cohorts for
404 verification, one recruited in the UK and one in South Africa. From the subset of proteins analysed by
405 ELISA or luminex, seven proteins were successfully validated. LBP, CFHR5, CRP and SAA were
406 consistently increased in TB cases in both cohorts. Statistically significant differences were observed
407 despite the wide inter-individual variation in biomarker concentrations, which is expected from clinical
408 TB which has a wide spectrum of disease severity. ILF2 was only verified in the MIMIC cohort due to
409 sample exhaustion, while LTN and LRG were only evaluated in the South African cohort. CFHR5
410 (complement factor H-related protein 5), ILF2 (Interleukin Enhancer Binding Factor 2) and LTN (E3
411 ubiquitin-protein ligase listerin) are novel protein candidate biomarkers for TB identified by the
412 discovery phase and all were successfully verified. Consistent with our findings, a recent report
413 identified ILF2 as a potential biomarker in paediatric TB by bioinformatic mining of gene expression
414 datasets (61).

415 Evaluation of the performance of a subset of markers indicated that combination rather than
416 individual markers provided a better diagnostic ability. In the UK-based cohort, ROC analysis
417 demonstrated that the multi-marker panel comprising CFHR5, LRG1, CRP, LBP and SAA1
418 performed well in receiver operator curve analysis against healthy controls (AUC = 0.93). However,
419 the discriminatory power was reduced but still significant when compared against other respiratory
420 diseases (AUC = 0.81). Clinically, differentiation against other respiratory conditions is the key
421 comparator for TB diagnosis. Host biomarkers are often limited by lack of specificity and our findings
422 reinforce the importance of choosing correct control groups for verification analysis (18). In the South
423 African cohort including patients with and without HIV infection, the multi-marker panel comprising
424 LBP, CFHR5, CRP and SAA yielded its best performance when TB patients were compared to other

425 respiratory diseases (AUC=0.98). This is an important finding from a clinical perspective, as
426 diagnosing TB in HIV-infected patients is generally more challenging than in non-
427 immunocompromised individuals (2). Furthermore, performance of our panel in both cohorts (UK and
428 South Africa) comparing ATBI to ORD groups was similar to a different recently validated host
429 response signature (IL6, IL8, IL18 and VEGF, AUC=0.80) (62). This suggests our preliminary
430 signature can be further refined by testing of remaining highly significant candidates that have not yet
431 been studied. The primary difference between the groups is that the UK cohort were hospitalised
432 patients, whereas the South African cohort were outpatients, and therefore the better performance in
433 South Africa may reflect the fact the patients were less unwell. For utility of a point-of-care test,
434 outpatients with respiratory symptoms will be the primary target group.

435 Significant efforts have been directed to define an optimal plasma protein biosignature for active TB
436 and recently, extensive testing of candidate proteins identified by predefined discovery panels, such
437 those measured with luminex, have shown that multi-component or multi-factorial signatures could
438 give a greater performance than immunological markers despite the heterogeneity of clinical
439 presentation (62, 63). Inclusion of novel markers that represent the biological diversity of the host
440 response to the *Mtb* infection in diagnostic panels may be crucial to achieve the analytical
441 performance required to translate to effective point-of-care devices. From our top list of 46 proteins
442 identified by both LIMMA and WGCNA from the discovery phase (Supplementary table 5), 21
443 proteins are entirely novel candidates and involved in a wide range of biological processes.
444 Consequently, verification and integration with known markers may improve the performance of the
445 existing signatures. This list recapitulated several potential diagnostic biomarkers identified in a range
446 of reported plasma proteomic TB signatures (11, 13, 14, 20, 44, 64), including one signature for TB
447 progression (17), one for cured pulmonary tuberculosis (21) and one for multidrug-resistant TB (65),
448 demonstrating the ability of our proteomic and bioinformatic approach to detect proteins associated to
449 the disease status, independent of differences in discovery platforms or patient cohorts. However,
450 further verification of all the newly reported candidates that we identify is required to refine the
451 current panel.

452 Translation of such markers to point-of-care tests with adequate performance will require the
453 development of multiplex lateral flow assays, and such platforms are currently emerging (66, 67),
454 though will require careful development. Any assay used as a rule-out test would need population-
455 based studies to confirm the specificity against standard current clinical practice and emerging blood
456 protein-based signatures. Due to the overlap between TB and other respiratory conditions, the host
457 biomarkers identified are potentially best utilised as a rule-out triage test prior to performing more
458 specific and expensive rule-in tests (68). In the future, analysis of other proteins that are differentially
459 abundant will become increasingly achievable, given the continuous advancements of LC-MS

460 methods in terms of throughput and analytical confidence. When combined with machine learning
461 approaches, LC-MS based assays may transform specificity and sensitivity in the diagnosis of TB.

462 In summary, we developed a non-depletion based proteomic methodology to deeply profile plasma
463 and identify novel biomarkers. We present a unique statistical and bioinformatic pipeline for
464 discovery and selection of candidates for verification that utilises both statistical significance and also
465 correlation of expression patterns to clinical traits. We report numerous novel analytes, with potential
466 to be translated for clinical utility. We have verified a subset of biomarkers from segment 4 by
467 independent antibody-based assays to generate a preliminary diagnostic panel, and similar
468 interrogation of segments 1 to 3 is likely to generate further novel biomarkers. Taken together,
469 developing these host biomarkers into a multiplex lateral flow assay has potential for a near-patient
470 TB rule-out test that fulfils the WHO product characteristics. Such an assay could be a powerful tool
471 to address the global TB pandemic.

472

473 **Methods**

474 **Study Participants**

475 This study included participants from three different cohorts. The participants from the South African
476 cohort were recruited at Ubuntu HIV/TB clinic in Cape Town and were black-African ethnicity from
477 June 2012 to February 2014. Written informed consent was obtained, HIV testing was offered, and
478 chest radiographs were performed as per routine practice. The diagnosis of active TB was based on
479 sputum smear or culture positivity, Gene Xpert results (where available) and chest X-ray examination.
480 For the control group, all sputum samples were smear and culture negative for acid-fast bacilli (AFB).
481 Plasma samples from this cohort were retrospectively selected from a cohort collected and previously
482 described (7). Participants from this cross-sectional study were categorised into six groups: i) HIV-
483 uninfected patients without active TB infection (HIV- ATBI -); ii) HIV-uninfected patients with
484 active TB infection (HIV- ATBI +); iii) HIV-uninfected patients without active TB but with
485 symptoms attributable to other respiratory infectious disease (HIV- ORD); iv) HIV-infected without
486 active TB infection (HIV+ ATBI-); v) HIV-infected with active TB infection (HIV+ ATBI+) and vi)
487 HIV-uninfected patients without active TB but with symptoms attributable to other respiratory disease
488 (HIV+ ORD). Microbiological confirmation of the infectious agent was not available for the HIV-
489 /HIV+ ORD groups due to limitations in local diagnostic capability. A randomly selected subset of
490 11 plasma samples from male participants belonging to the groups HIV- ATBI - and HIV- ATBI +
491 was used for discovery (Supplementary table 1). A larger set of 203 samples from all six groups and
492 including those used for discovery constituted the South African verification cohort and demographic
493 description of this group has been previously reported with CONSORT diagram (7).

494 Participants from the Peruvian discovery cohort were prospectively recruited at clinics in Lima, Peru
495 to match demographic features such as gender, age and BMI of participants from the South Africa
496 cohort. Recruitment was conducted during 2015. The diagnosis of active TB was based on a TB
497 symptom questionnaire, sputum smear positivity, culture positivity using microscopic-observation
498 drug-susceptibility (MODS) culture and chest X-ray. Healthy control individuals were Quantiferon
499 negative. In total, 10 samples from this cohort were selected for the discovery stage of this study
500 (Supplementary table 1).

501 A second independent cohort was included for verification of proteomic candidates comprising a
502 subset of 118 participants from the Multifunctional Integrated Microsystem for rapid point-of-care TB
503 Identification (MIMIC) cross-sectional study conducted in the United Kingdom. Recruitment was
504 performed from June 2014 to February 2017. All the participants were HIV uninfected and four
505 categories were defined for this cohort: i) Healthy controls (HC), ii) Latent TB infection (LTBI); iii)
506 Active TB infection (ATBI) and iv) Other respiratory diseases (ORD). Healthy controls were
507 asymptomatic individuals without a history of previous active TB or TB contact and no evidence of

508 TB infection on routine screening tests (negative interferon-gamma release assay and/or tuberculin
509 skin test result). Participants with latent TB infection were defined based on a positive interferon-
510 gamma release assay and/or tuberculin skin test result, without evidence of active disease after clinical
511 evaluation. All active pulmonary TB cases were individuals with symptomatic respiratory infection
512 that were microbiologically-confirmed to have TB based on any of the following criteria: sputum
513 smear positive, sputum culture positive for *Mtb*, or PCR test positive for *Mtb*. The control group other
514 respiratory diseases (ORD) were symptomatic individuals with microbiologically-confirmed
515 respiratory tract infection caused by a pathogen (viral or bacterial) other than *Mtb*, without a history of
516 previous active TB (Supplementary table 7). The microbiological composition of this group was 31%
517 influenza A/B, 15% *Streptococcus pneumoniae*, 8% respiratory syncytial virus, 8% *Staphylococcus*
518 *aureus*, 4% *Mycoplasma pneumoniae*, 4% *Human Metapneumovirus*, 4% H1N1 Influenza A, 4%
519 Methicillin-resistant *Staphylococcus aureus* and 22% unidentified organism.

520 **Plasma processing**

521 Venous blood was collected in sodium citrate vacutainer tubes and plasma prepared according to
522 standard operating procedures at the site of recruitment and stored at -80°C. Aliquots of 120µl of
523 plasma were liquid-fixed with 380µl of 7M guanidine hydrochloride and 10% methanol and stored at
524 -20°C until size exclusion chromatography fractionation was performed for the discovery stage.
525 Aliquots of 20µl of the individual samples available for discovery including control and active TB
526 groups was combined to generate a master pool aimed to control batch effects across different MS
527 experiments. All the plasma samples included in the verification stage were divided into 100µl
528 aliquots to reduce freeze-thaw cycles when received and stored at -80°C until analysis.

529 **Multidimensional plasma proteomic analysis**

530 *High-performance size exclusion chromatography (HP-SEC)*

531 A general overview of the plasma proteomic method is presented in Supplementary figure 1A. Plasma
532 samples used for discovery, including four aliquots of the master pool, were individually subjected to
533 HP-SEC pre-fractionation under optimised conditions of the method reported previously (28). Five
534 columns were serially connected: 2 Shodex KW-804 columns, 8.0mm I.D. x 300mm; one Shodex
535 KW-802.5 column, 8.0mm I.D. x 300mm; and 2 Shodex KW-804 columns, operated at 45°C and
536 1.5mL/min under isocratic elution with 6M guanidine hydrochloride and 10% methanol. Four protein
537 HP-SEC segments were collected in a peak-dependent fashion detected at 280nm and then stored at
538 -20°C until further analysis. HP-SEC separations are presented in Supplementary figure 2A-E. The
539 BEH450 SEC protein standard Mix (Waters, UK) and an aliquot of one control plasma sample were
540 run for day-to-day quality control of the separation variation (Supplementary figure 2F). Variation of
541 retention times was within 2SD for all samples excepting one (Supplementary figure 2G). Protein

542 segments were dialysis-purified using 3KDa MWCO Slide-A-Lyzer cassettes according to
543 manufacturer's specifications (Thermo Fisher, Hemel Hempstead, UK) with exchanges of four
544 volumes of 4L of ultrapure water every 12h intervals in a cold room environment (4°C). The resulting
545 dialysates were completely lyophilised using the Edwards Modulyo EF4-174 freeze dryer and Thermo
546 Savant Micro Modulyo-115 benchtop freeze dryer. Protein extracts were stored at -80°C under argon
547 atmosphere.

548 *Trypsin digestion*

549 Total protein lyophilised extracts obtained from each HP-SEC segment were reconstituted with 0.5M
550 TEAB (triethylammonium bicarbonate) and 0.05% SDS (sodium dodecyl sulfate) and sonicated on ice.
551 Protein extracts were then centrifuged for 10 minutes at 16000xg and 4°C and protein content in the
552 supernatants was estimated using the Nanodrop ND-1000 spectrophotometer (Thermo Fisher
553 Scientific, Wilmington, USA) using the A280 program. 120µg of protein volume-adjusted were
554 reduced with 2µL of TCEP (50mM tris-2-carboxymethyl phosphine) and incubated for 1h at 60°C.
555 Reduced samples were then alkylated using 1µL of MMTS (200mM methylmethane thiosulphonate)
556 and incubated 10 minutes at room temperature. Digestion was conducted to a ratio of 1:40
557 enzyme/substrate with trypsin MS grade (Pierce, Thermo Fisher Scientific, UK) overnight for 16h at
558 37°C in dark.

559 *Stable isotope labelling*

560 iTRAQ 8-plex tags were equilibrated at room temperature and isopropanol was added accordingly to
561 ensure >60% organic phase during labelling. Each tag was added to the appropriate trypsinised
562 sample, then the labelling reaction was conducted for 2h at room temperature. The reaction was
563 stopped with 8µL of 5% ammonium hydroxylamine. Samples were dried and stored at -20°C until
564 chromatographic separation. The master pool was labelled using the tag 113 and the samples were
565 allocated randomly to the remaining tags as presented in the Supplementary figure 2A.

566 *Offline alkaline RP-HPLC peptide fractionation*

567 Offline peptide fractionation was based on high pH (0.08% v/v NH₄OH) reverse phase (RP)
568 chromatography using the Kromasil C₄ column (3.5µm, 2.1mm x 150mm) and on the Shimadzu
569 HPLC system previously described in the HP-SEC section. iTRAQ labelled tryptic peptides were
570 analytically reconstituted and pooled together with 100µL of mobile phase, centrifuged at 16000xg at
571 room temperature for 10 minutes. Supernatant was injected and separated at a flow rate 0.30mL/min
572 and 30°C. The fractions were collected in a peak-dependent fashion detected at 215nm. Peptide
573 fractions were dried at room temperature with a speedvac concentrator for 4–5h and stored at -20°C
574 until LC-MS analysis. Highly hydrophilic and hydrophobic fractions from the extreme regions of the

575 chromatographic traces were pooled and further cleaned using Gracepure SPE C18-AQ 100mg/1mL
576 cartridges (Grace, Columbia, USA).

577 *LC-MS analysis*

578 The LC–MS experiments were performed on the Dionex Ultimate 3000 UHPLC system coupled to
579 the high resolution nano-ESI-LTQ-Velos Pro Orbitrap-Elite mass spectrometer (Thermo Fisher
580 Scientific). HCD and CID fragmentation for each of the collected fractions was performed. For the
581 analytical separation the AcclaimPepMap RSLC, 75µm × 25 cm, nanoViper, C18, 2µm particle
582 column (Thermo Fisher Scientific) with trap cartridge retrofitted to a PicoTip emitter (FS360-20-10-
583 D-20-C7) was used for multistep gradient elution. MS characterization of eluting peptides was
584 conducted between 380 and 1500 m/z. The top ten +2 and +3 precursor ions were further
585 characterised by tandem MS. Full MS scans and MS/MS scans were acquired at a resolution of
586 30000FWMH (complete plasma proteome) or 60000FWMH (detailed analysis segment 4) for profile-
587 mode and 15000FWMH for centroid-mode, respectively, with the lock mass option enabled for the
588 445.120025m/z ion (DMSO). Data were acquired using Xcalibur software (Thermo Fisher Scientific).
589 Conditions for ionisation, CID and HCD fragmentation and ion detection were reported in a previous
590 work (28).

591 *MS data processing*

592 Target-decoy searching of raw mass spectra data was conducted with the Proteome Discoverer 1.4
593 software (Thermo Fisher Scientific). SequestHT was used for the target decoy search for tryptic
594 peptides, allowing two missed cleavages, 10ppm mass tolerance, and minimum peptide length of 6
595 amino acids. A maximum of 2 variable (3 equal) modifications; oxidation (M), deamidation (N, Q)
596 and phosphorylation (S, T, Y) were set as dynamic modifications. As static modifications were set:
597 iTRAQ8plex (Any N-terminal), Methylthio (C) and iTRAQ8plex (K). Fragment ion mass tolerance
598 was set to 0.02Da for the FT-acquired HCD spectra and 0.5Da for the IT-acquired CID spectra. FDR
599 was estimated with the Percolator (6.4Bit) and validation was based on *q-value* <0.01 for high
600 confidence or <0.05 for moderate confidence. All spectra were searched against a concatenated
601 FASTA file including the reviewed UniProtKB SwissProt human proteome and the reference
602 proteome (SwissProt and TrEMBL) for *Mtb* (strain ATCC 25618 / H37Rv), both retrieved on 04
603 August 2017. All peptide spectrum matches (PSM) of reporter ions and iTRAQ ratios were exported
604 to .txt at 1% FDR or 5% FDR peptide confidence and 50% co-isolation exclusion threshold. Protein
605 grouping was allowed and maximum parsimony principle was applied. Only unique peptides were
606 considered for quantification downstream analysis. Raw precursor ion intensities from unique
607 peptides were imported to R (version 3.3.1) and median-adjusted. Median-normalised peptide
608 intensities were log2-transformed and values were averaged to obtain the mean relative expression for

609 each protein. Only proteins with relative quantification reported in all the samples was included for
610 statistical analysis.

611 **ELISA and luminex assays**

612 Proteins selected for verification from the proteomic discovery experiments were measured in two
613 different cohorts using ELISA or luminex assays. ELISA measurements comprised candidates for
614 which there are commercially available kits, such as: RPGRI1L, FGL1, COMP, ILF2, KCNN2,
615 LTN1, LRG1 and SFTPB (2B Scientific Ltd, Upper Heyford, UK and Caltag Medsystems Ltd,
616 Buckingham, UK). One luminex multiplex assay was custom-made for analysis of LBP, COMP,
617 TNFSF11 and CFHR5 and two single-plexes for SAA1 and CRP (Protavio Ltd, Cambridge, UK). CV
618 for the ELISA assays was $\leq 12\%$ and for the luminex assays was $\leq 15\%$. Assays were performed
619 according to manufacturer's directions.

620 **ROC curves and AUC analysis**

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

627 **Statistics**

628 Differentially expressed proteins (DEPs) were determined using linear modelling LIMMA (69)
629 followed by FDR correction for multiple correction testing. WGCNA-based analysis was applied to
630 the datasets resulting from the detailed profile of segment 4 to interpret biologically relevant patterns
631 of protein expression in plasma of patients with pulmonary TB. The WGCNA R package was used to
632 explore the correlation relationships between clusters of highly correlated proteins (colour modules)
633 and specific sample traits. The batch effect was corrected in order to increase the analysis power with
634 ComBat (40). Networks of highly interconnected proteins were constructed using a soft-thresholding
635 power = 0.9 and modules were identified using a minimum module size of 15. Module significance
636 was calculated as a measurement of the correlation between biological traits, such as disease or group,
637 ethnicity and smoking status and the protein expression profiles. Visualisation tools available from
638 this package were used to identify modules strongly correlated to biologically relevant covariates.
639 Functional enrichment analysis was conducted using the option g:GOST available in the tool g:Profiler
640 (70). Only GO terms with an FDR adjusted *p-value* (cut-off 0.05) were considered. Significant GO

641 terms were summarised by removing redundant terms using the tool REVIGO (71). C-Net plots were
642 generated using the R package cluster profiler (72).

643 For ELISA and luminex measurements, differences between groups were analysed by Kruskal-Wallis
644 tests and using Dunn's multiple comparison correction. Data was analysed on Prism 8 (GraphPad, San
645 Diego, US). A *p-value* ≤ 0.05 was considered statistically significant. For the ROC analyses, the
646 nonparametric method was used to estimate the standard error of the area under the curve and the
647 confidence interval was set at 95%.

648

649 **Study approval**

650 All clinical studies were conducted according to declaration of Helsinki principles. All participants
651 gave written informed consent prior to inclusion in any of the clinical studies here included. The
652 South African cohort was recruited under the study approved by the University of Cape Town
653 Research Ethics Committee (HREC, REF 516/2011). The prospective enrolment of participants in the
654 Peruvian study was approved by the Universidad Peruana Cayetano Heredia Institutional Review
655 Board (SIDISI 65314). The MIMIC study was funded by the Technology Strategy Board UK /
656 Innovate UK and approved by the National Research Ethics Service Committee South Central (Ref 13
657 SC 0043). University of Southampton Ethics and Research Governance Online (ERGO) approval for
658 transporting samples to the United Kingdom was granted (17758).

659 **Author Contributions**

660 DGB was involved in the study design; performed the optimisation of the proteomic method and
661 conducted the plasma proteome profiling, analysed and integrated the data, the verification
662 experiments and wrote the majority of the manuscript. CoW wrote the R scripts used to normalise raw
663 peptides intensities, calculate protein expressions, and LIMMA analysis. NW recruited the South
664 African cohort and provided clinical annotation. MT recruited the MIMIC cohort and provided
665 clinical annotation. HS was involved in the experiments of verification using ELISA and luminex.
666 CUG recruited the Peruvian clinical cohort and provided clinical annotation. AM and JA provided
667 expertise in the plasma proteomic protocol. AV provided expert insight on the bioinformatic analysis
668 and the R scripts for WGCNA and ComBat. MB was involved in the validation experiments. RW,
669 SMJ and BM assisted with recruitment of patients to the cohorts. LT assisted in the luminex analysis.
670 CrW was involved in the study design and provided expertise on the bioinformatic pipeline design.
671 SG was involved in the study design and provided expertise and advice on the plasma proteomics
672 method and contributed to the manuscript writing process. PE was involved with the study design,
673 secured funding, and contributed to manuscript writing and edition.

674

675 **Acknowledgements**

676 This work was supported by Colciencias Scholarship 6171, Government of Colombia, and MRC
677 Global Challenges Research Fund MR/P023754/1, Confidence in Concept MC_PC16059 and
678 MR/R001065/1 and the Global-NAMRIP funding program. NFW was supported by Wellcome Trust
679 (094000) NIHR, Starter Grant for Clinical Lecturers (Academy of Medical Sciences UK, Wellcome,
680 Medical Research Council UK, British Heart Foundation, Arthritis Research UK, Royal College of
681 Physicians, and Diabetes UK) and British Infection Association. CU-G received support from the
682 Program for Advanced Research Capacities for AIDS in Peru (PARACAS) at Universidad Peruana
683 Cayetano Heredia (D43TW00976301) from Fogarty International Center at the U.S. National Institute
684 of Health (NIH). We are grateful to the Wellcome Centre for Infectious Diseases Research in Africa
685 clinical research team and to the participants, staff, and patients of Ubuntu Clinic and the Western
686 Cape Government: Health. PE is grateful of the support of the Southampton NIHR Biomedical
687 Research Centre. The MIMIC study, MT and SM-J were supported by a grant from the UK
688 Technology Strategy Board / Innovate UK (Grant no. 101556). MT was also supported by a Clinical
689 Lectureship by the National Institute for Health Research UK. The mass spectrometry proteomics data
690 have been deposited to the ProteomeXchange Consortium via the PRIDE (73) partner repository with
691 the dataset identifier PXD020212. The plasma proteomic discovery pipeline is currently patent-
692 pending.

693

694

695

696

697

698

699

700

701 **References**

- 702 1. World Health Organization. Geneva, Switzerland: WHO; 2015:79.
- 703 2. Walzl G, McNerney R, du Plessis N, Bates M, McHugh TD, Chegou NN, and Zumla A.
704 Tuberculosis: advances and challenges in development of new diagnostics and biomarkers.
705 *Lancet Infect Dis.* 2018;18(7):e199-e210.
- 706 3. Tebruegge M, Ritz N, Curtis N, and Shingadia D. Diagnostic Tests for Childhood
707 Tuberculosis: Past Imperfect, Present Tense and Future Perfect? *Pediatr Infect Dis J.*
708 2015;34(9):1014-9.
- 709 4. Kik SV, Denkinger CM, Casenghi M, Vadnais C, and Pai M. Tuberculosis diagnostics: which
710 target product profiles should be prioritised? *The European respiratory journal.*
711 2014;44(2):537-40.
- 712 5. Geyer PE, Holdt LM, Teupser D, and Mann M. Revisiting biomarker discovery by plasma
713 proteomics. *Mol Syst Biol.* 2017;13(9):942.
- 714 6. Urbanowski ME, Ihms EA, Bigelow K, Kubler A, Elkington PT, and Bishai WR. Repetitive
715 Aerosol Exposure Promotes Cavitory Tuberculosis and Enables Screening for Targeted
716 Inhibitors of Extensive Lung Destruction. *The Journal of infectious diseases.* 2018;218(1):53-
717 63.
- 718 7. Walker NF, Wilkinson KA, Meintjes G, Tezera LB, Goliath R, Peyper JM, Tadokera R,
719 Opondo C, Coussens AK, Wilkinson RJ, et al. Matrix Degradation in Human
720 Immunodeficiency Virus Type 1-Associated Tuberculosis and Tuberculosis Immune
721 Reconstitution Inflammatory Syndrome: A Prospective Observational Study. *Clin Infect Dis.*
722 2017;65(1):121-32.
- 723 8. Golichenari B, Velonia K, Nosrati R, Nezami A, Farokhi-Fard A, Abnous K, Behravan J, and
724 Tsatsakis AM. Label-free nano-biosensing on the road to tuberculosis detection. *Biosens*
725 *Bioelectron.* 2018;113:124-35.

- 726 9. Tsai TT, Huang CY, Chen CA, Shen SW, Wang MC, Cheng CM, and Chen CF. Diagnosis of
727 Tuberculosis Using Colorimetric Gold Nanoparticles on a Paper-Based Analytical Device.
728 *Acs Sensors*. 2017;2(9):1345-54.
- 729 10. Esterhuysen MM, Weiner J, 3rd, Caron E, Loxton AG, Iannaccone M, Wagman C, Saikali P,
730 Stanley K, Wolski WE, Mollenkopf HJ, et al. Epigenetics and Proteomics Join
731 Transcriptomics in the Quest for Tuberculosis Biomarkers. *MBio*. 2015;6(5).
- 732 11. Achkar JM, Cortes L, Croteau P, Yanofsky C, Mentinova M, Rajotte I, Schirm M, Zhou Y,
733 Junqueira-Kipnis AP, Kasprovicz VO, et al. Host Protein Biomarkers Identify Active
734 Tuberculosis in HIV Uninfected and Co-infected Individuals. *EBioMedicine*. 2015;2(9):1160-
735 8.
- 736 12. Zhang X, Liu F, Li Q, Jia H, Pan L, Xing A, Xu S, and Zhang Z. A proteomics approach to
737 the identification of plasma biomarkers for latent tuberculosis infection. *Diagn Microbiol*
738 *Infect Dis*. 2014;79(4):432-7.
- 739 13. Xu DD, Deng DF, Li X, Wei LL, Li YY, Yang XY, Yu W, Wang C, Jiang TT, Li ZJ, et al.
740 Discovery and identification of serum potential biomarkers for pulmonary tuberculosis using
741 iTRAQ-coupled two-dimensional LC-MS/MS. *Proteomics*. 2014;14(2-3):322-31.
- 742 14. Agranoff D, Fernandez-Reyes D, Papadopoulos MC, Rojas SA, Herbster M, Loosemore A,
743 Tarelli E, Sheldon J, Schwenk A, Pollak R, et al. Identification of diagnostic markers for
744 tuberculosis by proteomic fingerprinting of serum. *Lancet*. 2006;368(9540):1012-21.
- 745 15. Zhou F, Xu X, Wu S, Cui X, Fan L, and Pan W. Protein array identification of protein
746 markers for serodiagnosis of Mycobacterium tuberculosis infection. *Sci Rep*. 2015;5:15349.
- 747 16. Xu D, Li Y, Li X, Wei LL, Pan Z, Jiang TT, Chen ZL, Wang C, Cao WM, Zhang X, et al.
748 Serum protein S100A9, SOD3, and MMP9 as new diagnostic biomarkers for pulmonary
749 tuberculosis by iTRAQ-coupled two-dimensional LC-MS/MS. *Proteomics*. 2015;15(1):58-67.
- 750 17. Penn-Nicholson A, Hraha T, Thompson EG, Sterling D, Mbandi SK, Wall KM, Fisher M,
751 Suliman S, Shankar S, Hanekom WA, et al. Discovery and validation of a prognostic
752 proteomic signature for tuberculosis progression: A prospective cohort study. *PLoS Med*.
753 2019;16(4):e1002781.

- 754 18. MacLean E, Broger T, Yerliyaka S, Fernandez-Carballo BL, Pai M, and Denkinger CM. A
755 systematic review of biomarkers to detect active tuberculosis. *Nat Microbiol.* 2019;4(5):748-
756 58.
- 757 19. Seddon J, Kasprowicz V, Walker NF, Yuen HM, Sunpath H, Tezera L, Meintjes G,
758 Wilkinson RJ, Bishai WR, Friedland JS, et al. Procollagen III N-terminal propeptide and
759 desmosine are released by matrix destruction in pulmonary tuberculosis. *The Journal of*
760 *infectious diseases.* 2013;208(10):1571-9.
- 761 20. Song SH, Han M, Choi YS, Dan KS, Yang MG, Song J, Park SS, and Lee JH. Proteomic
762 profiling of serum from patients with tuberculosis. *Ann Lab Med.* 2014;34(5):345-53.
- 763 21. Wang C, Wei LL, Shi LY, Pan ZF, Yu XM, Li TY, Liu CM, Ping ZP, Jiang TT, Chen ZL, et
764 al. Screening and identification of five serum proteins as novel potential biomarkers for cured
765 pulmonary tuberculosis. *Sci Rep.* 2015;5:1561.
- 766 22. Sun H, Pan L, Jia H, Zhang Z, Gao M, Huang M, Wang J, Sun Q, Wei R, Du B, et al. Label-
767 Free Quantitative Proteomics Identifies Novel Plasma Biomarkers for Distinguishing
768 Pulmonary Tuberculosis and Latent Infection. *Front Microbiol.* 2018;9:1267.
- 769 23. Li C, He X, Li H, Zhou Y, Zang N, Hu S, Zheng Y, and He M. Discovery and verification of
770 serum differential expression proteins for pulmonary tuberculosis. *Tuberculosis (Edinb).*
771 2015.
- 772 24. Jiang TT, Shi LY, Wei LL, Li X, Yang S, Wang C, Liu CM, Chen ZL, Tu HH, Li ZJ, et al.
773 Serum amyloid A, protein Z, and C4b-binding protein beta chain as new potential biomarkers
774 for pulmonary tuberculosis. *PLoS One.* 2017;12(3):e0173304.
- 775 25. Hakimi A, Auluck J, Jones GD, Ng LL, and Jones DJ. Assessment of reproducibility in
776 depletion and enrichment workflows for plasma proteomics using label-free quantitative data-
777 independent LC-MS. *Proteomics.* 2014;14(1):4-13.
- 778 26. Yadav AK, Bhardwaj G, Basak T, Kumar D, Ahmad S, Priyadarshini R, Singh AK, Dash D,
779 and Sengupta S. A systematic analysis of eluted fraction of plasma post immunoaffinity
780 depletion: implications in biomarker discovery. *PLoS One.* 2011;6(9):e24442.

- 781 27. Garbis SD, Roumeliotis TI, Tyritzis SI, Zorpas KM, Pavlakis K, and Constantinides CA. A
782 novel multidimensional protein identification technology approach combining protein size
783 exclusion prefractionation, peptide zwitterion-ion hydrophilic interaction chromatography,
784 and nano-ultraperformance RP chromatography/nESI-MS2 for the in-depth analysis of the
785 serum proteome and phosphoproteome: application to clinical sera derived from humans with
786 benign prostate hyperplasia. *Anal Chem.* 2011;83(3):708-18.
- 787 28. Al-Daghri NM, Al-Attas OS, Johnston HE, Singhanian A, Alokail MS, Alkharfy KM, Abd-
788 Alrahman SH, Sabico SL, Roumeliotis TI, Manousopoulou-Garbis A, et al. Whole Serum 3D
789 LC-nESI-FTMS Quantitative Proteomics Reveals Sexual Dimorphism in the Milieu Interieur
790 of Overweight and Obese Adults. *J Proteome Res.* 2014;13(11):5094-105.
- 791 29. Zeidan B, Manousopoulou A, Garay-Baquero DJ, White CH, Larkin SET, Potter KN,
792 Roumeliotis TI, Papachristou EK, Copson E, Cutress RI, et al. Increased circulating resistin
793 levels in early-onset breast cancer patients of normal body mass index correlate with lymph
794 node negative involvement and longer disease free survival: a multi-center POSH cohort
795 serum proteomics study. *Breast Cancer Res.* 2018;20(1):19.
- 796 30. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, and von Mering C. Version 4.0 of
797 PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines.
798 *Proteomics.* 2015;15(18):3163-8.
- 799 31. PaxDB Team. PaxDb^{4.1}: Protein Abundance Database. <https://pax-db.org/dataset/9606/171/>.
800 Accessed May 20, 2019.
- 801 32. Anderson NL, and Anderson NG. The Human Plasma Proteome. *Molecular & Cellular*
802 *Proteomics.* 2002;1(11):845-67.
- 803 33. Pathan M, Keerthikumar S, Chisanga D, Alessandro R, Ang CS, Askenase P, Batagov AO,
804 Benito-Martin A, Camussi G, Clayton A, et al. A novel community driven software for
805 functional enrichment analysis of extracellular vesicles data. *J Extracell Vesicles.*
806 2017;6(1):1321455.

- 807 34. D'Angelo G, Chaerkady R, Yu W, Hizal DB, Hess S, Zhao W, Lekstrom K, Guo X, White
808 WI, Roskos L, et al. Statistical Models for the Analysis of Isobaric Tags Multiplexed
809 Quantitative Proteomics. *J Proteome Res.* 2017;16(9):3124-36.
- 810 35. Kammers K, Cole RN, Tiengwe C, and Ruczinski I. Detecting Significant Changes in Protein
811 Abundance. *EuPA Open Proteom.* 2015;7:11-9.
- 812 36. Smyth GK. Linear models and empirical bayes methods for assessing differential expression
813 in microarray experiments. *Stat Appl Genet Mol Biol.* 2004;3:Article3.
- 814 37. Cohen Freue GV, Meredith A, Smith D, Bergman A, Sasaki M, Lam KK, Hollander Z,
815 Opushneva N, Takhar M, Lin D, et al. Computational biomarker pipeline from discovery to
816 clinical implementation: plasma proteomic biomarkers for cardiac transplantation. *PLoS*
817 *Comput Biol.* 2013;9(4):e1002963.
- 818 38. Levin Y. The role of statistical power analysis in quantitative proteomics. *Proteomics.*
819 2011;11(12):2565-7.
- 820 39. Law CW, Chen Y, Shi W, and Smyth GK. voom: Precision weights unlock linear model
821 analysis tools for RNA-seq read counts. *Genome Biol.* 2014;15(2):R29.
- 822 40. Johnson WE, Li C, and Rabinovic A. Adjusting batch effects in microarray expression data
823 using empirical Bayes methods. *Biostatistics.* 2007;8(1):118-27.
- 824 41. Langfelder P, and Horvath S. WGCNA: an R package for weighted correlation network
825 analysis. *BMC Bioinformatics.* 2008;9:559.
- 826 42. Yu G, Wang LG, Han Y, and He QY. clusterProfiler: an R package for comparing biological
827 themes among gene clusters. *OmicS.* 2012;16(5):284-7.
- 828 43. Sigal GB, Segal MR, Mathew A, Jarlsberg L, Wang M, Barbero S, Small N, Haynesworth K,
829 Davis JL, Weiner M, et al. Biomarkers of Tuberculosis Severity and Treatment Effect: A
830 Directed Screen of 70 Host Markers in a Randomized Clinical Trial. *EBioMedicine.*
831 2017;25:112-121.
- 832 44. De Groote MA, Sterling DG, Hraha T, Russell T, Green LS, Wall K, Kraemer S, Ostroff R,
833 Janjic N, and Ochsner UA. Discovery and Validation of a Six-Marker Serum Protein

- 834 Signature for the Diagnosis of Active Pulmonary Tuberculosis. *J Clin Microbiol.* 2017;55
835 (10):3057-71
- 836 45. Manousopoulou A, Al-Daghri NM, Sabico S, Garay-Baquero DJ, Teng J, Alenad A, Alokail
837 MS, Athanasopoulos N, Deligeoroglou E, Chrousos GP, et al. Polycystic Ovary Syndrome
838 and Insulin Physiology: An Observational Quantitative Serum Proteomics Study in
839 Adolescent, Normal-Weight Females. *Proteomics Clin Appl.* 2019:e1800184.
- 840 46. Manousopoulou A, Hayden A, Mellone M, Garay-Baquero DJ, White CH, Noble F, Lopez
841 M, Thomas GJ, Underwood TJ, and Garbis SD. Quantitative proteomic profiling of primary
842 cancer-associated fibroblasts in oesophageal adenocarcinoma. *Br J Cancer.*
843 2018;118(9):1200-7.
- 844 47. Arul AB, and Robinson RAS. Sample Multiplexing Strategies in Quantitative Proteomics.
845 *Anal Chem.* 2019;91(1):178-89.
- 846 48. Huang T, Armbruster MR, Coulton JB, and Edwards JL. Chemical Tagging in Mass
847 Spectrometry for Systems Biology. *Anal Chem.* 2019;91(1):109-25.
- 848 49. Chen C, Yan T, Liu L, Wang J, and Jin Q. Identification of a Novel Serum Biomarker for
849 Tuberculosis Infection in Chinese HIV Patients by iTRAQ-Based Quantitative Proteomics.
850 *Front Microbiol.* 2018;9:330.
- 851 50. Johnston HE, Carter MJ, Cox KL, Dunscombe M, Manousopoulou A, Townsend PA, Garbis
852 SD, and Cragg MS. Integrated Cellular and Plasma Proteomics of Contrasting B-cell Cancers
853 Reveals Common, Unique and Systemic Signatures. *Mol Cell Proteomics.* 2017;16(3):386-
854 406
- 855 51. Larkin SE, Johnston HE, Jackson TR, Jamieson DG, Roumeliotis TI, Mockridge CI, Michael
856 A, Manousopoulou A, Papachristou EK, Brown MD, et al. Detection of candidate biomarkers
857 of prostate cancer progression in serum: a depletion-free 3D LC/MS quantitative proteomics
858 pilot study. *Br J Cancer.* 2016;115(9):1078-86.
- 859 52. Manousopoulou A, Hamdan M, Fotopoulos M, Garay-Baquero DJ, Teng J, Garbis SD, and
860 Cheong Y. Integrated Eutopic Endometrium and Non-Depleted Serum Quantitative

861 Proteomic Analysis Identifies Candidate Serological Markers of Endometriosis. *Proteomics*
862 *Clin Appl.* 2019;13(3):e1800153.

863 53. Brown J, Clark K, Smith C, Hopwood J, Lynard O, Toolan M, Creer D, Barker J, Breen R,
864 Brown T, et al. Variation in C - reactive protein response according to host and mycobacterial
865 characteristics in active tuberculosis. *BMC Infect Dis.* 2016;16:265.

866 54. Gopal R, Monin L, Torres D, Slight S, Mehra S, McKenna KC, Fallert Junecko BA, Reinhart
867 TA, Kolls J, Baez-Saldana R, et al. S100A8/A9 proteins mediate neutrophilic inflammation
868 and lung pathology during tuberculosis. *Am J Respir Crit Care Med.* 2013;188(9):1137-46.

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

871 56. Jacobs R, Maasdorp E, Malherbe S, Loxton AG, Stanley K, van der Spuy G, Walzl G, and
872 Chegou NN. Diagnostic Potential of Novel Salivary Host Biomarkers as Candidates for the
873 Immunological Diagnosis of Tuberculosis Disease and Monitoring of Tuberculosis Treatment
874 Response. *Plos One.* 2016;11(8).

875 57. Sathyamoorthy T, Tezera LB, Walker NF, Brilha S, Saraiva L, Mauri FA, Wilkinson RJ,
876 Friedland JS, and Elkington PT. Membrane Type 1 Matrix Metalloproteinase Regulates
877 Monocyte Migration and Collagen Destruction in Tuberculosis. *J Immunol.* 2015;195(3):882-
878 91.

879 58. Cliff JM, Kaufmann SH, McShane H, van Helden P, and O'Garra A. The human immune
880 response to tuberculosis and its treatment: a view from the blood. *Immunological reviews.*
881 2015;264(1):88-102.

882 59. Russell DG, Cardona PJ, Kim MJ, Allain S, and Altare F. Foamy macrophages and the
883 progression of the human tuberculosis granuloma. *Nat Immunol.* 2009;10(9):943-8.

884 60. Inoue M, Niki M, Ozeki Y, Nagi S, Chadeka EA, Yamaguchi T, Osada-Oka M, Ono K, Oda
885 T, Mwende F, et al. High-density lipoprotein suppresses tumor necrosis factor alpha
886 production by mycobacteria-infected human macrophages. *Sci Rep.* 2018;8(1):6736.

887 61. Cheng L, Han Y, Zhao X, Xu X, and Wang J. Identifying pathway modules of tuberculosis in
888 children by analyzing multiple different networks. *Exp Ther Med.* 2018;15(1):755-60.

- 889 62. Ahmad R, Xie L, Pyle M, Suarez MF, Broger T, Steinberg D, Ame SM, Lucero MG, Szucs
890 MJ, MacMullan M, et al. A rapid triage test for active pulmonary tuberculosis in adult
891 patients with persistent cough. *Sci Transl Med.* 2019;11(515).
- 892 63. Chegou NN, Sutherland JS, Malherbe S, Crampin AC, Corstjens PL, Geluk A, Mayanja-
893 Kizza H, Loxton AG, van der Spuy G, Stanley K, et al. Diagnostic performance of a seven-
894 marker serum protein biosignature for the diagnosis of active TB disease in African primary
895 healthcare clinic attendees with signs and symptoms suggestive of TB. *Thorax.*
896 2016;71(9):785-794
- 897 64. Xu D, Li Y, Li X, Wei LL, Pan Z, Jiang TT, Chen ZL, Wang C, Cao WM, Zhang X, et al.
898 Serum protein S100A9, SOD3 and MMP9 as new diagnostic biomarkers for pulmonary
899 tuberculosis by iTRAQ-coupled two-dimensional LC-MS/MS. *Proteomics.* 2015;15(1):58-67.
- 900 65. Wang C, Liu CM, Wei LL, Shi LY, Pan ZF, Mao LG, Wan XC, Ping ZP, Jiang TT, Chen ZL,
901 et al. A Group of Novel Serum Diagnostic Biomarkers for Multidrug-Resistant Tuberculosis
902 by iTRAQ-2D LC-MS/MS and Solexa Sequencing. *Int J Biol Sci.* 2016;12(2):246-56.
- 903 66. He PJW, Katis IN, Eason RW, and Sones CL. Rapid Multiplexed Detection on Lateral-Flow
904 Devices Using a Laser Direct-Write Technique. *Biosensors (Basel).* 2018;8(4).
- 905 67. Kim H, Chung DR, and Kang M. A new point-of-care test for the diagnosis of infectious
906 diseases based on multiplex lateral flow immunoassays. *Analyst.* 2019;144(8):2460-6.
- 907 68. Dheda K, Barry CE, 3rd, and Maartens G. Tuberculosis. *Lancet.* 2016;387(10024):1211-26.
- 908 69. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK. limma powers
909 differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids*
910 *Res.* 2015;43(7):e47.
- 911 70. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, and Vilo J. g:Profiler: a web
912 server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic*
913 *Acids Res.* 2019;47(W1):W191-98.
- 914 71. Supek F, Bosnjak M, Skunca N, and Smuc T. REVIGO Summarizes and Visualizes Long
915 Lists of Gene Ontology Terms. *Plos One.* 2011;6(7).

- 916 72. Yu GC, Wang LG, Han YY, and He QY. clusterProfiler: an R Package for Comparing
917 Biological Themes Among Gene Clusters. *Omics-a Journal of Integrative Biology*.
918 2012;16(5):284-7.
- 919 73. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti
920 A, Griss J, Mayer G, Eisenacher M, et al. The PRIDE database and related tools and resources
921 in 2019: improving support for quantification data. *Nucleic Acids Res*. 2019;47(D1):D442-
922 D50.
- 923
- 924

925 **Fig. 1. Overview of the plasma proteomic discovery and validation strategy of potential TB**
926 **biomarkers.**

927 (A) Identification and quantification of plasma proteins was performed using a quantitative
928 multidimensional protein identification approach which comprises a series of fractionation steps at
929 both protein (denaturing HP-SEC) and peptide level (offline high pH C4 HPLC followed by online
930 low pH C18 UPLC). Initial plasma prefractionation using HP-SEC produces 5 segments depending on
931 the molecular size. Only segments 1 to 4 were included in this study since these include most of the
932 protein contents. (B) Bioinformatic processing prioritised markers, which were then measured by
933 ELISA or luminex in plasma or serum samples from two cohorts. Discovery and validation stages
934 involved multiple ethnicities.

935

936 **Fig. 2. In-depth quantitative plasma proteome profiling in TB.**

937 (A) Violin plots with median and interquartile range show molecular weight frequency distributions
938 of proteins quantified (peptide confidence $\leq 1\%$ FDR) in each independent HP-SEC segment. The
939 number of proteins with relative quantitative data in all profiled samples is indicated. Four plasma
940 samples from TB patients, three healthy controls and one master pool were analysed. (B) Abundance
941 of quantified proteins from all HP-SEC segments. Only proteins with circulating levels reported in the
942 reference PaxDb1.4 protein abundance database or in the literature were annotated. Proteins
943 considered as classical plasma proteins are indicated in red, tissue leakage proteins in green, proteins
944 with signalling functions in purple and proteins associated to extracellular vesicles in yellow.
945 Concentrations of detected proteins span 11 orders of magnitude. (C) Principal component analysis
946 based on quantified proteins from all HP-SEC segments of eight profiled samples. iTRAQ tags and
947 groups are indicated. Overall, TB patients were separated from healthy controls by the principal
948 component PC1 and PC2, collectively explaining the 62% of total variance. The TB sample labelled
949 with tag 121 clustered with the healthy control samples. The master pool, a combination of all
950 samples, was located in the centre of the samples. (D) Log₂ transformed relative protein expression
951 heatmap of all proteins profiled in the four HP-SEC segments. Purple indicates TB patients and green
952 healthy controls. Pearson correlation was used for clustering of proteins and Spearman for samples.
953 Two clusters were defined based on the relative protein expression and GO analysis of these was
954 performed using g:Profiler. Cyan: downregulated proteins; Magenta: upregulated proteins.

955

956 **Fig. 3. Detailed profiling of segment 4 identifies a differential plasma proteome in TB infection.**

957 Analyses of common quantified proteins (peptide confidence FDR ≤ 0.01) derived from HP-SEC
958 segment 4 across three iTRAQ experiments studying 10 controls and 11 TB patients (n=426 proteins).
959 (A) Volcano plot representation of plasma proteins differentially expressed in TB defined by LIMMA
960 with FDR correction (q-value ≤ 0.05). Red indicates upregulated proteins and blue downregulated.
961 Gene names of significantly regulated proteins with \log_2 fold change $\geq |0.5|$ are shown. (B) WGCNA

962 cluster dendrogram of quantified proteins into distinctive modules defined by dendrogram branch
963 cutting. Colour modules indicate protein clusters of highly interconnected proteins associated to the
964 disease status. Correlation score and significance demonstrates that module turquoise is strongly
965 correlated to TB status. (C) Gene ontology enrichment of proteins included in the module turquoise
966 (n=189). Dots represent the top 20 enriched cellular component organisation terms. Dot colour
967 indicates significance (*p-value* Benjamini-Hochberg adjusted) and size represents the number of
968 differential proteins in the significant gene list associated with the GO term.

969

970 **Fig. 4. Physiological changes in pulmonary TB are reflected in the plasma proteome.**

971 Functional enrichment analysis of the biological processes was performed on the one hundred and
972 eighty nine proteins strongly associated to the TB status and identified by WGCNA. Gene-concept
973 network (c-net plot) depicts the linkages of proteins and the top 20 biological process terms enriched
974 in the turquoise module. Up-regulated and down-regulated proteins were included. Green-to-red
975 coding next to the network indicates the log₂FC. Proteins in bold were selected for validation.

976

977 **Fig. 5. Top candidate biomarkers for active TB link to multiple biological processes.** Chord plot
978 for plasma proteins strongly correlated to active TB and identified by combining outputs from
979 WGCNA and LIMMA. This plot links these proteins via ribbons to their associated biological
980 processes. Blue-to-red coding next to the proteins indicates the log₂FC. Gene ontology enrichment for
981 biological process was performed in g:Profiler and only significant terms (FDR *q-value* ≤ 0.05) are
982 shown. Plot generated with the R package GOplots.

983

984 **Fig. 6. Novel TB biomarkers validate in an independent UK cohort of mixed ethnicity.**

985 Two novel TB biomarkers were significantly upregulated in TB infection measured by luminex or
986 ELISA in serum from an independent UK-based cohort. (A) CFHR5 (Complement factor H related
987 protein 5) is increased in TB, and also significantly increased in other respiratory diseases (ORDs).
988 Four known TB potential markers were measured and were significantly elevated in TB: (B) LRG1
989 (Leucine-rich alpha-2-glycoprotein). (C) LBP (Lipopolysaccharide binding protein), (D) SAA1
990 (Serum amyloid A1), (E) CRP (C-reactive protein). (F) ILF2 (interleukin enhancer binding factor 2),
991 a novel analyte from segment 3, was elevated in TB and ORDs. Box displays 25% and 75%
992 percentiles with line showing median, and whiskers displaying minimum to maximum values.
993 Differences were considered significant when *p-value* < 0.05 and calculated from Kruskal-Wallis test
994 and Dunn's multiple comparison test. HC: Healthy controls (n=30), LTBI: Latent TB infection
995 (n=30), PTBI: Pulmonary TB infection (n=32) and OR: Other respiratory diseases (n=26).

996

997 **Fig. 7. Combination of five protein markers discriminates TB patients in a UK-based cohort**

998 Receiver operator characteristic (ROC) curves were generated using SPSS v.25, for individual
999 proteins (CFHR5, LBP, SAA, CRP and ILF2) and after binary logistic regression for combined
1000 analytes. AUC was estimated under nonparametric assumption. TB was set as the positive test
1001 outcome and the test direction such that larger test result indicates a more positive test. ROC curve for
1002 TB infection vs. healthy controls shows good discrimination, with the multiplex panel most
1003 discriminatory **(A)**, while the ROC curve for TB infection vs. ORD shows individual analytes are not
1004 differentiating, but a combined multiplex generates an AUC of 0.813 **(B)**.

1005

1006 **Fig. 8. CFHR5 validates as a new diagnostic marker of TB in HIV-coinfection, and multiplex**
1007 **analysis performs well against other respiratory conditions.**

1008 **(A)** CFHR5 was significantly upregulated during active TB infection in a previously reported South
1009 African cohort, in both HIV uninfected and HIV infected individuals. Three other potential TB
1010 markers were also elevated: **(B)** LBP, **(C)** SAA1 and CRP (previously reported). Box displays 25%
1011 and 75% percentiles with line showing median, and whiskers displaying minimum to maximum
1012 values. Differences were considered significant when p-value <0.05 and calculated from Kruskal-
1013 Wallis test and Dunn's multiple comparison test. HC: Healthy controls (n=60), PTBI: Pulmonary TB
1014 infection (n=39) and ORD: Other respiratory diseases (n=22). HIV indicates HIV co-infection. HC-
1015 HIV (n=16), ATBI-HIV (n=53) and ORD-HIV (n=13)

1016

1017 **Fig. 9. Combination of four protein markers discriminates TB patients with HIV-coinfection**

1018 Receiver operator characteristic (ROC) curves were generated using SPSS v.25, for individual
1019 proteins (CFHR5, LBP, SAA1 and CRP) and after binary logistic regression for combined analytes.
1020 ROC curve for TB infection vs. ORD in HIV uninfected individuals shows optimal performance from
1021 the combined host panel, with AUC of 0.888 **(A)**. Analysis of TB infection vs. ORD in HIV co-
1022 infected individuals produced an AUC of 0.976 from the combined panel **(B)**.

1023

1024

1025

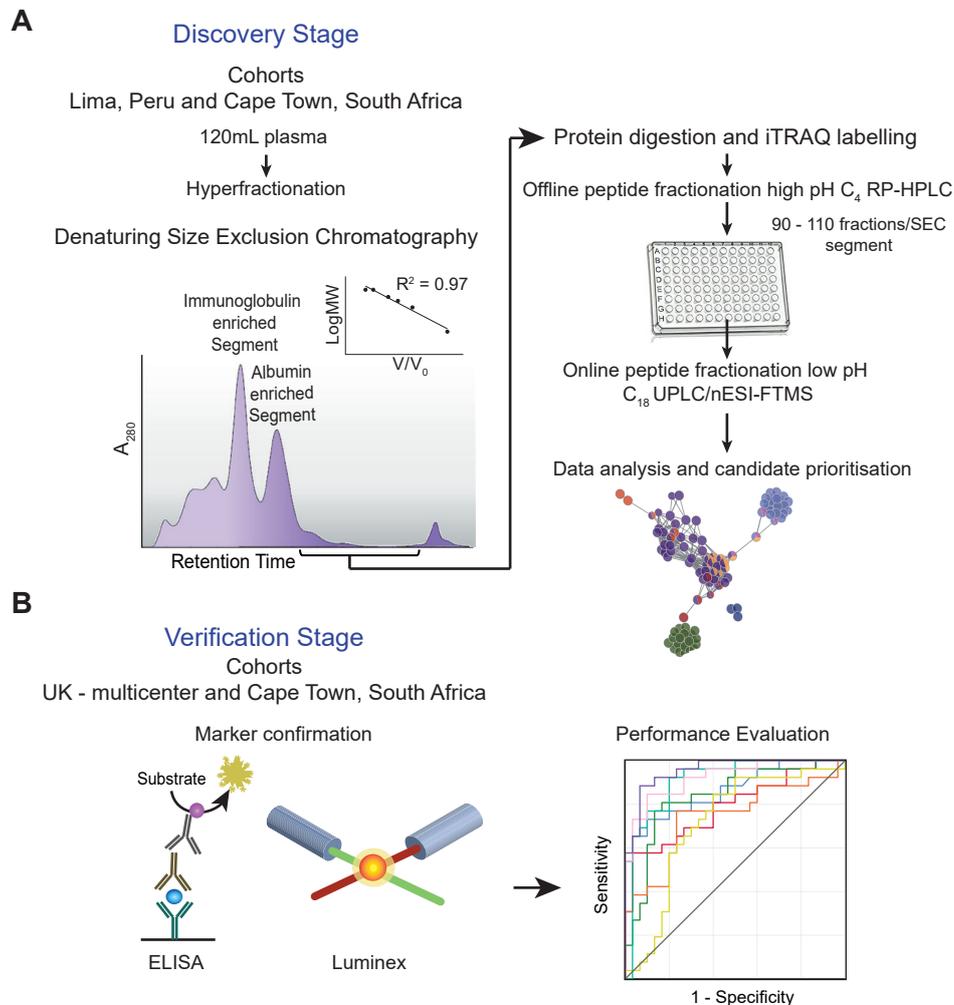


Fig. 1. Overview of the plasma proteomic discovery and validation strategy of potential TB biomarkers.

(A) Identification and quantification of plasma proteins was performed using a quantitative multidimensional protein identification approach which comprises a series of fractionation steps at both protein (denaturing HP-SEC) and peptide level (offline high pH C₄ HPLC followed by online low pH C₁₈ UPLC). Initial plasma prefractionation using HP-SEC produces 5 segments depending on the molecular size. Only segments 1 to 4 were included in this study since these include most of the protein contents. (B) Bioinformatic processing prioritised markers, which were then measured by ELISA or luminex in plasma or serum samples from two cohorts. Discovery and validation stages involved multiple ethnicities.

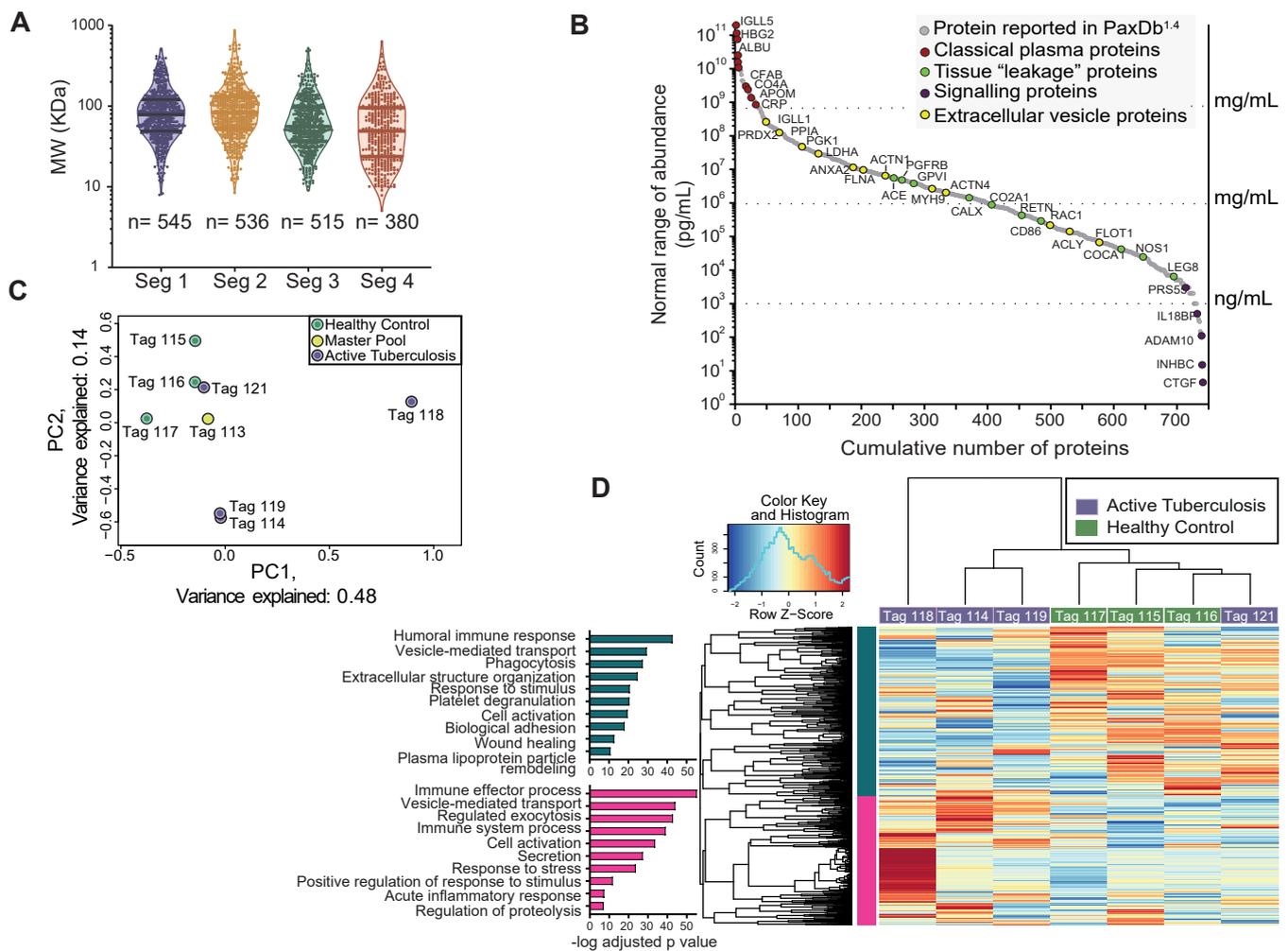


Fig. 2. In-depth quantitative plasma proteome profiling in TB.

(A) Violin plots with median and interquartile range show molecular weight frequency distributions of proteins quantified (peptide confidence $\leq 1\%$ FDR) in each independent HP-SEC segment. The number of proteins with relative quantitative data in all profiled samples is indicated. Four plasma samples from TB patients, three healthy controls and one master pool were analysed. (B) Abundance of quantified proteins from all HP-SEC segments. Only proteins with circulating levels reported in the reference PaxDb1.4 protein abundance database or in the literature were annotated. Proteins considered as classical plasma proteins are indicated in red, tissue leakage proteins in green, proteins with signalling functions in purple and proteins associated to extracellular vesicles in yellow. Concentrations of detected proteins span 11 orders of magnitude. (C) Principal component analysis based on quantified proteins from all HP-SEC segments of eight profiled samples. iTRAQ tags and groups are indicated. Overall, TB patients were separated from healthy controls by the principal component PC1 and PC2, collectively explaining the 62% of total variance. The TB sample labelled with tag 121 clustered with the healthy control samples. The master pool, a combination of all samples, was located in the centre of the samples. (D) Log₂ transformed relative protein expression heatmap of all proteins profiled in the four HP-SEC segments. Purple indicates TB patients and green healthy controls. Pearson correlation was used for clustering of proteins and Spearman for samples. Two clusters were defined based on the relative protein expression and GO analysis of these was performed using g:Profiler. Cyan: downregulated proteins; Magenta: upregulated proteins.

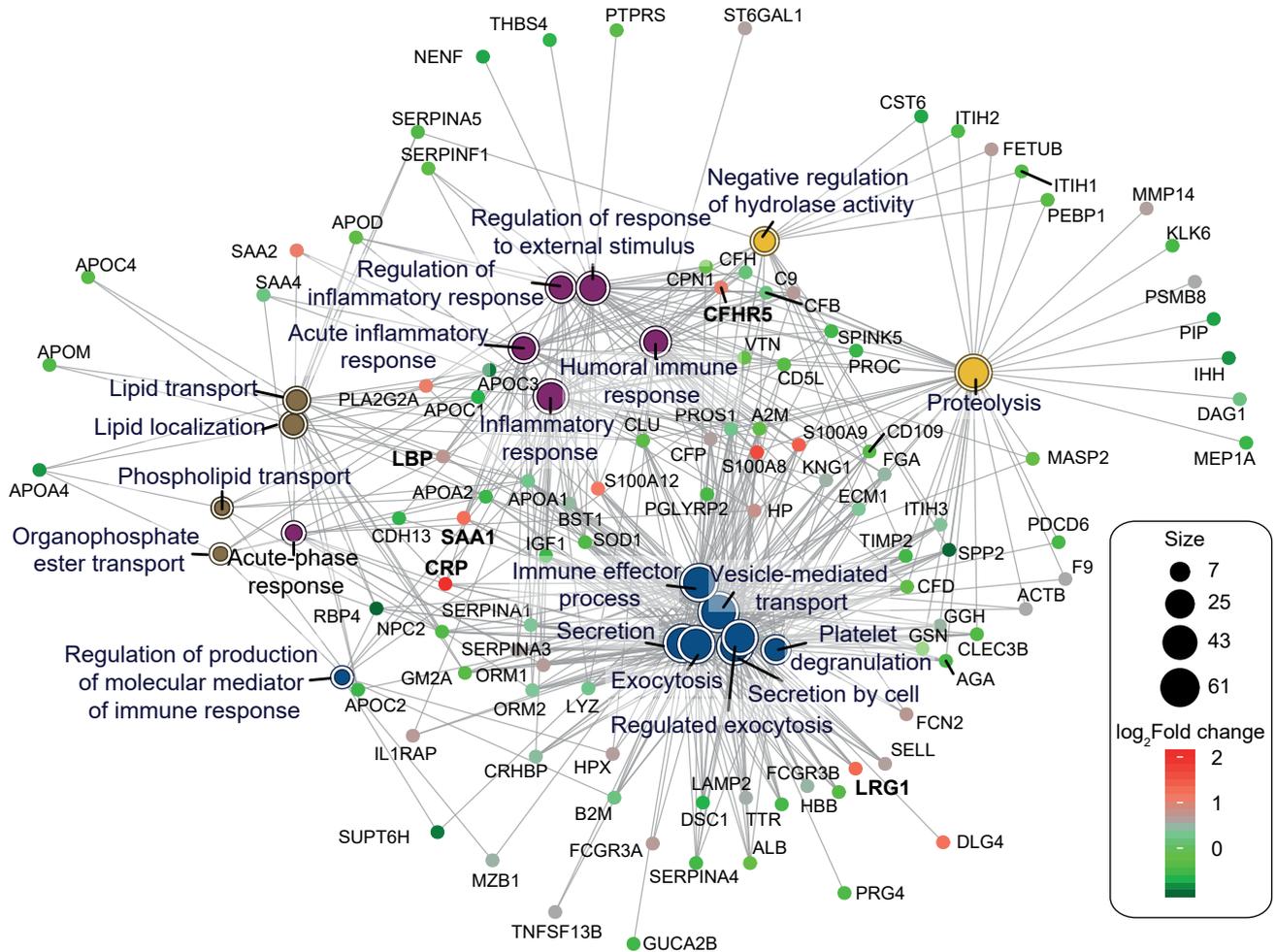


Fig. 4. Physiological changes in pulmonary TB are reflected in the plasma proteome.

Functional enrichment analysis of the biological processes was performed on the one hundred and eighty nine proteins strongly associated to the TB status and identified by WGCNA. Gene-concept network (c-net plot) depicts the linkages of proteins and the top 20 biological process terms enriched in the turquoise module. Up-regulated and down-regulated proteins were included. Green-to-red coding next to the network indicates the \log_2FC . Proteins in bold were selected for validation.

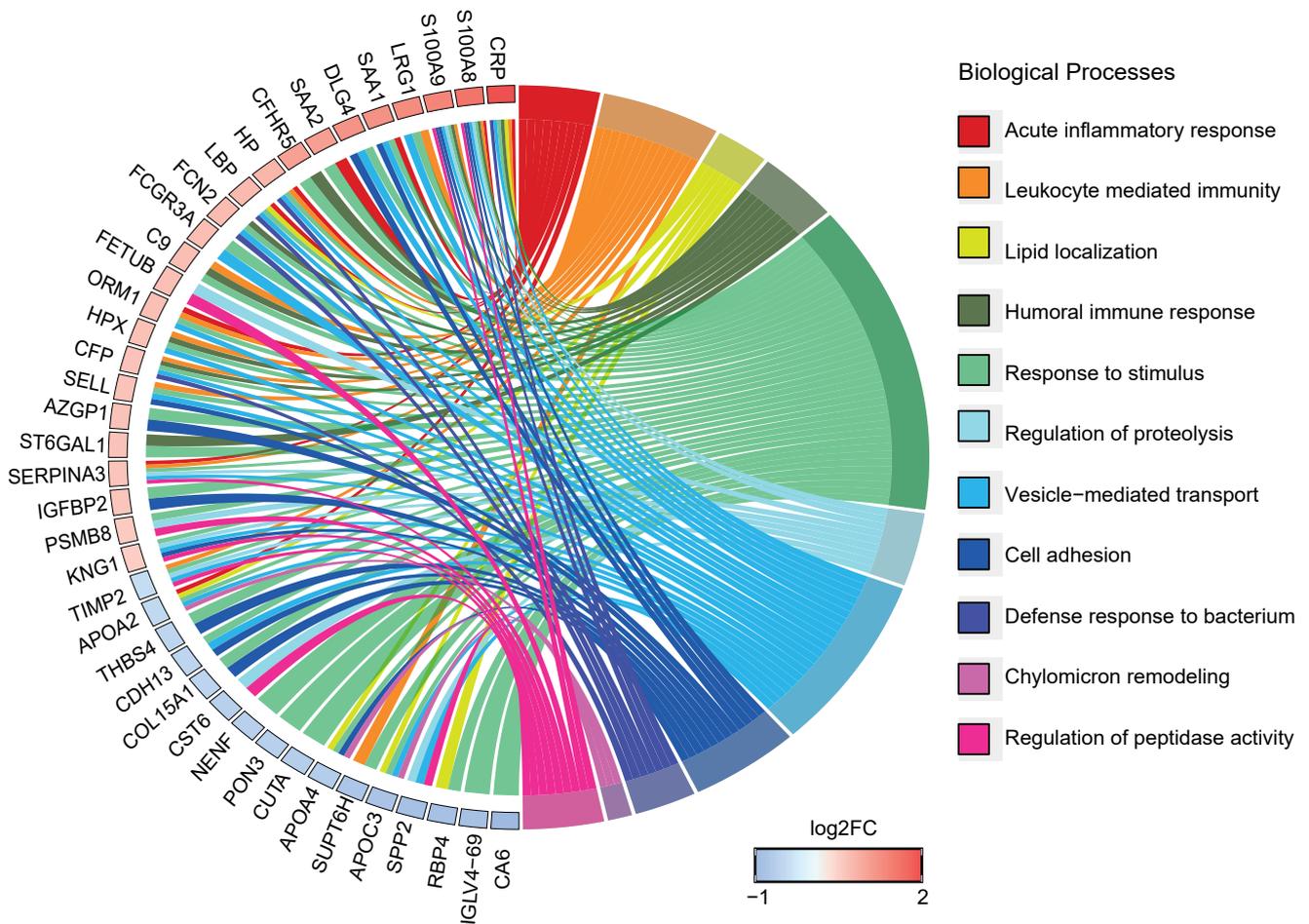


Fig. 5. Top candidate biomarkers for active TB link to multiple biological processes.

Chord plot for plasma proteins strongly correlated to active TB and identified by combining outputs from WGCNA and LIMMA. This plot links these proteins via ribbons to their associated biological processes. Blue-to-red coding next to the proteins indicates the log₂FC. Gene ontology enrichment for biological process was performed in g:Profiler and only significant terms (FDR *q-value* ≤ 0.05) are shown. Plot generated with the R package GOplots.

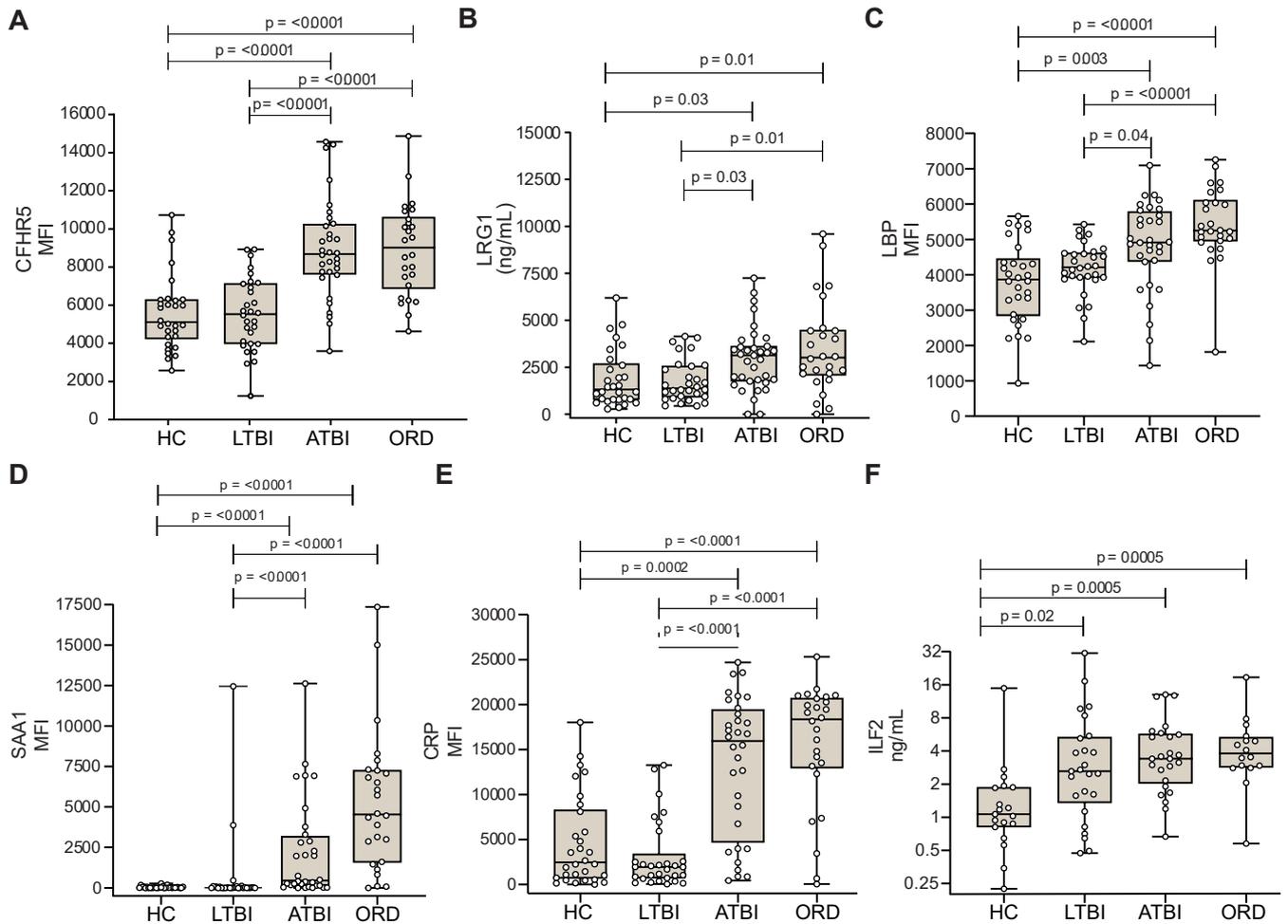


Fig. 6. Novel TB biomarkers validate in an independent UK cohort of mixed ethnicity.

Two novel TB biomarkers were significantly upregulated in TB infection measured by luminex or ELISA in serum from an independent UK-based cohort. **(A)** CFHR5 (Complement factor H related protein 5) is increased in TB, and also significantly increased in other respiratory diseases (ORDs). Four known TB potential markers were measured and were significantly elevated in TB: **(B)** LRG1 (Leucine-rich alpha-2-glycoprotein). **(C)** LBP (Lipopolysaccharide binding protein), **(D)** SAA1 (Serum amyloid A1), **(E)** CRP (C-reactive protein). **(F)** ILF2 (interleukin enhancer binding factor 2), a novel analyte from segment 3, was elevated in TB and ORDs. Box displays 25% and 75% percentiles with line showing median, and whiskers displaying minimum to maximum values. Differences were considered significant when p -value < 0.05 and calculated from Kruskal-Wallis test and Dunn's multiple comparison test. HC: Healthy controls ($n = 30$), LTBI: Latent TB infection ($n = 30$), PTBI: Pulmonary TB infection ($n = 32$) and OR: Other respiratory diseases ($n = 26$).

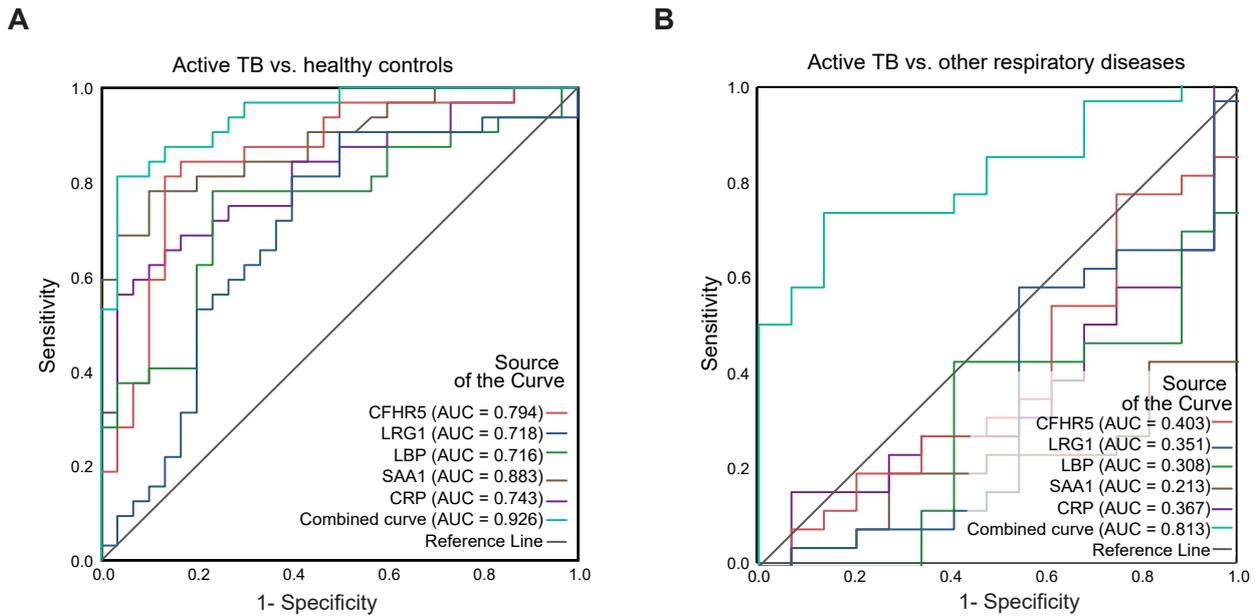


Fig. 7. Combination of five protein markers discriminates TB patients in a UK-based cohort

Receiver operator characteristic (ROC) curves were generated using SPSS v.25, for individual proteins (CFHR5, LBP, SAA, CRP and ILF2) and after binary logistic regression for combined analytes. AUC was estimated under nonparametric assumption. TB was set as the positive test outcome and the test direction such that larger test result indicates a more positive test. ROC curve for TB infection vs. healthy controls shows good discrimination, with the multiplex panel most discriminatory **(A)**, while the ROC curve for TB infection vs. ORD shows individual analytes are not differentiating, but a combined multiplex generates an AUC of 0.813 **(B)**.

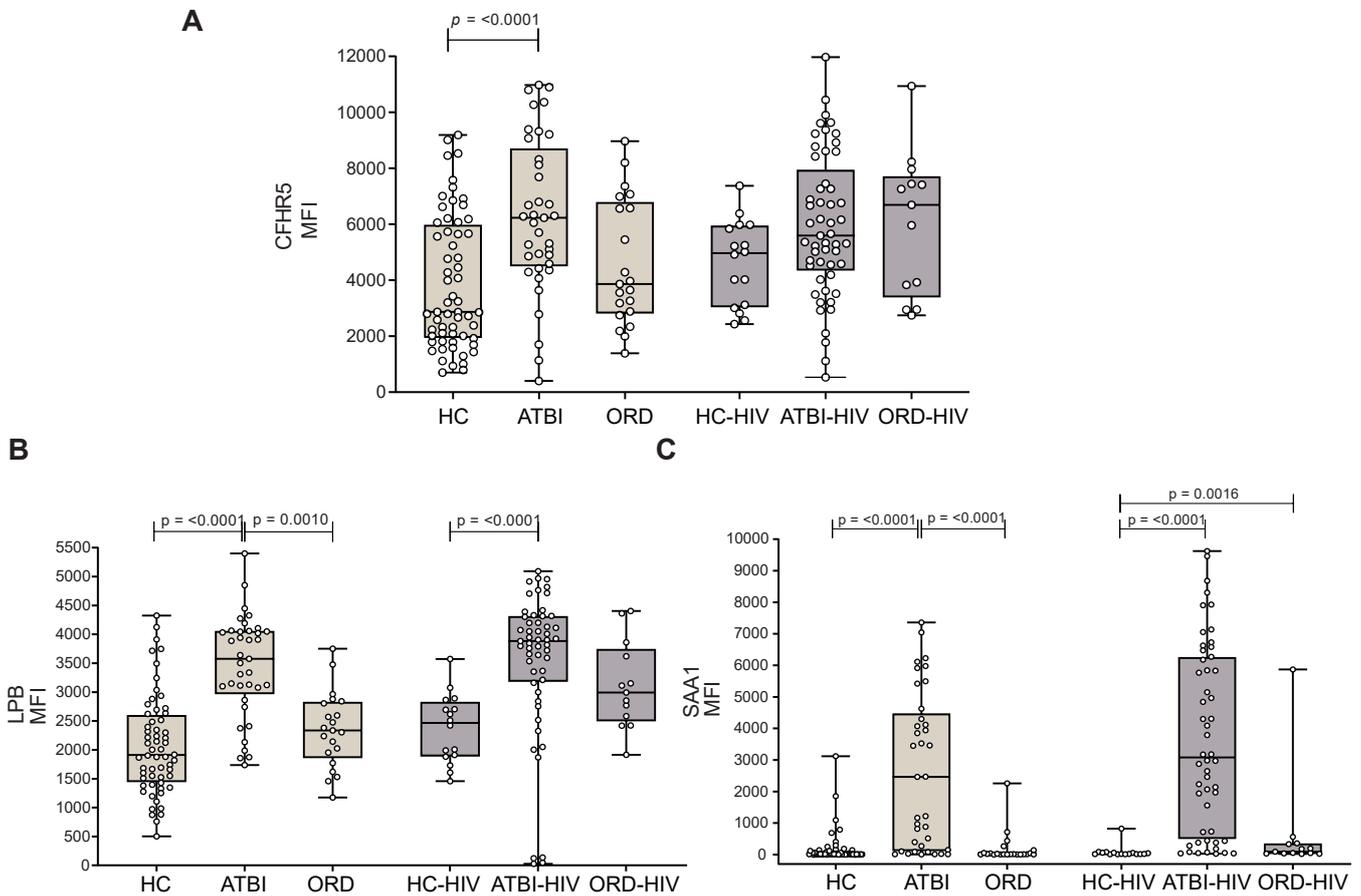


Fig. 8. CFHR5 validates as a new diagnostic marker of TB in HIV-coinfection, and multiplex analysis performs well against other respiratory conditions.

(A) CFHR5 was significantly upregulated during active TB infection in a previously reported South African cohort, in both HIV uninfected and HIV infected individuals. Three other potential TB markers were also elevated: (B) LBP, (C) SAA1 and CRP (previously reported). Box displays 25% and 75% percentiles with line showing median, and whiskers displaying minimum to maximum values. Differences were considered significant when p -value < 0.05 and calculated from Kruskal-Wallis test and Dunn's multiple comparison test. HC: Healthy controls ($n = 60$), PTBI: Pulmonary TB infection ($n = 39$) and ORD: Other respiratory diseases ($n = 22$). HIV indicates HIV co-infection. HC-HIV ($n = 16$), ATBI-HIV ($n = 53$) and ORD-HIV ($n = 13$)

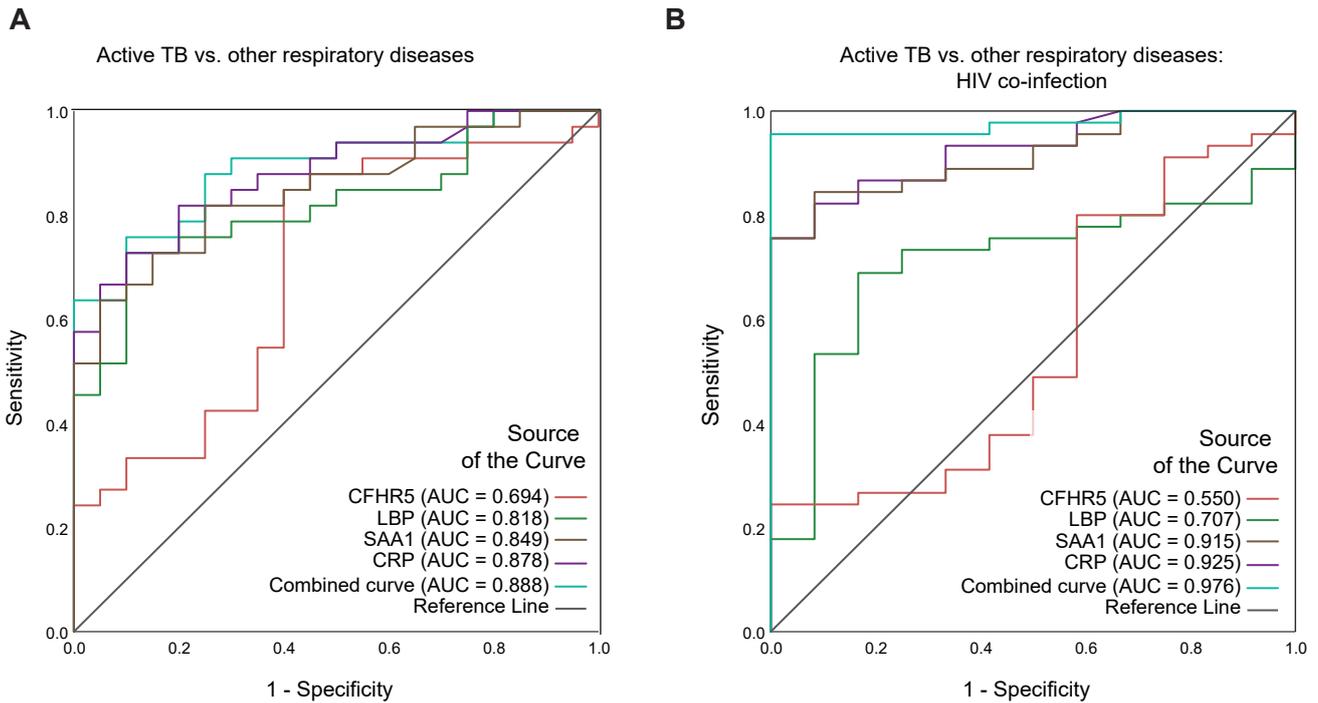


Fig. 9. Combination of four protein markers discriminates TB patients with HIV-coinfection

Receiver operator characteristic (ROC) curves were generated using SPSS v.25, for individual proteins (CFHR5, LBP, SAA1 and CRP) and after binary logistic regression for combined analytes. ROC curve for TB infection vs. ORD in HIV uninfected individuals shows optimal performance from the combined host panel, with AUC of 0.888 (**A**). Analysis of TB infection vs. ORD in HIV co-infected individuals produced an AUC of 0.976 from the combined panel (**B**).