Enrichment-assisted PhIP-seq discovers a serum antibody peptide signature for active tuberculosis

Haoyun Xu[a#], Mingxiang Huang[b#], Zhaowei Xu[c], Jiaoxiang Wu[d], Chongxiang Tong[e], Yunhong Tan[f], Quan Wang[g], Xundi Bao[h], Sheng Tan[a], Shujuan Guo[a], Zigan Sha[a], Jiahao Xie[a], Hongzhen Di[a], Mingliang Ma[i], Jinhua Wu[i], Yang Li[i], Huan Qi[a]*, Sheng-ce Tao[a]*

[a] Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, No. 800, Dongchuan Road, Minhang District, Shanghai, P. R. China, 200240

[b] Fujian Medical University Clinical Teaching Hospital, Department of laboratory Medicine, Fuzhou Pulmonary Hospital, No. 8, Beiyuan Road, Cangshan District, Fuzhou City, Fujian Province, P. R. China, 350008

[c] Key Laboratory of Gastrointestinal Cancer (Fujian Medical University), Ministry of Education, No. 1, Xue Yuan Road, University Town, Fuzhou City, Fujian Province, P. R. China, 350122

[d] Department of Infectious Diseases and Department of Clinical Laboratory, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, No. 1111 XianXia Road, Shanghai, P. R. China, 200336

[e] Laboratory Department of Lanzhou Pulmonary Hospital, No. 968, North Binhe Middle Road, Lanzhou City, Gansu Province, P. R. China, 730046

[f] Hunan Institute for Tuberculosis Control, No. 519, Xianjiahu Road, Yuelu District, Changsha, Hunan Province, P. R. China, 410013

[g] Sixth People's Hospital of Xinjiang Uygur Autonomous Region, No. 100, Beiyixiang, Kashi East Road, Xinshi District, Urumqi City, Xinjiang Uygur Autonomous Region, P. R. China, 830011

[h] Institute of Microbiological Testing, Anhui Provincial Center for Disease Control and Prevention, No. 12560, Fanhua Avenue, Economic & Technological Development Zone, Hefei, Anhui Province, P. R. China, 230601

[i] AbCode Biotechnology Co., Ltd, Room 702, Building 1, No. 58, Yuanmei Road, Minhang District, Shanghai, P. R. China

[#] These authors contributed equally

[*] Corresponding author: [taosc@sjtu.edu.cn; qihuan@sjtu.edu.cn]

In brief

Xu *et al.* develop enrichment-assisted PhIP-seq to recover tuberculosis-reactive serum antibody signals and nominate a compact peptide signature for active tuberculosis. Internal testing supports candidate biomarker potential, while disease-control analysis against *M. intracellulare* defines a key boundary for clinical translation.

Highlights

- ePhIP-seq enriches tuberculosis-reactive serum antibodies before profiling the antibody reactome
- A phage library displaying 86,287 overlapping *M. tuberculosis* peptides supports systematic antibody-reactive peptide discovery
- A compact 3-marker candidate antibody signature for tuberculosis diagnosis

Summary

Serological diagnosis of active tuberculosis has been limited by inconsistent performance of historical antibody tests, partly reflecting incomplete and candidate-driven antigen discovery. Here, we developed enrichment-assisted phage immunoprecipitation sequencing (ePhIP-seq) to systematically identify serum IgG-reactive peptide biomarkers for active tuberculosis. We constructed a proteome-wide *Mycobacterium tuberculosis* H37Rv T7 phage-display library containing 86,287 overlapping 30-amino-acid peptides and enriched H37Rv-reactive serum antibodies before PhIP-seq screening. In a discovery cohort of 30 patients with active tuberculosis and 30 healthy controls, ePhIP-seq identified 277 differential peptide candidates, from which 21 were selected for orthogonal peptide microarray validation. 15 peptides showed higher IgG reactivity in active tuberculosis and were used for model development. A compact 3-marker candidate signature, Model C359, comprising MTB-N-9, MTB-N-10, and MTB-N-14, achieved an apparent AUC of 0.951 and a bootstrap optimism-corrected AUC of 0.942 in the model-building cohort. In a held-out internal test set, Model C359 achieved an AUC of 0.886 for active tuberculosis versus healthy controls. In exploratory disease-control analysis against pulmonary *Mycobacterium intracellulare* infection, the AUC was 0.672. These findings nominate a compact serum antibody peptide signature for further validation and support ePhIP-seq as a reusable antigen-discovery strategy for diseases in which pathogen-specific antibody signals may be weak or diluted within total serum IgG.

Keywords

Tuberculosis; *Mycobacterium tuberculosis*; H37Rv; Biomarker; PhIP-seq

Introduction

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (MTB), remains one of the most important infectious diseases worldwide. According to the World Health Organization (WHO) Global Tuberculosis Report 2025, an estimated 10.7 million people developed TB in 2024 and TB caused approximately 1.23 million deaths, including deaths among people living with human immunodeficiency virus (HIV).[1] Although major advances have been made in molecular diagnosis and treatment, delayed or missed diagnosis continues to contribute to ongoing transmission, disease progression, and mortality.[2] The diagnostic gap is particularly important for patients with paucibacillary disease, extrapulmonary disease, HIV co-infection, or difficulty producing sputum.[3–5]

Current diagnostic strategies for TB include sputum smear microscopy, mycobacterial culture, nucleic-acid amplification tests, radiological assessment, immunological tests

for TB infection, and drug-resistance testing. These methods have complementary strengths and also limitations. Culture remains slow,[6] smear microscopy lacks sensitivity,[2,7] and nucleic-acid amplification tests depend on specimen availability, infrastructure, and appropriate implementation.[8] WHO guidelines and target-product profiles continue to emphasize the need for rapid, accurate tests that can be deployed closer to the point of care and that can improve access to diagnosis.[9,10] A robust blood-based diagnostic test would be especially valuable because serum is easier to collect, process, store, and transport than many respiratory specimens.

Antibody-based serum diagnostics have long been attractive for TB because they can, in principle, be simple, rapid, low-cost, and compatible with point-of-care formats. However, historical commercial serological tests for TB showed variable sensitivity and specificity,[11–13] leading WHO to recommend against their use.[14] One reason for this poor performance is that many prior efforts relied on a limited number of candidate antigens rather than systematic sampling of the MTB antigenic repertoire. More recent studies have used higher-throughput platforms, such as protein arrays and proteome-level antigen discovery, to identify more diverse antibody targets, but clinical translation remains challenging.[15,16]

Phage immunoprecipitation sequencing (PhIP-seq) provides a scalable strategy for profiling antibody reactivity against large peptide libraries. The method combines synthetic oligonucleotide-encoded peptide libraries, phage display, immunoprecipitation, and next-generation sequencing, enabling high-throughput identification of antibody-bound peptides.[17,18] PhIP-seq and related phage-display methods have been applied to autoantigen discovery, viral serology, and antibody epitope mapping.[19–23] For TB, a proteome-wide peptide-resolution strategy has potential advantages: it can identify linear antibody epitopes across the H37Rv proteome, capture non-classical antigens, and generate short peptide biomarkers that potentially easier to clinical translation than whole proteins.

However, serum contains an extremely diverse antibody repertoire,[24,25] and pathogen-specific antibodies may represent only a small fraction of total IgG. In conventional PhIP-seq, immunoprecipitation using Protein A/G can capture both antigen-specific and unrelated antibodies, potentially hampering sensitivity when disease-relevant antibodies are rare. We therefore hypothesized that enriching MTB-reactive serum antibodies before PhIP-seq would improve the discovery of TB-associated antibody-binding peptides. In this study, we constructed a high-coverage H37Rv proteome-wide peptide phage library (TbMap), developed an H37Rv lysate-based antibody-enrichment workflow, discovered candidate peptide biomarkers by PhIP-seq, validated them using a peptide microarray, and built diagnostic models for ATB discrimination.

Results

ePhIP-seq and construction of a proteome-wide H37Rv peptide phage library

To improve the sensitivity of PhIP-seq, we first developed ePhIP-seq, as illustrated in **Figure 1A**. Using MTB as a model, we first constructed a proteome-wide peptide phage library covering the annotated protein-coding sequences of MTB strain H37Rv (**Figure 1A**). Compared with conventional PhIP-seq procedure, an additional pre-enrichment step was incorporated into the standard workflow. In this step, antibodies

against MTB were enriched by incubating serum with biotinylated MTB H37Rv cell lysate, followed by capture with streptavidin magnetic beads. The following parts of ePhIP-seq were the same as those of PhIP-seq, including phage immunoprecipitation, sequencing and analysis.

The library was designed as 86,287 overlapping 30-amino-acid peptides with 15-amino-acid overlap, representing 4,025 H37Rv proteins. This design provided peptide-level sampling of the H37Rv proteome while retaining short linear sequences suitable for phage display and downstream peptide-biomarker development.

After oligonucleotide synthesis, insert preparation, phage packaging, amplification, and purification, the library was evaluated by dot blot and next-generation sequencing. Dot blot using an anti-FLAG antibody showed that approximately half of plaques carried detectable C-terminal FLAG signal, consistent with successful display of designed peptide inserts in a substantial fraction of the library (**Figure 1B**). Sequencing-based quality control showed a relatively uniform clone abundance distribution (**Figure 1C**). Of the 86,287 designed peptides, 82,529 were detected by sequencing, corresponding to 96% peptide-level coverage (**Figure 1D**). At the protein level, 85% of H37Rv proteins were fully covered (**Figure 1E**). Sequencing-depth analysis showed that the Chao1-based library-complexity estimate approached saturation at high read depth, supporting adequate sequencing depth for library characterization (**Figure S1**). These results established a high-coverage H37Rv peptide phage library suitable for serum antibody profiling.

Enrichment of H37Rv-reactive serum antibodies by ePhIP-seq

We reasoned that antibodies reactive with MTB antigens may represent a small fraction of the total serum antibody repertoire and that enriching pathogen-reactive antibodies could improve PhIP-seq sensitivity. To enrich H37Rv-reactive antibodies, H37Rv total lysate was biotinylated, immobilized on streptavidin beads, and used to capture serum antibodies in ePhIP-seq. Experimental validation confirmed successful lysate biotinylation and bead coupling (**Figure 2A**). Pilot experiments using ATB and HC sera showed that H37Rv-reactive antibodies could be captured and eluted from serum samples (**Figure 2B**).

The PhIP-seq discovery cohort included 30 ATB patients and 30 HCs matched for age and sex (**Figure 2C**). After antibody enrichment, sera were profiled against the H37Rv peptide phage library.

A sequential filtering strategy was employed to identify potential biomarkers. First, 1,641 peptides were retained due to their higher normalized read count across all ATB patients compared with library input. Second, 1,154 peptides exhibited stronger immune reactivity in the ATB group than in the HC group were retained. Third, 277 differential peptides were identified by *t*-test. From these, 21 peptides were selected for orthogonal validation based on multiple criteria (**Figure 2D**). The selected peptides showed stronger reactivity or diagnostic potential in the discovery analysis (**Figure 2E,** and **Table S3**), providing a focused candidate set for validation.

Development of a peptide microarray for orthogonal candidate validation

To validate ePhIP-seq-derived peptide candidates, we designed a microarray-based pipeline (**Figure 3A**). The 21 selected peptides were expressed as peptide-rHF fusion proteins and printed onto peptide microarrays (**Figure 3B, C**). SDS-PAGE and

Western blot analyses confirmed successful expression and purification for the candidate fusion proteins, although smaller bands were observed for some constructs and are interpreted conservatively as possible degradation products (**Figure 3B**, and **Figure S2A**). Microarray quality-control experiments confirmed successful printing and immobilization of candidate antigens and control proteins, with appropriate spot signal and morphology (**Figure 3C**).

During assay optimization, serum samples showed reactivity against the rHF carrier protein, which could introduce non-peptide-specific background. We therefore developed a serum pretreatment strategy to remove rHF-reactive antibodies (**Figure 3A**). rHF-background signal decreased after pretreatment, and 1:320 serum dilution showed comparable rHF signal to background levels (**Figure S2B, C**). These optimization experiments supported the use of the peptide microarray for candidate validation.

A parsimonious 3-marker model retained higher diagnostic performance

Using the optimized peptide microarray workflow, we evaluated the 21 candidate peptides in model-development cohort of 50 ATB patients and 50 HCs with matched age and sex (**Figure 4A**). After systematic construction (**Figure 4B**), a parsimonious 3-marker model was retained.

First, as shown in the **Figure S3**, 15 peptides showed higher IgG reactivity in ATB sera than in HC sera. These peptides were retained for diagnostic model construction and biological interpretation. The 15 retained candidates were MTB-N-2, MTB-N-4, MTB-N-8, MTB-N-9, MTB-N-10, MTB-N-11, MTB-N-12, MTB-N-13, MTB-N-14, MTB-N-15, MTB-N-16, MTB-N-17, MTB-N-18, MTB-N-19, and MTB-N-20. MTB-N-14, MTB-N-10 and MTB-N-9 showed the strongest single-marker performance. MTB-N-14 achieved an AUC of 0.886, MTB-N-10 achieved an AUC of 0.845, and MTB-N-9 achieved an AUC of 0.708. These results indicated that individual peptides have diagnostic potential, however, combinations may have better performance.

Second, after 5-fold cross-validation, 50 single or multi-marker combinations were retained for further evaluation. Subsequently, 100 repeated runs were performed, leading to the selection of 10 candidate combinations. As shown in **Figure 4C** and **Table S4**, the best combination within the corresponding marker number showed that increasing the number of markers did not substantially improve AUC. Therefore, a 3-marker model was selected as a parsimonious final model, *i. e.*, Model C359, comprising MTB-N-9, MTB-N-10, and MTB-N-14. The bootstrap optimism-corrected AUC based on 500 bootstrap resamples was 0.942, indicating only mild optimism. In addition, the nested cross-validation showed that MTB-N-14 was the most stable marker, MTB-N-9 and MTB-N-10 were also contributed consistently to the selected model (**Figure 4D**). In the model-development cohort, Model C359 prediction scores clearly separated ATB from HC participants (**Figure 4E**), and achieved an AUC of 0.951 (95% CI: 0.910-0.992), which was much higher than the individual markers (**Figure 4F**). At the Youden index-derived cutoff of 0.408, Model C359 achieved a sensitivity of 0.940, specificity of 0.880, and accuracy of 0.910 (**Table S4**). In addition to Model C359, other candidate models with different numbers of markers were also evaluated, and their detailed performance metrics are provided in **Table S4**. Although several more complex models showed comparable apparent or bootstrap-corrected AUCs, Model C359 achieved strong discrimination

with fewer markers and balanced sensitivity and specificity. Therefore, we selected this model as the final diagnostic model.

MTB-N-9, MTB-N-10, and MTB-N-14 in the parsimonious model mapped to Rv0080, Rv0243/FadA2, and Rv2025c, respectively (**Figure 4G**). Their predicted/localized cellular locations and solvent-accessible surface area analysis supported the structural plausibility of antibody recognition, although these estimates do not prove *in vivo* antigen exposure. These findings support the current strategy: emphasizing the 3-marker model as a compact signature.

Internal test set and exploratory cohort evaluation

We next evaluated Model C359 performance beyond the model-building analysis. In **Figure 5**, the diagnostic Model C359 was applied to an internal test set ATB versus HC cohort and achieved an AUC of 0.886 (95% CI: 0.791–0.982) for ATB versus HC (**Figure 5A**–**C**) and an exploratory AUC of 0.672 (95% CI: 0.535–0.809) for ATB versus *M. intracellulare* pulmonary disease (**Figure 5D**–**F**).

In addition, other candidate models were also evaluated in the internal test set and provided in **Figure S4**. This results further suggested that the Model C359 could achieve comparable AUC with fewer markers.

The *p* value for comparison between ATB and *M. intracellulare* pulmonary disease was 0.018. These results suggest that the serum antibody peptide signature retains diagnostic signal outside the initial model-building cohort, while the modest disease-control AUC indicates that differential diagnosis against nontuberculous mycobacterial disease remains a limitation.

The 15 peptide set spans diverse MTB protein functions

The 15 antibody-reactive peptides mapped to proteins with diverse annotations (**Table S3**). Several candidates were associated with cell-envelope or membrane-related functions, including Rv2025c, a conserved membrane protein with possible cation-transport annotation, Rv1230c, a possible membrane protein, Rv0982/MprB, a membrane-associated two-component sensor kinase, and Rv2350c/PlcB, a membrane-associated phospholipase C. Other candidates were linked to lipid or envelope metabolism, including Rv0243/FadA2, a probable acetyl-CoA acyltransferase involved in lipid degradation, and Rv2962c, a possible glycosyltransferase with reported macrophage-killing-resistance relevance.

Stress-response and host-adaptation biology was represented by Rv0982/MprB, Rv0737, Rv0080, Rv0049, and Rv2231A/VapC16. Rv0982/MprB is a strong biological anchor because MprAB is a stress-responsive two-component system that regulates stress-associated sigma factors SigB and SigE.[26–28] Rv2350c/PlcB and Rv2467/PepN provided additional virulence-associated anchors. Mycobacterial phospholipases C have been implicated in virulence in experimental infection models,[29–31] and PepN has been reported to reach host macrophage cytosol and modulate virulence-related phenotypes.[32] Rv3347c/PPE55 belonged to the PE/PPE family, a mycobacteria-enriched antigenic family frequently discussed in immune interaction, antigenic variation, and vaccine contexts.[33,34] Rv2540c/AroF mapped to the shikimate pathway, which is absent from mammals and important for microbial metabolism.[35]

Several validated candidates were conserved hypothetical or poorly characterized proteins, including Rv0049 and Rv0080, or proteins with less direct diagnostic precedent, such as Rv2886c, a probable resolvase. Their inclusion should be interpreted as hypothesis-generating rather than mechanistic proof. Collectively, the candidate set was not dominated by a single classical TB antigen. Instead, ePhIP-seq identified antibody-reactive peptides across membrane, metabolism, stress, PE/PPE, virulence-associated, and poorly characterized protein classes, supporting the value of our systematic peptide-level discovery strategy.

Discussion

In this study, we developed a workflow, ePhIP-seq, for discovering serum antibody peptide biomarkers for ATB. Our work has three main findings. First, we constructed a high-coverage H37Rv proteome-wide peptide phage library (TbMap) and confirmed its quality by dot blot and next-generation sequencing. Second, we showed that ePhIP-seq can nominate candidate ATB-associated peptide reactivities from serum. Third, we selected 15 candidate peptides using an orthogonal peptide microarray and derived a compact 3-marker model with high discrimination in the model-building analysis and further validation in a held-out internal test set cohort.

The key methodological feature of this study is the antibody-enrichment step before PhIP-seq. Serum IgG is highly diverse,[24,25] and most circulating antibodies are unrelated to MTB. Conventional Protein A/G-based immunoprecipitation can capture antigen-bound antibodies but does not selectively enrich pathogen-specific antibodies before library incubation. By using biotinylated H37Rv lysate to capture H37Rv-reactive antibodies, we attempted to increase the relative abundance of TB-relevant antibody and reduce unrelated antibody background. This strategy is conceptually similar to other enrichment steps used when the target signal is rare, but here it is applied upstream of a proteome-wide peptide-display screen. The present data support the feasibility of this approach, although the incremental benefit of antibody enrichment over unenriched PhIP-seq should be quantified more directly in future studies.

The study also addresses a longstanding limitation of TB serology: antigen selection. Historical serological tests performed poorly, leading WHO to recommend against available commercial serodiagnostic tests.[14] This does not mean that humoral immunity is irrelevant in TB. Rather, it underscores the need for systematic, well-validated antigen discovery and careful clinical evaluation. Recent work increasingly recognizes that antibodies can reflect infection state and may contribute to TB immunity or biomarker development.[36,37] Our approach differs from conventional candidate-antigen serology by sampling the H37Rv proteome at peptide resolution and by validating PhIP-seq-derived peptides in an independent microarray assay.

The final candidate set provides a useful biological signal beyond model performance. The 15 peptides mapped to proteins involved in multiple aspects of MTB biology. Membrane or cell-envelope-associated candidates may be more likely to generate antibody responses because they can be exposed, released, or processed during infection. Lipid-metabolism and cell-envelope-remodelling proteins, such as FadA2 and PlcB, are consistent with the central role of mycobacterial lipid biology in host adaptation. Stress-response proteins, especially MprB, link the signature to cell-envelope stress and host-environment adaptation. PE/PPE family proteins, represented

by PPE55, are biologically plausible antibody targets because this family is widely discussed in immune modulation, antigenic variation, and host-pathogen interaction. Poorly characterized proteins in the signature are also important: they suggest that an unbiased peptide-level approach can identify non-classical antigens that would not necessarily be prioritized by conventional antigen selection.

Among the three markers in the parsimonious model, MTB-N-9 mapped to Rv0080, MTB-N-10 mapped to Rv0243/FadA2, and MTB-N-14 mapped to Rv2025c. Rv2025c is annotated as a conserved membrane protein with possible transporter function. Rv0243/FadA2 is a probable acetyl-CoA acyltransferase involved in lipid degradation, and local structural/localization analysis suggests association with cytosol, peptidoglycan-based cell wall, and plasma membrane annotations. Rv0080 was one of the poorly characterized proteins, whereas most other candidates appear to be newly reported as antibody-related TB diagnostic candidates.

The 3-marker model is attractive for translation because a smaller marker set is easier to convert into lower-complexity assays such as ELISA panels or lateral-flow-compatible formats. A compact signature may reduce cost, simplify manufacturing, and improve robustness. We therefore present the 3-marker model as a parsimonious signature derived from the 15-candidate set, while retaining the other model with more markers as the complete benchmark.

The internal test set and disease-control analyses highlight both promise and limitations. The result suggests that the diagnostic model retains signal in an internal ATB versus HC comparison and in an exploratory ATB versus *M. intracellulare* pulmonary disease. This is encouraging, as the serum antibody peptide signature retains diagnostic signal outside the initial model-building cohort. The modest AUC for ATB versus *M. intracellulare* pulmonary disease leaves room for improvement in clinical application, but is biologically interpretable. Sequence homology analysis reveals that MTB-N-10 and MTB-N-14 are significantly similar to acetyl-CoA C-acetyltransferase and cation diffusion facilitator family transporter of *M. intracellulare*, respectively. MTB-N-9 also shares identical parts with defense island system associated with restriction-modification (DISARM) anti-phage system protein DrmE domain-containing protein. Consequently, the observed phenomenon likely reflects genuine antibody recognition of evolutionarily conserved epitopes. For clinical translation, species-specific markers should be incorporated when the assay is applied in *M. intracellulare*-endemic regions.

Future work should proceed in several directions. First, the final signature should be evaluated in pre-specified validation cohorts with locked model parameters, transparent reporting of confidence intervals, sensitivity, specificity, and cutoff selection, and comparison with clinically relevant reference standards. Second, the candidates should be converted into simpler assay formats such as ELISA, multiplex bead assays, or lateral-flow-compatible prototypes. Third, ATB, latent tuberculosis infection, multidrug-resistant tuberculosis, nontuberculous mycobacterial disease, and HC cohorts should be intentionally designed to test whether peptide-level antibody signatures can stratify active tuberculosis from clinically adjacent states. Fourth, functional studies of selected proteins, particularly Rv2025c, Rv0982/MprB, Rv2350c/PlcB, Rv2467/PepN, Rv2962c, and Rv3347c/PPE55, may clarify why these peptides are immunoreactive during active disease.

In conclusion, ePhIP-seq provides a systematic strategy for discovering serum antibody peptide biomarkers for ATB. By combining pathogen-specific antibody enrichment, proteome-wide peptide display, orthogonal microarray validation, and biological annotation, this study identified a functionally diverse 15-peptide candidate set and a compact 3-marker serum antibody signature. The approach is not yet clinically deployable, but it provides a promising route for biomarker discovery in TB and potentially other infections where disease-relevant humoral signals are present but difficult to detect.

Limitations of the study

This study has several limitations. First, the sample size remains modest, and the cohorts were collected from limited geographic settings. Larger, multi-centre, prospectively collected cohorts are needed to evaluate robustness across age, sex, geography, HIV status, comorbidities, disease severity, smear status, and prior TB exposure. Second, the diagnostic model was developed in a case-control setting and requires pre-specified validation with locked model parameters before clinical interpretation. Third, the peptide phage library and peptide microarray focus primarily on linear peptide epitopes and do not capture conformational epitopes or post-translational modifications. Fourth, the current study does not establish mechanistic roles for the candidate proteins; antibody recognition should be interpreted as diagnostic association and biological plausibility rather than proof of antigen function in disease.

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Sheng-ce Tao (taosc@sjtu.edu.cn).

Materials availability

Requests for reagents should be directed to and will be fulfilled by the lead contact after signing a materials transfer agreement.

Data and code availability

- PhIP-seq raw sequencing data have been deposited in the NCBI SRA (PRJNA1515451) and are publicly available as of the date of publication.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Acknowledgements

This study was partially supported by National Key R&D Program of China (2023YFC2307200, and 2024YFA1307600), the Fundamental Research Funds for the Central Universities (YG2023ZD04), National Natural Science Foundation of China (32271492, and 82300971), Fuzhou Middle-aged Talents Research Program (2025-S-rc5), Tianshan Talents program (High_level Medical and Health personnel Training program) of the Xinjiang uygur Autonomous Region (TsYc202301B135), Prevention and Control of Emerging and Major Infectious Diseases-National Science and Technology Major Project 2025ZD01999900 (Supported by Major Project of Guangzhou National Laboratory, Grant No. GZNL2023A01005). We thank Professor Jianzhong Xi from the College of Future Technology, Peking University, for synthesizing the oligo pool encoding the whole proteome of MTB, and Professor Ting Xin from the Institute of Animal Science of Chinese Academy, Agricultural Sciences, for providing the MTB H37Rv strain lysate.

Author contributions

Conceptualization: Sheng-ce Tao, Huan Qi.

Data curation: Huan Qi, Haoyun Xu, Mingliang Ma, Jinhua Wu, Yang Li.

Formal analysis: Huan Qi, Haoyun Xu, Zigan Sha.

Funding acquisition: Sheng-ce Tao, Huan Qi, Mingxiang Huang, Jiaoxiang Wu, Quan Wang.

Investigation: Huan Qi, Haoyun Xu, Zhaowei Xu, Jiahao Xie.

Methodology: Huan Qi, Sheng-ce Tao.

Project administration: Sheng-ce Tao, Huan Qi, Shujuan Guo.

Resources: Mingxiang Huang, Jiaoxiang Wu, Chongxiang Tong, Yunhong Tan, Quan Wang, Xundi Bao, Sheng Tan.

Software: Jinhua Wu, Huan Qi.

Supervision: Sheng-ce Tao, Huan Qi.

Visualization: Haoyun Xu, Huan Qi.

Writing – original draft: Sheng-ce Tao, Huan Qi, Haoyun Xu.

Writing – review & editing: Sheng-ce Tao, Huan Qi.

Declaration of interests

Sheng-ce Tao is a co-founder of AbCode Biotechnology Co., Ltd. Sheng-ce Tao, Huan Qi, Shujuan Guo and Haoyun Xu are co-inventors on a pending patent application (202510411480.9) related to TbMap construction described in this manuscript. The other authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Banana Pro in order to suggest image concepts for the schematic diagrams **Figure 1A** and **Figure 3A**. ChatGPT was used to improve readability and language of the manuscript. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

STAR★Methods

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
ANTI-FLAG® M2 antibody, mouse monoclonal	Merck	Cat# F1804; RRID: AB_262044
IRDye® 800CW donkey anti-mouse IgG secondary antibody	LI-COR Biotech	Cat# 926-32212; RRID: AB_621847
Anti-histidine tagged antibody, clone HIS.H8	Merck	Cat# 05-949; RRID: AB_492660
Cy™3 AffiniPure® goat anti-mouse IgG, Fcγ fragment specific	Jackson ImmunoResearch	Cat# 115-165-071; RRID: AB_2338687
Cy™3 AffiniPure® goat anti-human IgG, Fcγ fragment specific	Jackson ImmunoResearch	Cat# 109-165-008; RRID: AB_2337720
Bacterial and virus strains		
Escherichia coli BLT5403	Merck	Cat# 70550-M
Escherichia coli BL21(DE3)	Sangon Biotech	Cat# B528414-0100
Escherichia phage T7	Merck	Cat# 70550-M
Chemicals, peptides, and recombinant proteins		
DIFCO™ LB broth Miller 10Kg pail	BD	Cat# 214906
Chloroform	General-reagent	Cat# G75915B
TWEEN ® 20	Merck	Cat# P2287
EcoRI-HF®	New England Biolabs	Cat# R3101V
HindIII-HF®	New England Biolabs	Cat# R3104V
Q5® hot start high-fidelity DNA polymerase	New England Biolabs	Cat# M0493L
T4 DNA ligase	New England Biolabs	Cat# M0202V
Agarose, regular	Sangon Biotech	Cat# A620014

Continued

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
Carbenicillin disodium	Sangon Biotech	Cat# A600469
dNTP 10 mM mixture	Sangon Biotech	Cat# A610056
Glycerol	Sangon Biotech	Cat# A600232
IPTG, isopropyl β-D-thiogalactopyranoside	Sangon Biotech	Cat# A600168
Magnesium sulfate, heptahydrate	Sangon Biotech	Cat# A500864
Non-fat powdered milk	Sangon Biotech	Cat# A600669
PEG 8000	Sangon Biotech	Cat# A660434
Sodium chloride	Sangon Biotech	Cat# A501218
Tris	Sangon Biotech	Cat# A600194
4SGelred, 10000× in water	Sangon Biotech	Cat# A616697
50× TAE buffer	Sangon Biotech	Cat# B548101
DL500 DNA marker	Takara	Cat# 3590
Mycobacterium tuberculosis strain H37Rv lysate	Gene Optimal	Cat# GOMY0061
IRDye® 800CW streptavidin	LI-COR Biotech	Cat# 926-32230
Bovine serum albumin	Merck	Cat# SRE0096
Glycine	Sangon Biotech	Cat# A502065
Sodium dodecyl sulfate	Sangon Biotech	Cat# A600485
Methanol	Sinopharm	Cat# 10014118
Dynabeads™ M-280 streptavidin	Thermo Fisher Scientific	Cat# 11206D
Dynabeads™ protein G	Thermo Fisher Scientific	Cat# 10004D
EZ-Link™ Sulfo-NHS-LC-biotin	Thermo Fisher Scientific	Cat# 21335
PageRuler™ prestained protein ladder	Thermo Fisher Scientific	Cat# 26616
TieChui™ E. coli lysis buffer	ACE	Cat# BR0005

Continued

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
Deposited data		
PhIP-seq sequencing data	This paper	DOI: 10.17632/pdbr2mkdnd.1
Microarray signal matrices	This paper	DOI: 10.17632/pdbr2mkdnd.1
Oligonucleotides		
Primers	This paper	See Table S2
Software and algorithms		
SnapGene	Dotmatics	https://www.snapgene.com/
GraphPad	GraphPad	https://www.graphpad.com/
BioRender	BioRender	https://www.biorender.com/
R software	R Foundation	https://cran.rproject.org
Python	Python Software Foundation	https://www.python.org/
GenePix	Molecular Devices	N/A
Others		
T7Select®10-3 cloning kit	Merck	Cat# 70550-M
Fast silver stain kit	Beyotime	Cat# P0017S
FlaPure gel purification kit	Genesand	Cat# GE706
PATH® protein microarray slides, no barcode	Grace Bio-Labs	Cat# GBL805025
Zeba™ spin desalting columns, 7K MWCO	Thermo Fisher Scientific	Cat# 89882
Odyssey® infrared imaging system	LI-COR Biotech	N/A
Marathon argus microarray spotter	Arrayjet	N/A
LuxScan	CapitalBio	N/A

Experimental model and study participant details

Study design

This was a retrospective biomarker discovery and validation study designed to identify serum IgG-reactive peptide biomarkers for active pulmonary tuberculosis. The workflow consisted of four major stages: (1) construction and quality control of an H37Rv proteome-wide phage-displayed peptide library; (2) enrichment-assisted PhIP-seq discovery of candidate antibody-binding peptides in an active tuberculosis (ATB) versus healthy control (HC) discovery cohort; (3) orthogonal validation of selected candidates using a peptide microarray; and (4) diagnostic-model construction and evaluation using independent disease-control cohorts.

Participants and serum samples

As shown in **Table 1**, all the serum samples were assigned to four cohorts according to their primary purpose: a discovery cohort for candidate marker discovery by ePhIP-seq, comprising 30 ATB patients and 30 HCs; a model-development cohort for final model construction by peptide microarray, comprising pre-specified 50 ATB patients and 50 HCs; a held-out internal test set for evaluation the final model performance, comprising pre-specified 27 ATB patients and 27 HCs; exploratory cohort for evaluation the final model ability for differentiating the ATB from *M. intracellulare* infection, comprising 32 ATB patients and 32 patients with *M. intracellulare* infection. Age and sex in every cohort were matched between groups.

Table 1. Overview of study cohorts and analytical roles.

Cohort	Platform	ATB	HC	*M. intracellulare* infection	Main purpose
Discovery cohort	ePhIP-seq	30	30	0	Candidate marker discovery
Model-development cohort	Peptide microarray	50	50	0	Model construction
Internal test set cohort	Peptide microarray	27	27	0	Evaluation the final model performance on ATB vs HC
Exploratory cohort	Peptide microarray	32	0	32	Evaluation the final model ability for differentiating the ATB from *M. intracellulare* infection

Serum samples from ATB patients and patients with *M. intracellulare* infection were collected from Fuzhou Pulmonary Hospital between April 2024 to March 2025. ATB patients met the diagnostic criteria of the Chinese health industry standard WS 288-2017 for pulmonary tuberculosis and were confirmed to be susceptible to rifampicin and isoniazid by phenotypic drug-susceptibility testing. Patients were newly

diagnosed, had no previous history of TB, and had not received anti-TB therapy before blood collection. Patients with *M. intracellulare* infection were diagnosed based on melting curve analysis. The HCs were collected from Tongren Hospital affiliated to Shanghai Jiao Tong University School of Medicine between March to October 2021. HCs had no history of TB, acute infection, or uncontrolled chronic disease. The study was approved by the ethics review committee for human research at Shanghai Jiao Tong University, approval number B20230338I.

Method details

H37Rv proteome-wide peptide library design

Annotated protein sequences of H37Rv were used to design overlapping peptides. Each peptide was 30 amino acids in length, with a 15-amino-acid overlap between adjacent peptides. Based on 4,025 H37Rv protein sequences, 86,287 peptides were designed. The amino acid sequence information of all the peptides was in **Table S1**. Oligonucleotides consisting the peptide-coding region flanked by the left consensus sequences 5'-GGGAATTCCGCTGCG-3' and the right consensus sequence 5'-GATTACAAGGATGAC-3' were synthesized via the CustomArray's OligoArray platform and cloned into a T7Select 10-3b phage-display system (Merck, cat. no. 70550-M). The H37Rv reference strain and gene annotations were based on H37Rv genome resources and cross-checked against Mycobrowser annotations where needed.[38,39]

Library construction and quality control

The oligonucleotide pool was amplified with a high-template, low-cycle PCR strategy to reduce amplification bias. Restriction sites, a FLAG tag, and stop codons were introduced through extension PCR with the primers Oligo-MTB-F and Oligo-MTB-R (sequences are available in **Table S2**). PCR products were analyzed by 2.5% agarose gel electrophoresis, and band corresponding to the expected size of 182 bp was excised and purified. The purified products were digested with *Eco*R I and *Hin*d III and purified again. The digestion efficiency was subsequently evaluated by TA cloning. Only purified products that had been completely digested were ligated into T7 phage vector arms, packaged *in vitro*, amplified using plate-based amplification, purified, and stored for downstream ePhIP-seq. Library titer was determined by plaque assay.

Initially, the quality of MTB phage display library was assessed by dot blot using anti-FLAG antibody (Merck, cat. no. F1804, RRID: AB_262044). To comprehensively evaluate the library quality, sequencing libraries for NGS were prepared by two rounds of PCR. Briefly, the peptide-coding region was amplified from the library using one S-series primer and one N-series primer. The primer sequences are listed in **Table S2**, with primer IDs containing "S" or "N" indicating the corresponding primer series. The resulting PCR products were subsequently indexed with one Illumina index combinations through extension PCR. The final PCR product was analyzed by 2.5% agarose gel electrophoresis, and band corresponding to the expected size of 325 bp was excised and purified. Subsequently, the purified PCR product was sequenced on Illumina NovaSeq platform using paired end sequencing. Reads were filtered, paired reads were merged by sample identifiers, and only reads with fully consistent paired-end sequences and expected library structure were

retained for analysis. Nucleotide sequences were translated into peptide sequences and matched to the peptide set.

Sequencing depth was evaluated using Chao1-based diversity estimation.[17,39] Clone abundance distribution, peptide-level coverage, and protein-level coverage were calculated from next-generation sequencing data.

Biotinylation of H37Rv lysate and enrichment of H37Rv-reactive serum antibodies

To enrich antibodies reactive with MTB, H37Rv total lysate (Gene Optimal, cat. no. GOMY0061-1 mg) was biotinylated according to the manufacturer protocol (Thermo Fisher Scientific, cat. no. 21335). Free biotin was removed using desalting columns (Thermo Fisher Scientific, cat. no. 89882). For each sample, 2 μg biotinylated H37Rv proteins were mixed with 10 μL streptavidin magnetic beads (Thermo Fisher Scientific, cat. no. 11206D). Then the beads were washed and resuspended in 30 μL TBST. Subsequently, incubated with 30 μL serum to capture H37Rv-reactive antibodies. After washing with TBST, bound antibodies were eluted under acidic conditions and neutralized before PhIP-seq. Biotinylation, bead coupling, and antibody enrichment were evaluated by SDS-PAGE or Western blot.

PhIP-seq screening and candidate selection

The enriched antibodies were incubated overnight at 4°C with 10 μL phage library (8×10^7 pfu), which corresponded to approximately 1,000-fold coverage of the library diversity. Antibody-bound phages were captured using 10 μL Protein G beads (Thermo Fisher Scientific, cat. no. 10004D) for 4 h at 4°C. The beads were then washed sequentially with TBST and ultrapure water. Negative controls containing only the phage library and Protein G beads were included in parallel. Phage DNA inserts from immunoprecipitated samples were amplified, indexed, pooled, and sequenced as previously described. In the first-round PCR, each sample was amplified using a unique combination of one S-series and one N-series primer. All the first-round PCR products were pooled and purified as the template for the second-round PCR. After purification, the product of second-round PCR was sequenced on the Illumina NovaSeq platform using paired end sequencing.

Raw sequencing data was filtered and aligned to the phage library. Sequencing reads from each serum sample were normalized to a total of 100,000 reads across all peptides. Library quality-control sequencing data were normalized similarly and used as the input abundance for each peptide. Candidate peptides were retained if the normalized read count for a given peptide in every ATB sample exceeded its corresponding input abundance, showed higher reads in ATB patients than in HCs, and had a *t*-test p value < 0.05. Information on the corresponding source proteins and available structures were then retrieved; when structures were unavailable, AlphaFold3-predicted structures were used. The solvent-accessible surface area (SASA) of each peptide region was calculated in PyMOL using the get_area function. Potential serum biomarker peptides were prioritized based on literature support,[40,41] peptide overlap within the same protein, SASA, input abundance, and statistical significance.

Recombinant peptide-rHF antigen production

The 21 selected peptide sequences were expressed as fusion proteins with the recombinant human ferritin heavy chain protein (rHF) in expression vector pET-32a. An additional His tag was appended to the N-terminus of final recombinant protein. Recombinant proteins were produced in *Escherichia coli*. Briefly, the bacteria were collected after induction with 1 mM IPTG for 4 h at 37°C and lysed with lysis buffer (ACE, cat. no. BR0005-02). Inclusion bodies were collected and resuspended in denaturing buffer (containing 20 mM Tris-HCl, 500 mM NaCl, and 3% SDS, pH 8.0) for 1 h at 37°C. The supernatant was incubated with Ni NTA Beads 6FF (Smart-Lifesciences, cat. no. SA005100) for 2 h at 37°C. To remove impurities and facilitate the reassembly of rHF, the agarose beads were washed sequentially with denaturing buffer and renaturing buffer (containing 20 mM Tris-HCl and 500 mM NaCl, pH 8.0). The proteins were eluted with elution buffer (containing 20 mM Tris-HCl, 500 mM NaCl, 500 mM imidazole, and 20% glycerol, pH 8.0), and subsequently assessed by SDS-PAGE and Western blot. Negative control and rHF were included as control proteins where appropriate.

Peptide microarray fabrication and serum IgG detection

Purified peptide-rHF fusion proteins and control proteins were printed onto PATH slides (Grace Bio-Labs, Oregon, USA) via Super Marathon printer (ArrayJet, UK). Printing quality was assessed using antibody against His tag (Merck, cat. no. 05-949, RRID: AB_492660) and corresponding fluorescent secondary antibody (LI-COR Biotech, cat. no. 926-32212, RRID: AB_621847). Because serum samples showed background reactivity against the rHF carrier protein, an rHF-reactive antibody-removal step was developed. Sera were diluted and pre-absorbed using rHF-immobilized agarose beads before microarray incubation. Microarrays were blocked, incubated with pretreated sera, washed, probed with fluorescent anti-human IgG detection antibody, scanned, and quantified with LuxScan-10K/A (CapitalBio, China) as previously described with minor modifications.[42] Data were extracted using GenePix Pro 6.0 software (Molecular Devices, CA, USA).

Diagnostic model construction

Among the microarray signals, 15 of 21 peptides were higher in the ATB group and were therefore retained for further analysis. Logistic regression models were constructed using all possible one- to five-marker combinations, resulting in 4,943 candidate models.

All candidate models were first screened using five-fold cross-validation. Models were ranked separately according to the number of included markers, and the top 10 models for each marker number were retained. These models were then evaluated using 100 repeated stratified 7:3 split validations. In each repetition, 70% of samples were used for model fitting and 30% for validation, and validation AUCs were recorded. For each marker number, the two models with the best overall performance were retained, resulting in 10 final candidate models.

The final model was selected by considering discrimination, stability, calibration, prediction error, and model simplicity. Performance was assessed using AUC, sensitivity, specificity, and accuracy. The optimal cutoff for the final logistic regression model was determined by maximizing the Youden index, defined as

sensitivity + specificity − 1. At this cutoff, classification-based performance metrics were calculated, including sensitivity, specificity, and accuracy. Bootstrap optimism correction with 500 resamples, nested cross-validation, and held-out internal test set cohort evaluation were performed to assess optimism, robustness, and generalizability.

Functional annotation of validated candidate peptides

The final 15 peptide candidates were mapped to H37Rv proteome using the peptide-to-protein mapping. Functional annotations were compiled from Mycobrowser and UniProt, and focused literature sources. Candidate proteins were grouped into biological themes: cell-envelope or membrane-associated proteins, lipid or envelope metabolism, stress adaptation and regulatory proteins, virulence or intracellular-survival-associated proteins, PE/PPE family proteins, and non-classical or poorly characterized proteins. Functional annotation was used to support biological interpretation but not to infer causality from antibody recognition.

Statistical analysis

Continuous variables were summarized as mean and SD or median and interquartile range, as appropriate. Group comparisons for clinical covariates used two-sided *t*-tests or appropriate non-parametric tests as indicated. Peptide reactivity differences were assessed using *t*-tests in PhIP-seq discovery and Mann-Whitney tests in microarray validation, with multiple-testing correction where indicated. AUCs and corresponding 95% confidence intervals were calculated using the built-in nonparametric method in GraphPad Prism. All tests were two-sided unless otherwise specified.

References

1. World Health Organization (2025). Global tuberculosis report 2025 (World Health Organization).

2. Trajman, A., Campbell, J.R., Kunor, T., Ruslami, R., Amanullah, F., Behr, M.A., and Menzies, D. (2025). Tuberculosis. Lancet *405*, 850–866. https://doi.org/10.1016/S0140-6736(24)02479-6.

3. Nathavitharana, R.R., Garcia-Basteiro, A.L., Ruhwald, M., Cobelens, F., and Theron, G. (2022). Reimagining the status quo: How close are we to rapid sputum-free tuberculosis diagnostics for all? eBioMedicine *78*, 103939. https://doi.org/10.1016/j.ebiom.2022.103939.

4. Broger, T., Koeppel, L., Huerga, H., Miller, P., Gupta-Wright, A., Blanc, F.-X., Esmail, A., Reeve, B.W.P., Floridia, M., Kerkhoff, A.D., et al. (2023). Diagnostic yield of urine lipoarabinomannan and sputum tuberculosis tests in people living with HIV: A systematic review and meta-analysis of individual participant data. Lancet Glob Health *11*, e903–e916. https://doi.org/10.1016/S2214-109X(23)00135-3.

5. Walzl, G., McNerney, R., du Plessis, N., Bates, M., McHugh, T.D., Chegou, N.N., and Zumla, A. (2018). Tuberculosis: Advances and challenges in development of new diagnostics and biomarkers. Lancet Infect Dis *18*, e199–e210. https://doi.org/10.1016/S1473-3099(18)30111-7.

6. Acharya, B., Acharya, A., Gautam, S., Ghimire, S.P., Mishra, G., Parajuli, N., and Sapkota, B. (2020). Advances in diagnosis of tuberculosis: An update into molecular diagnosis of *Mycobacterium tuberculosis*. Mol Biol Rep *47*, 4065–4075. https://doi.org/10.1007/s11033-020-05413-7.

7. Steingart, K.R., Henry, M., Ng, V., Hopewell, P.C., Ramsay, A., Cunningham, J., Urbanczik, R., Perkins, M., Aziz, M.A., and Pai, M. (2006). Fluorescence versus conventional sputum smear microscopy for tuberculosis: A systematic review. Lancet Infect Dis *6*, 570–581. https://doi.org/10.1016/S1473-3099(06)70578-3.

8. World Health Organization (2014). Xpert MTB/RIF implementation manual: Technical and operational “how-to”: Practical considerations (World Health Organization).

9. World Health Organization (2025). WHO consolidated guidelines on tuberculosis. Module 3: Diagnosis (World Health Organization).

10. World Health Organization (2024). Target product profile for tuberculosis diagnosis and detection of drug resistance (World Health Organization).

11. Steingart, K.R., Henry, M., Laal, S., Hopewell, P.C., Ramsay, A., Menzies, D., Cunningham, J., Weldingh, K., and Pai, M. (2007). Commercial serological antibody detection tests for the diagnosis of pulmonary tuberculosis: A systematic review. PLoS Med *4*, e202. https://doi.org/10.1371/journal.pmed.0040202.

12. Steingart, K.R., Henry, M., Laal, S., Hopewell, P.C., Ramsay, A., Menzies, D., Cunningham, J., Weldingh, K., and Pai, M. (2007). A systematic review of commercial serological antibody detection tests for the diagnosis of extrapulmonary tuberculosis. Thorax *62*, 911–918. https://doi.org/10.1136/thx.2006.075754.

13. Steingart, K.R., Flores, L.L., Dendukuri, N., Schiller, I., Laal, S., Ramsay, A., Hopewell, P.C., and Pai, M. (2011). Commercial serological tests for the diagnosis of active pulmonary and extrapulmonary tuberculosis: An updated systematic review and meta-analysis. PLoS Med *8*, e1001062. https://doi.org/10.1371/journal.pmed.1001062.

14. World Health Organization (2011). Commercial serodiagnostic tests for diagnosis of tuberculosis: Policy statement (World Health Organization).

15. Land, K.J., Boeras, D.I., Chen, X.-S., Ramsay, A.R., and Peeling, R.W. (2019). REASSURED diagnostics to inform disease control strategies, strengthen health systems and improve patient outcomes. Nat Microbiol *4*, 46–54. https://doi.org/10.1038/s41564-018-0295-3.

16. Song, L., Wallstrom, G., Yu, X., Hopper, M., Van Duine, J., Steel, J., Park, J., Wiktor, P., Kahn, P., Brunner, A., et al. (2017). Identification of antibody targets for tuberculosis serology using high-density nucleic acid programmable protein arrays. Mol Cell Proteomics *16*, S277–S289. https://doi.org/10.1074/mcp.M116.065953.

17. Larman, H.B., Zhao, Z., Laserson, U., Li, M.Z., Ciccia, A., Gakidis, M.A.M., Church, G.M., Kesari, S., LeProust, E.M., Solimini, N.L., et al. (2011). Autoantigen discovery with a synthetic human peptidome. Nat Biotechnol *29*, 535–541. https://doi.org/10.1038/nbt.1856.

18. Mohan, D., Wansley, D.L., Sie, B.M., Noon, M.S., Baer, A.N., Laserson, U., and Larman, H.B. (2018). PhIP-Seq characterization of serum antibodies using oligonucleotide-encoded peptidomes. Nat Protoc *13*, 1958–1978. https://doi.org/10.1038/s41596-018-0025-6.

19. Bodansky, A., Mettelman, R.C., Sabatino, J.J., Vazquez, S.E., Chou, J., Novak, T., Moffitt, K.L., Miller, H.S., Kung, A.F., Rackaityte, E., et al. (2024). Molecular mimicry in multisystem inflammatory syndrome in children. Nature *632*, 622–629. https://doi.org/10.1038/s41586-024-07722-4.

20. Xu, G.J., Kula, T., Xu, Q., Li, M.Z., Vernon, S.D., Ndung’u, T., Ruxrungtham, K., Sanchez, J., Brander, C., Chung, R.T., et al. (2015). Comprehensive serological profiling of human populations using a synthetic human virome. Science *348*, aaa0698. https://doi.org/10.1126/science.aaa0698.

21. Monaco, D.R., Sie, B.M., Nirschl, T.R., Knight, A.C., Sampson, H.A., Nowak-Wegrzyn, A., Wood, R.A., Hamilton, R.G., Frischmeyer-Guerrerio, P.A., and Larman, H.B. (2021). Profiling serum antibodies with a pan allergen phage library identifies key wheat allergy epitopes. Nat Commun *12*, 379. https://doi.org/10.1038/s41467-020-20622-1.

22. Zamecnik, C.R., Sowa, G.M., Abdelhak, A., Dandekar, R., Bair, R.D., Wade, K.J., Bartley, C.M., Kizer, K., Augusto, D.G., Tubati, A., et al. (2024). An autoantibody signature predictive for multiple sclerosis. Nat Med *30*, 1300–1308. https://doi.org/10.1038/s41591-024-02938-3.

23. Sundell, G.N., and Tao, S.-C. (2024). Phage immunoprecipitation and sequencing—A versatile technique for mapping the antibody reactome. Mol Cell Proteomics *23*, 100831. https://doi.org/10.1016/j.mcpro.2024.100831.

24. Perelson, A.S., and Oster, G.F. (1979). Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. J Theor Biol *81*, 645–670. https://doi.org/10.1016/0022-5193(79)90275-3.

25. Briney, B., Inderbitzin, A., Joyce, C., and Burton, D.R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature *566*, 393–397. https://doi.org/10.1038/s41586-019-0879-y.

26. He, H., Hovey, R., Kane, J., Singh, V., and Zahrt, T.C. (2006). MprAB is a stress-responsive two-component system that directly regulates expression of sigma factors SigB and SigE in *Mycobacterium tuberculosis*. J Bacteriol *188*, 2134–2143. https://doi.org/10.1128/jb.188.6.2134-2143.2006.

27. Manganelli, R., Cioetto-Mazzabò, L., Segafreddo, G., Boldrin, F., Sorze, D., Conflitti, M., Serafini, A., and Provvedi, R. (2023). SigE: A master regulator of *Mycobacterium tuberculosis*. Front Microbiol *14*, 1075143. https://doi.org/10.3389/fmicb.2023.1075143.

28. Zahrt, T.C., Wozniak, C., Jones, D., and Trevett, A. (2003). Functional analysis of the *Mycobacterium tuberculosis* MprAB two-component signal transduction system. Infect Immun *71*, 6962–6970. https://doi.org/10.1128/iai.71.12.6962-6970.2003.

29. Raynaud, C., Guilhot, C., Rauzier, J., Bordat, Y., Pelicic, V., Manganelli, R., Smith, I., Gicquel, B., and Jackson, M. (2002). Phospholipases C are involved in the virulence of *Mycobacterium tuberculosis*. Mol Microbiol *45*, 203–217. https://doi.org/10.1046/j.1365-2958.2002.03009.x.

30. Assis, P.A., Espíndola, M.S., Paula-Silva, F.W., Rios, W.M., Pereira, P.A., Leão, S.C., Silva, C.L., and Faccioli, L.H. (2014). *Mycobacterium tuberculosis* expressing phospholipase C subverts PGE2 synthesis and induces necrosis in alveolar macrophages. BMC Microbiol *14*, 128. https://doi.org/10.1186/1471-2180-14-128.

31. Bakala N'Goma, J.C., Schué, M., Carrière, F., Geerlof, A., and Canaan, S. (2010). Evidence for the cytotoxic effects of *Mycobacterium tuberculosis* phospholipase C towards macrophages. Biochim Biophys Acta Mol Cell Biol Lipids *1801*, 1305–1313. https://doi.org/10.1016/j.bbalip.2010.08.007.

32. Sharma, N., Aggarwal, S., Kumar, S., Sharma, R., Choudhury, K., Singh, N., Jayaswal, P., Goel, R., Wajid, S., Yadav, A.K., et al. (2019). Comparative analysis of homologous aminopeptidase PepN from pathogenic and non-pathogenic mycobacteria reveals divergent traits. PLoS One *14*, e0215123. https://doi.org/10.1371/journal.pone.0215123.

33. Guo, F., Wei, J., Song, Y., Li, B., Qian, Z., Wang, X., Wang, H., and Xu, T. (2023). Immunological effects of the PE/PPE family proteins of *Mycobacterium tuberculosis* and related vaccines. Front Immunol *14*, 1255920. https://doi.org/10.3389/fimmu.2023.1255920.

34. Zhang, Z., Dong, L., Li, X., Deng, T., and Wang, Q. (2025). The PE/PPE family proteins of *Mycobacterium tuberculosis*: Evolution, function, and prospects for tuberculosis control. Front Immunol *16*, 1606311. https://doi.org/10.3389/fimmu.2025.1606311.

35. Nunes, J.E.S., Duque, M.A., de Freitas, T.F., Galina, L., Timmers, L.F.S.M., Bizarro, C.V., Machado, P., Basso, L.A., and Ducati, R.G. (2020). *Mycobacterium tuberculosis* shikimate pathway enzymes as targets for the rational design of anti-tuberculosis drugs. Molecules *25*, 1259. https://doi.org/10.3390/molecules25061259.

36. Wang, Q., Nag, D., Baldwin, S.L., Coler, R.N., and McNamara, R.P. (2024). Antibodies as key mediators of protection against *Mycobacterium tuberculosis*. Front Immunol *15*, 1430955. https://doi.org/10.3389/fimmu.2024.1430955.

37. Miles, J.R., Lu, P., Bai, S., Aguillón-Durán, G.P., Rodríguez-Herrera, J.E., Gunn, B.M., Restrepo, B.I., and Lu, L.L. (2025). ESAT-6 and CFP-10 reactive IgG in patients with tuberculosis inhibits intracellular bacteria. Cell Rep. *44*, 116653. https://doi.org/10.1016/j.celrep.2025.116653.

38. Chitale, P., Lemenze, A.D., Fogarty, E.C., Shah, A., Grady, C., Odom-Mabey, A.R., Johnson, W.E., Yang, J.H., Eren, A.M., Brosch, R., et al. (2022). A comprehensive update to the *Mycobacterium tuberculosis* H37Rv reference genome. Nat Commun *13*, 7068. https://doi.org/10.1038/s41467-022-34853-x.

39. Kapopoulou, A., Lew, J.M., and Cole, S.T. (2011). The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis *91*, 8–13. https://doi.org/10.1016/j.tube.2010.09.006.

40. Liu, Z., Qie, S., Li, L., Xiu, B., Yang, X., Dai, Z., Zhang, X., Duan, C., Que, H., Zhao, P., et al. (2016). Identification of novel RD1 antigens and their combinations for diagnosis of sputum smear-/culture+ TB patients. Biomed Res Int *2016*, 7486425. https://doi.org/10.1155/2016/7486425.

41. You, X., Li, R., Wan, K., Liu, L., Xie, X., Zhao, L., Wu, N., Deng, X., Wang, L., and Zeng, Y. (2017). Evaluation of Rv0220, Rv2958c, Rv2994 and Rv3347c of *Mycobacterium tuberculosis* for serodiagnosis of tuberculosis. Microb Biotechnol *10*, 604–611. https://doi.org/10.1111/1751-7915.12697.

42. Li, Y., Li, C., Guo, S., Guo, W., Jiang, H., Li, H., and Tao, S. (2020). Longitudinal serum autoantibody repertoire profiling identifies surgery-associated biomarkers in lung adenocarcinoma. eBioMedicine *53*, 102674. https://doi.org/10.1016/j.ebiom.2020.102674.

Figure and table legends

Main figure legends

Figure 1. ePhIP-seq workflow and quality control of TbMap.

(A) Overview of the TbMap construction, serum sample preparation, H37Rv-reactive antibody enrichment, phage immunoprecipitation, sequencing, and downstream data analysis workflow. (B) Dot blot assay using anti-FLAG antibody to evaluate peptide display in the constructed phage library. (C) Sequencing-based assessment of library clone distribution. (D) Peptide-level coverage of the designed H37Rv peptide library. (E) Protein-level coverage across the H37Rv proteome.

Figure 2. H37Rv-specific antibody enrichment and ePhIP-seq discovery of candidate peptides associated with active tuberculosis.

(A) Experimental validation of H37Rv lysate biotinylation and immobilization on streptavidin beads for enrichment of *M. tuberculosis*-reactive serum antibodies. (B) Assessment of H37Rv-reactive antibody capture and enrichment from representative ATB and HC serum samples. (C) Age and sex distribution of the ATB and HC participants included in the discovery analysis. (D) Sequential candidate-selection workflow used to filter ePhIP-seq signals from the designed peptide pool to 277 candidate peptides and then to 21 peptides selected for orthogonal microarray validation. (E) Discovery-stage signal summary for the selected candidate peptides in ATB and HC samples.

Figure 3. Orthogonal peptide microarray platform for validation of ePhIP-seq-derived candidate peptides.

(A) Workflow for recombinant expression, purification, microarray immobilization, and serum IgG detection of candidate peptide-rHF fusion proteins selected from the ePhIP-seq discovery stage. (B)

SDS-PAGE analysis of recombinant candidate peptide-rHF proteins and control proteins used for peptide microarray construction. (C) Representative peptide microarray image showing candidate peptide-rHF proteins, rHF carrier control, negative peptide control, BSA-Cy3 control, and blank spots.

Figure 4. Validation of candidate peptides and construction of a compact 3-marker diagnostic model.

(A) Age and sex distribution of ATB and HC participants in the model-building cohort. (B) Procedures for model construction. (C) The highest apparent AUC of combination within each marker number and corresponding optimized AUC. (D) The selected frequency of each marker in the nest cross-validation. (E) Distribution of Model C359 prediction scores in ATB and HC participants. (F) ROC analysis of Model C359 and its three component markers, MTB-N-9, MTB-N-10, and MTB-N-14, for ATB versus HC discrimination in the model-building analysis. (G) Source proteins, predicted/localized cellular location, statistical significance, and SASA values for the three markers in Model C359. The model consists of MTB-N-9 from Rv0080, MTB-N-10 from Rv0243/FadA2, and MTB-N-14 from Rv2025c.

Figure 5. Internal test set and exploratory disease-control evaluation of the 3-marker diagnostic model.

(A) Age and sex distribution of ATB and HC participants in the internal test set cohort. (B) Distribution of Model C359 prediction scores in the internal test set ATB and HC samples. (C) ROC curve for Model C359 in the internal test set ATB versus HC test cohort. (D) Age and sex distribution of ATB and participants with *M. intracellulare* infection in the exploratory disease-control cohort. (E) Distribution of Model C359 prediction scores in ATB and participants with *M. intracellulare* infection. (F) ROC curve for Model C359 in the exploratory ATB versus participants with *M. intracellulare* infection.

Supplementary figure legends

Figure S1. Sequencing-depth and library-complexity assessment of TbMap.

Relationship between sequencing read depth and estimated clone number in TbMap. The curve supports adequate sequencing depth for library characterization and provides additional quality-control evidence for the library summarized in **Figure 1**.

Figure S2. Extended candidate antigen quality control and rHF-background reduction for the peptide microarray assay.

(A) Western blot analysis of purified candidate peptide-rHF fusion proteins and control proteins used in the peptide microarray. (B) Representative microarray images showing rHF-reactive signal after serial rHF-based serum pretreatment or dilution conditions. (C) Quantification of rHF-associated background under the tested conditions. These experiments support antigen-production quality control and assay optimization for the orthogonal peptide microarray validation workflow shown in **Figure 3**.

Figure S3. The fluorescence intensity of 21 peptides in the model-development cohort.

Figure S4. The AUC of 10 candidate models in the internal test set cohort.

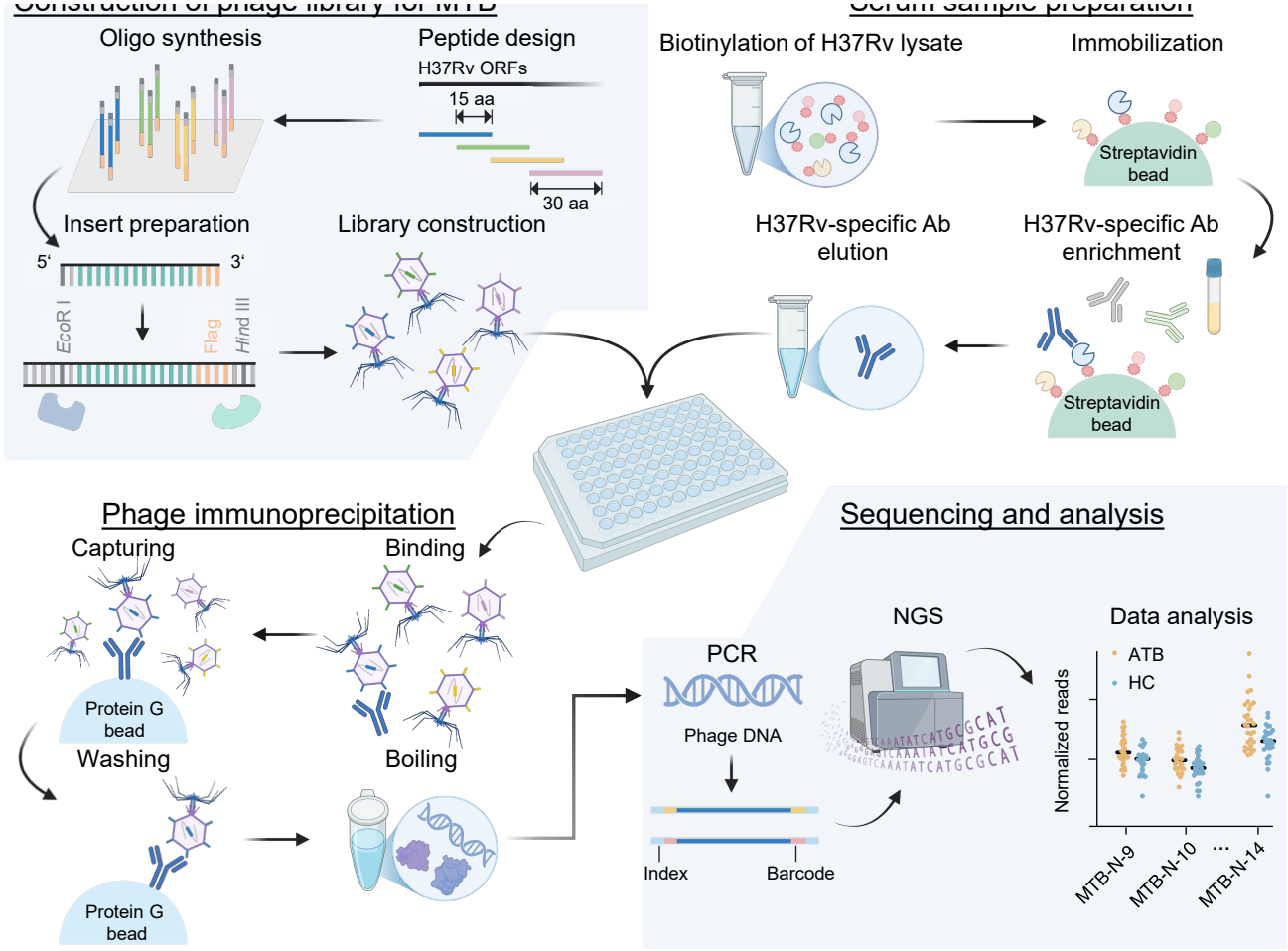
Supplementary table legends

Table S1. The amino acid sequences of the peptides in TbMap.

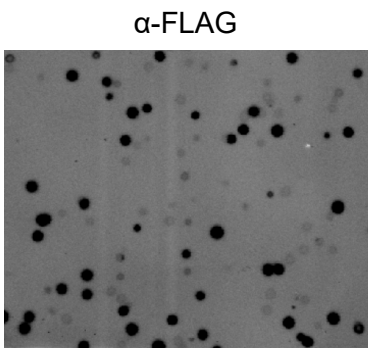
Table S2. Primers used in this study.

Table S3. Properties of 21 candidate peptides identified by ePhIP-seq after H37Rv-reactive antibody enrichment.

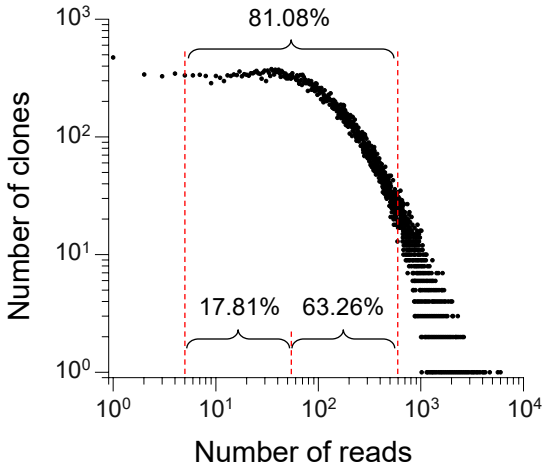
Table S4. The characteristics of the 10 candidate models.



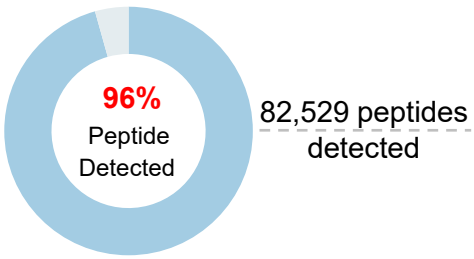
B



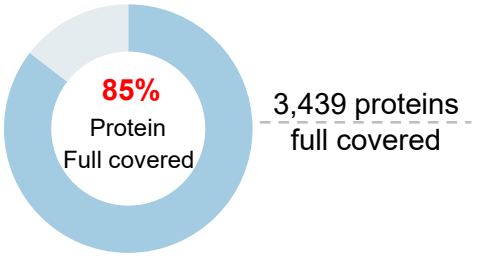
C



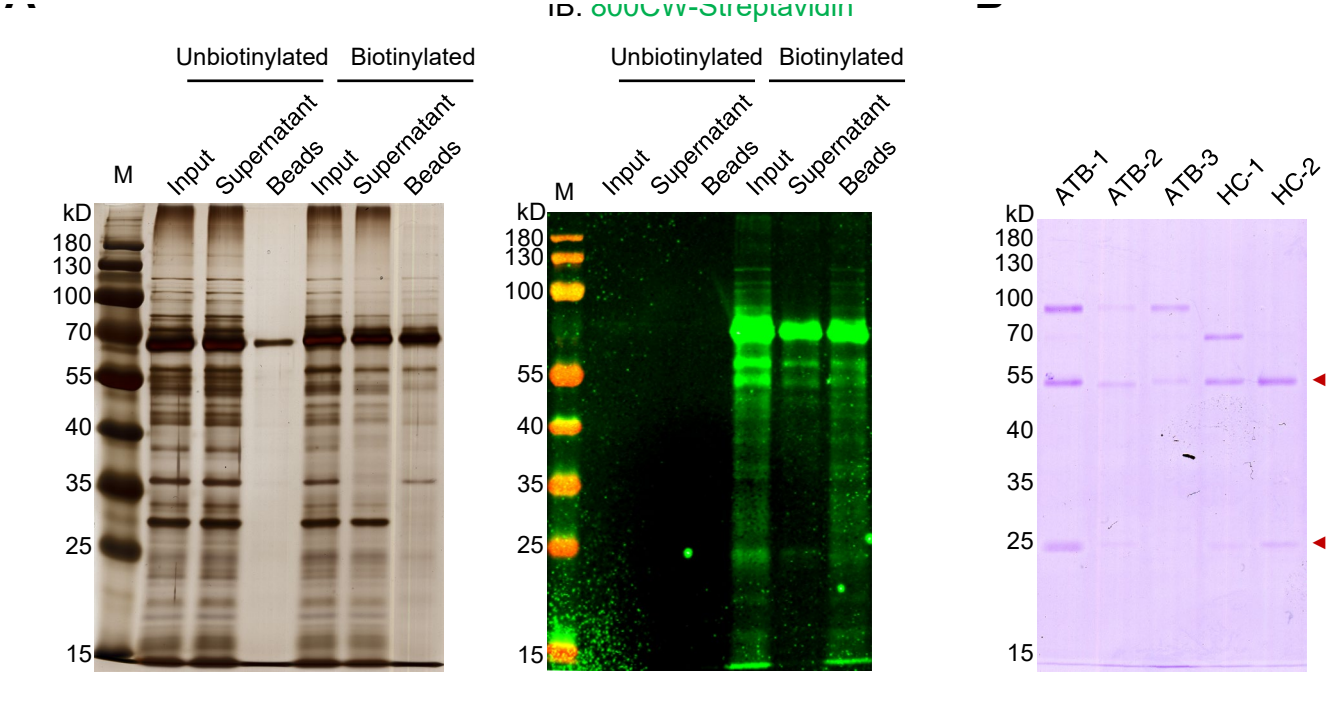
D

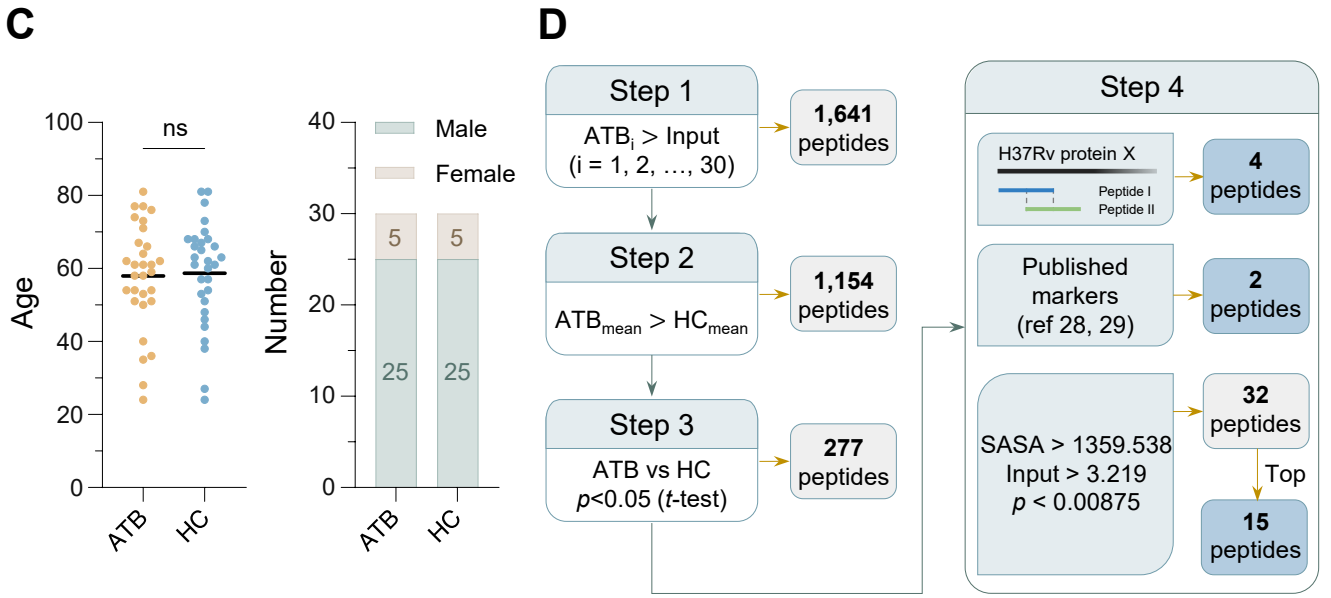


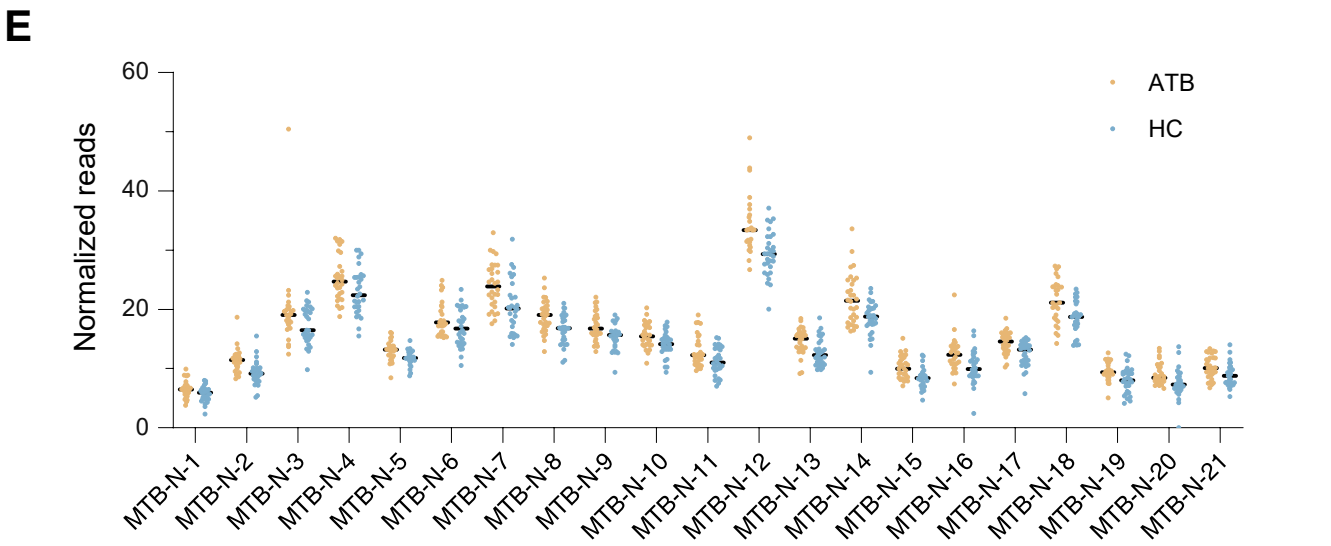
E



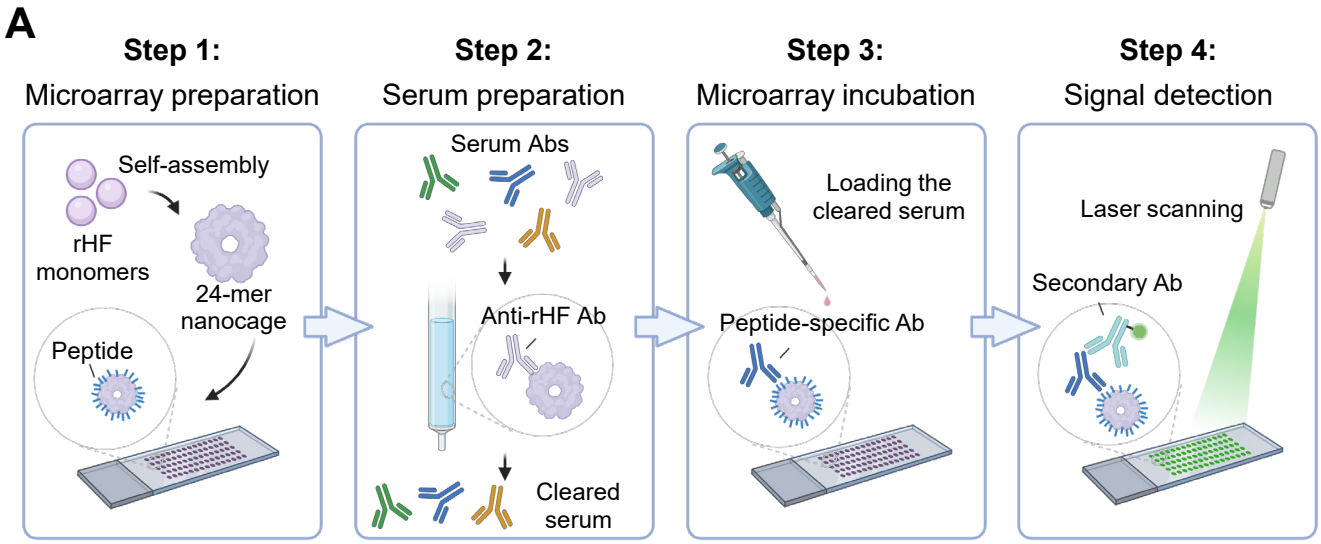
Xu *et al.*, Figure 1

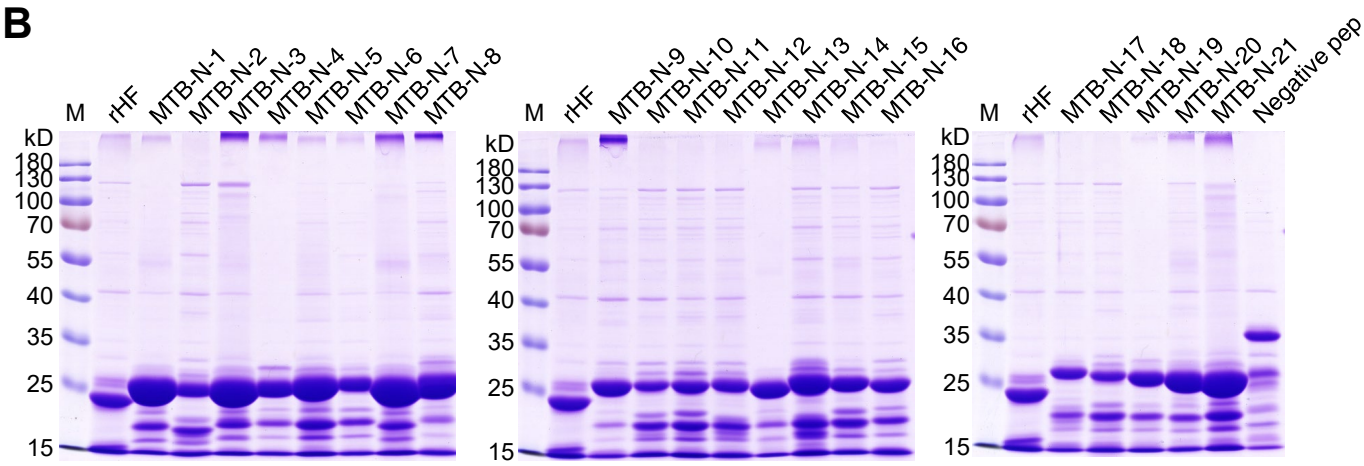


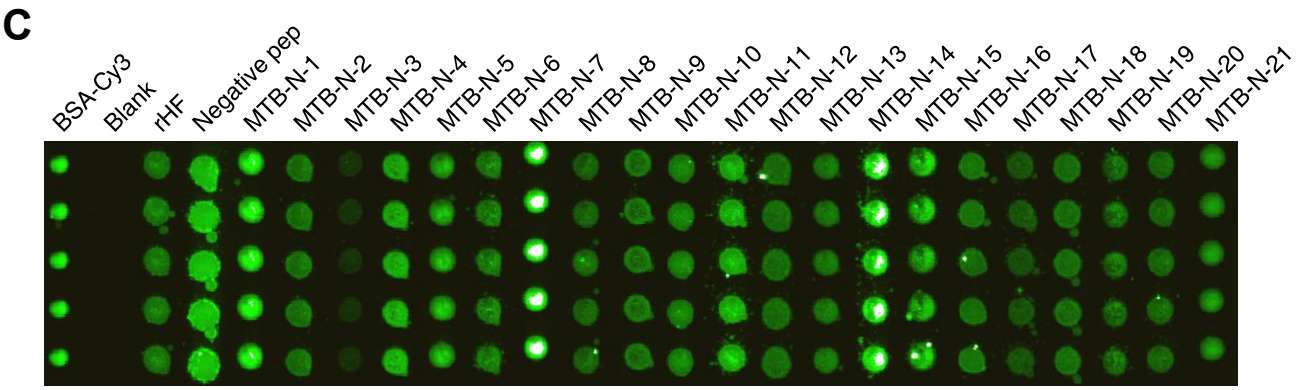




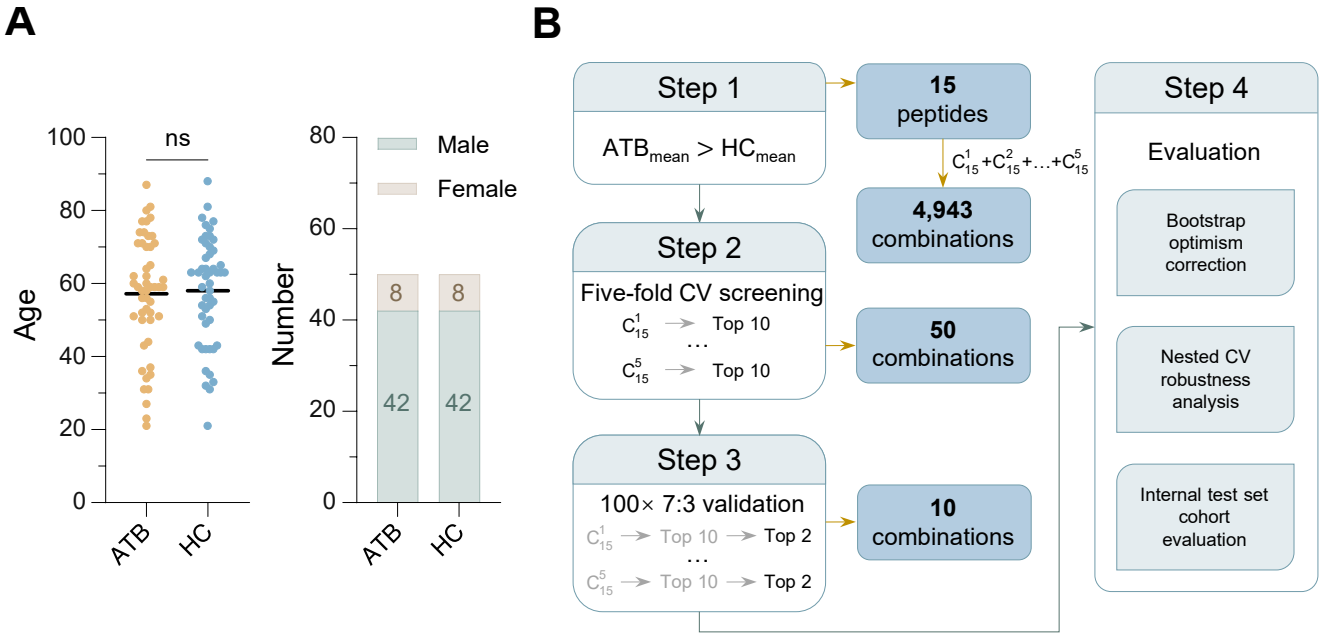
Xu *et al.*, Figure 2

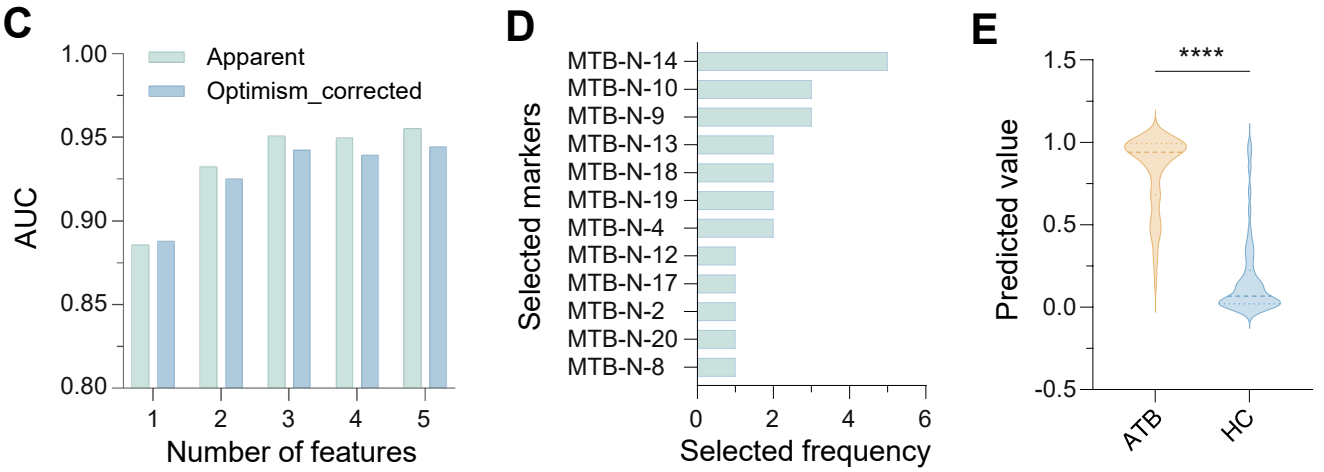


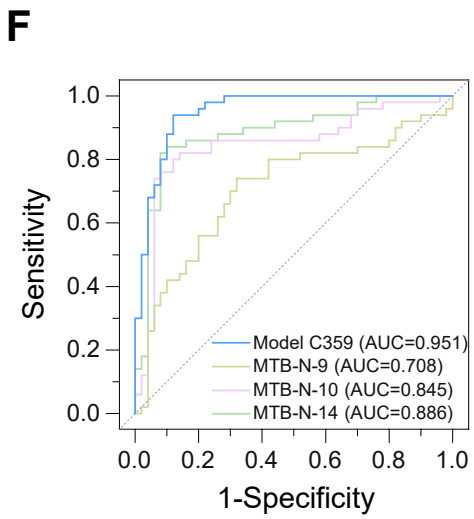




Xu *et al.*, Figure 3



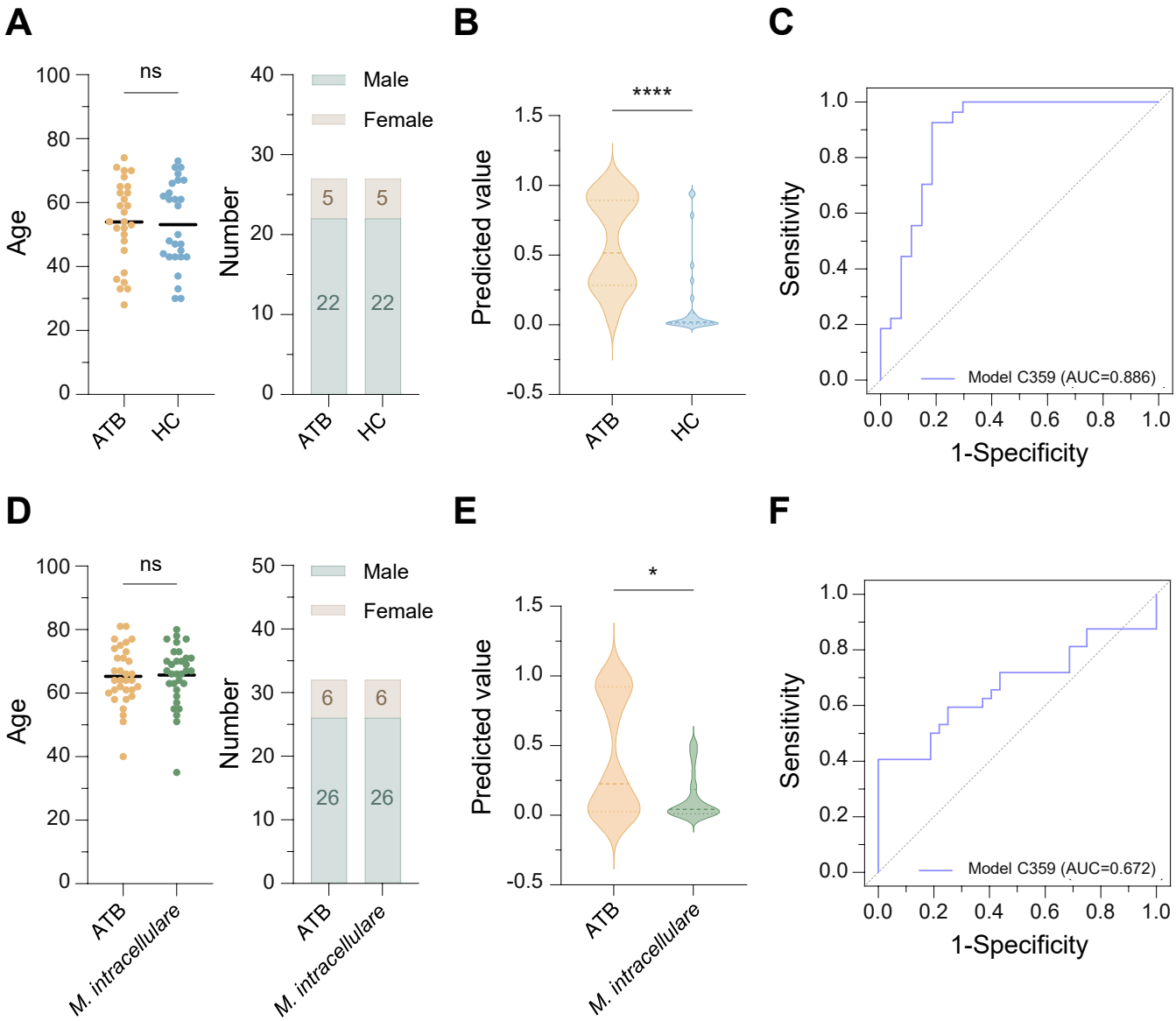




G

Peptide	Protein	Location	*p* value	SASA
MTB-N-9	Rv0080	ND	0.006	1907.232
MTB-N-10	Rv0243	Cytosol, peptidoglycan-based cell wall, plasma membrane	<0.0001	1858.423
MTB-N-14	Rv2025c	Cell membrane	<0.0001	1623.897

Xu *et al.*, Figure 4



Xu *et al.*, Figure 5