Pozzoli *et al. Genome Biology* (2025) 26:387
https://doi.org/10.1186/s13059-025-03854-6

Genome Biology

RESEARCH Open Access

Mycobacterium tuberculosis uses intrinsically disordered, fast evolving proteins to interact with conserved host factors

Uberto Pozzoli[1*], Diego Forni[1], Federica Arrigoni[2], Rachele Cagliani[1], Luca De Gioia[2] and Manuela Sironi[3*]

*Correspondence:
Uberto Pozzoli
uberto.pozzoli@lanostrafamiglia.it
Manuela Sironi
manuela.sironi@unimib.it
[1]Computational Biology Unit, Scientific Institute IRCCS E. MEDEA, Bosisio Parini 23842, Italy
[2]Department of Biotechnology and Biosciences, University of Milan-Bicocca, Milan 20126, Italy
[3]School of Medicine and Surgery, University of Milano-Bicocca, Monza 20900, Italy

Abstract

Background Intrinsically disordered protein regions (IDRs) are implicated in diverse cellular processes in eukaryotes and, in these organisms, they cover up to 40% of the proteome. Surprisingly little is known about IDRs in bacterial proteomes. Specifically, a number of questions remain unanswered, such as the role of these regions in host–pathogen interactions, their adaptive potential and evolutionary trajectories, as well as their biophysical properties. Here we focus on *Mycobacterium tuberculosis* and take advantage of the fact that, due to its extreme epidemiological relevance, several large-scale analyses are available.

Results After benchmarking different disorder prediction tools, we integrate multiple levels of biological information to show that IDR-containing proteins are involved in virulence, in the modulation of host immune response, and in lipid metabolism. *Mycobacterium tuberculosis* IDRs are fast evolving and poorly antigenic, and they display specific sequence-ensemble-function relationships. Conversely, human proteins that interact with *Mycobacterium tuberculosis* are evolutionary constrained, widely expressed, and highly connected in the human interactome map. This indicates that the classical arms race paradigm is not universal in host–pathogen interactions. We also extend analysis to 540 human-infecting bacteria and we underscore wide variations in IDR representation and conformational properties.

Conclusions Our data point to a role of IDRs in contributing to bacterial virulence, interaction with the human host, and control of immune responses. Although this awaits experimental validation, we suggest that *Mycobacterium tuberculosis* also uses IDRs to sense and interact with its environment. Herein, we provide a database of bacterial IDRs, together with relevant parameters, for public use.

Keywords *Mycobacterium tuberculosis*, Intrinsically disordered protein regions, Host–pathogen interaction, Evolution, Conformational features, Bacterial proteomes

Background

Mycobacterium tuberculosis (Mtb), the causative agent of tuberculosis (TB), is estimated to infect about a quarter of the world population and represents a major cause of death worldwide (https://www.who.int/news-room/fact-sheets/detail/tuberculosis). Mtb prod

uces an active form of the disease (active TB, ATB) or an asymptomatic form known as latent TB (LTB), which is the most common [1, 2]. The bacilli are transmitted via aerosolization and, once in the lungs, they infect tissue-resident alveolar macrophages and other cell populations [3, 4]. The development of the host immune response can result in the elimination of the bacterium or produce a granulomatous reaction to contain the infection in a tubercle, leading to LTB [4, 5]. Infection reactivation can cause a variety of manifestations, including cavitary lung disease, the most common form. Individuals with cavitary TB transmit the bacterium more efficiently than other disease forms, suggesting that the ability of Mtb to establish latency and to subsequently reactivate in an infectious form represent a strategy to persist in human populations [4, 6–8]. In fact, Mtb cannot rely on a zoonotic reservoir for reintroduction into human populations, as it is fully adapted to its human host. The association of Mtb with humans is long-standing and the bacterium is thought to have emerged about 70,000 years ago in Africa [9].

Because of its host-specificity and transmission strategy, Mtb has evolved a plethora of mechanisms to both evade and take advantage of host immune responses [4]. Some of these depend on specialized membrane or secreted proteins that interfere with macrophage functions. These include the PE–PPE/PGRS family of proteins, which are specific to mycobacteria and, despite an overall reductive genome evolution, have expanded considerably with the gain of virulence of Mtb [10–13]. Some PE–PPE/PGRS were reported to encompass intrinsically disordered regions (IDR), although no systematic analysis was conducted [14, 15].

IDRs have been implicated in diverse cellular processes in eukaryotes and, in these organisms, they cover up to 40% of the proteome [16]. Proteins with IDRs are often intracellular and function as interaction hubs in protein–protein interaction (PPI) networks [16, 17]. In contrast, IDRs are less abundant in bacteria and their functions are poorly understood [16, 18]. Investigation of individual IDRs in bacterial proteins revealed their involvement in core cellular processes including transcription [19], polar organization [20, 21], DNA repair [22], RNA metabolism [23], cell division [24], flagellum-dependent motility [25], and cell wall homeostasis [26], but also in mechanisms of resistance and adaptation such as toxin-antitoxin systems [27–29] protection against antimicrobial peptides [30], and biofilm formation [31]. Nonetheless, systematic analyses of IDRs in bacterial systems are missing and the function, evolution and contribution to pathogenicity of these regions remain unexplored.

Results

Differences among predictors in estimating IDR fraction in the Mtb proteome

To investigate the IDR content in the proteome of Mtb, we selected the reference proteome of strain H37Rv (3995 proteins) and we used two IDR prediction methods (Additional file 1: Table S1). Specifically, we applied Metapredict V2, a deep-learning-based approach that combines different predictors to generate consensus disorder scores, and an AlphaFold2-based method, which uses per-residue predicted local difference test (pLDDT) scores (see Methods) [32–35]. Metapredict estimated 10.5% of the Mtb proteome to be disordered, whereas the average IDR fraction predicted by AlphaFold2 amounted to 7.8%. Inspection of the proteins accounting for the differences indicated that the overwhelming majority of them belong to the PE-PGRS family (Fig. 1A): out of more than 35,000 codons predicted to be in IDRs by Metapredict but not by AlphaFold2,

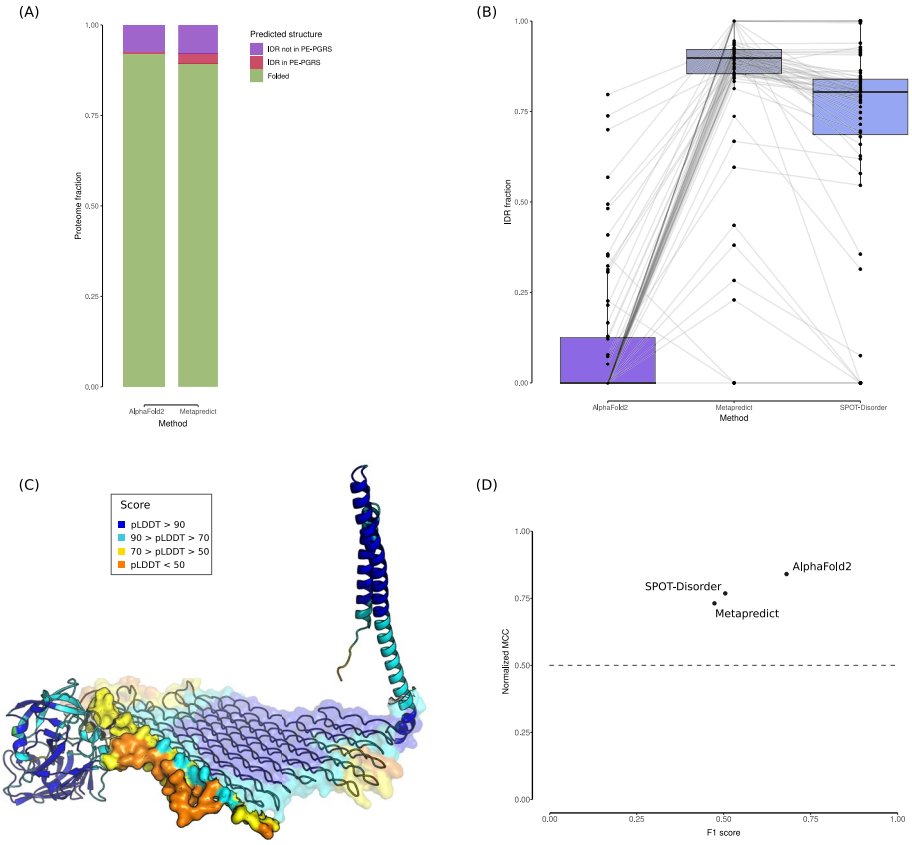

Fig. 1 Disorder prediction in the Mtb proteome. **A** Metapredict and AlphaFold2 estimates of the fraction of residues in folded domains and in IDRs that are either or not located in PE-PGRS proteins. **B** IDR fractions predicted for 63 PE-PGRS proteins using three predictors: Metapredict, AlphaFold2, and SPOT-Disorder. Data are represented as standard box and whisker plots. **C** Ribbon diagram of the AlphaFold2 predicted structure of one representative PE-PGRS protein (PE_PGRS16, Uniprot: Q79FU3). Residues are colored as per pLDDT score. For this protein, AlphaFold2 identified a 72 AA long IDR (surface representation), whereas Metapredict predicted a 550 AA long IDR covering most of the PGII helix structures (surface in transparency). **D** Comparison of IDR prediction performance for the three methods. The unit-normalized Matthews correlation coefficient (MCC) is plotted against the F1 score. The hatched line (unit-normalized MCC = 0.5) corresponds to random prediction

81% are located in these proteins. We thus used a third predictor, SPOT-Disorder [36], to estimate the IDR content in the Mtb proteome. Results were consistent with Metapredict and estimated a much higher fraction of disorder than AlphaFold2 in the sixty-three PE-PGRS proteins (Fig. 1B).

PE-PGRS proteins are specific to mycobacteria and all share a glycine-rich domain of different size. Most codons in the PGRS domains were predicted to be disordered by SPOT-Disorder and Metapredict. Instead, in AlphaFold2 models, several PGRS domain regions were predicted to fold into polyglycine type II-like (PGII) conformations, flexible left-handed extended helices, which are not constrained by intra-helix hydrogen bonding as in the case of alpha helices [37, 38] (Fig. 1C). PGII helix structures are known to be rare in the proteomes of most organisms [39], but these conformations were experimentally solved in a few cases [39–41], indicating that they exist in natural proteins. Thus, these results suggest that Metapredict and SPOT-disorder incorrectly over-estimate the IDR fraction in the Mtb proteome mainly because of their failure to predict PGII helices. More generally, these data imply that investigation of IDR content benefits

from the comparison of different methods, especially in organisms that encode unique protein families or domains.

Besides PE-PGRS proteins, the concordance among the three disorder predictions was far from complete, with Kendall's correlation coefficients of IDR fraction per protein ranging from 0.43 (AlphaFold2 vs SPOT-Disorder) to 0.57 (AlphaFold2 vs Metapredict). To benchmark the prediction methods, we searched for information on Mtb proteins with an experimentally solved structure in the Protein Data Bank (PDB) [42]. Of 594 H37Rv proteins with at least one available structure, we selected those obtained with X-ray diffraction, with structural information covering the full protein, and that represented monomers or homomultimers (see Methods) ($n = 229$). From these structures, we annotated regions with missing residues—i.e., regions that were not solved in the crystal and are thus likely to be disordered [43]. As above, residue gaps longer than 29 amino acids were considered IDRs. We identified IDRs in 21 PDB structures (208 proteins were fully folded). To evaluate the performance of the three predictors, we calculated two metrics, the Matthews correlation coefficient (MCC) and the F1 score, which were developed for binary classifiers [44–47]. Results indicated that the AlphaFold2-based method achieved a good performance and both indices were higher than those calculated for Metapredict or SPOT-Disorder (Fig. 1D).

IDRs are enriched in secretory/membrane Mtb proteins that interact with the host

Given the results above, we used AlphaFold2 predictions in all the following analyses. In the Mtb proteome, we found 1,123 proteins (28.4%) to have at least one IDR (Additional file 1: Table S1). We first asked whether IDRs are more common in bacterial essential proteins or in proteins that are dispensable for optimal growth. We found that 16.6% of essential proteins have at least one IDR, whereas this percentage amounts to 22.2% in non essential proteins, a statistically significant difference (Fisher's Exact Test, $p = 0.0005$). In fact, when we compared the fraction of protein sequence in IDRs, this was significantly higher in nonessential proteins compared to the essential ones (Brunner Munzel test, p-value $= 1.61 \times 10^{-6}$) (Fig. 2A) (Additional file 1: Table S1).

We next sought to investigate the association between IDR fraction and the biological function or cellular localization of Mtb proteins. We used Brunner Munzel tests to evaluate the significance of whether proteins associated to a given gene ontology (GO) term have a higher IDR fraction than proteins not associated to the examined term. Our GO analysis showed significant enrichments of IDR fraction in proteins that are involved in different processes, including lipid transport, lipid biosynthetic processes, autophosphorylation, and response to the host immune system. The latter term had the highest number of contributing elements. Proteins with a high IDR fraction were also more likely to function as kinases and to be either associated to the membrane or secreted (Fig. 2B). To investigate IDR location in proteins that perform the same molecular function (kinases) or are involved in the same biological processes (lipid biosynthesis or transport), we used the SMART database (https://smart.embl.de/) to map functional domains. Results indicated that in most kinases, IDRs are located within linker regions between domains or are either C- or N-terminal (Additional file 2: Fig. S1). Conversely, IDRs were mostly embedded in functional domains (especially the wax ester synthase-like Acyl-CoA acyltransferase domain) in proteins involved in lipid biosynthesis. For

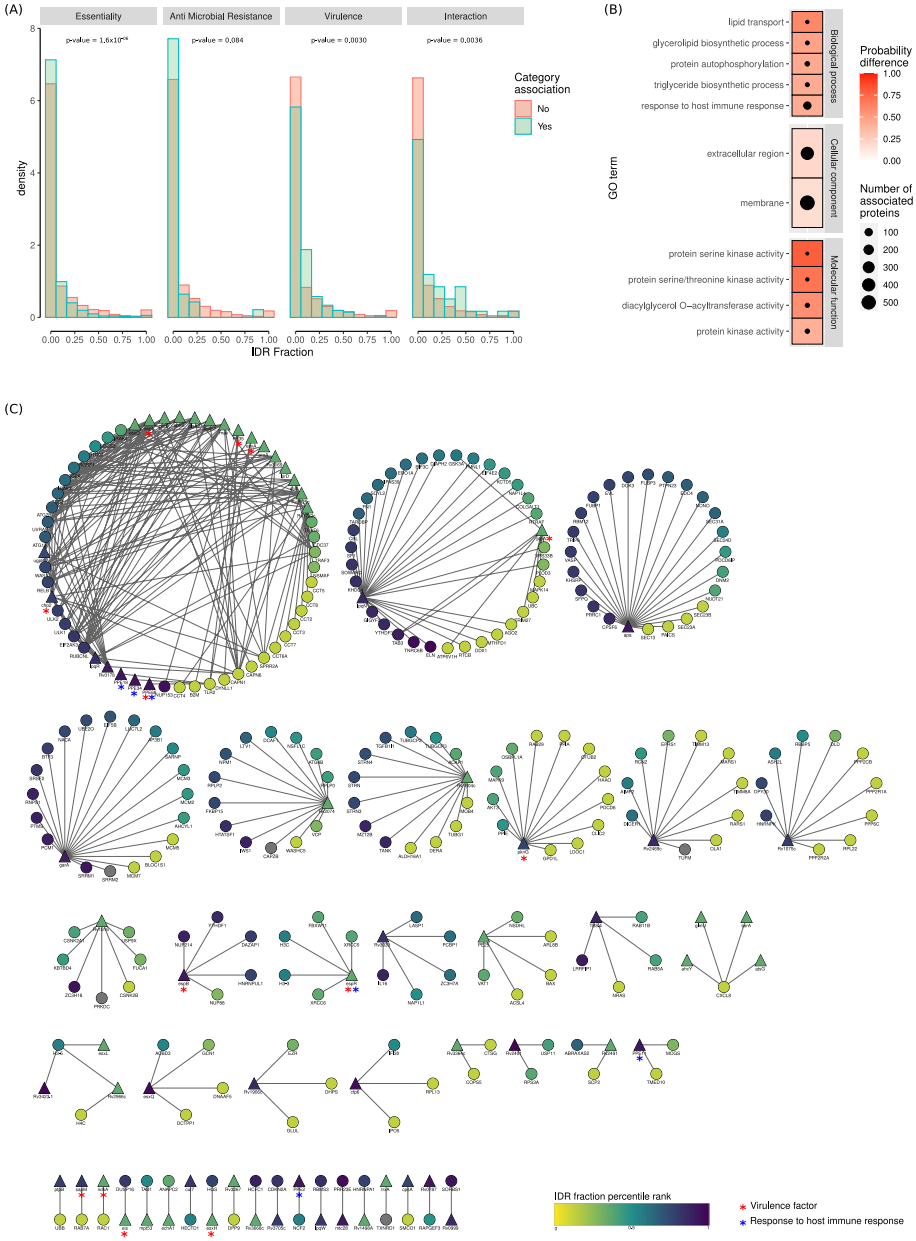

Fig. 2 Function and interactions of IDRs in the Mtb proteome. **A** Comparison of IDR fraction in different functional categories: essential vs non-essential proteins, proteins that are involved in AMR vs those that are not, proteins that contribute to Mtb virulence vs those that do not, and protein that engage in PPIs with human proteins vs those that are not known to participate to such interactions. All *p* values derive from Brunner Munzel tests. See also Table S1. **B** GO analysis of IDR fraction. Only significant terms are shown (*p* value < 0.01 after FDR correction). The red scale represents the probability difference from Brunner Munzel tests. The number of proteins associated with any term is represented by circle diameters. **C** Representation of known Mtb-human PPIs. Nodes represent proteins (triangles for Mtb and circles for human), edges indicate physical interactions deriving from different sources (see text and Methods). Proteins are color-coded based on the percentile rank of IDR fraction in the respective proteome. Interacting proteins that represent virulence factors are marked with a red asterisk, those associated with the GO term "response to host immune response" with a blue asterisk

lipid transporters, the situation was more heterogeneous in terms of both domain representation and IDR location (Additional file 2: Fig. S1).

We next sought to explore IDR content in functional classes that are not covered by GO terms. Analysis of proteins involved in antimicrobial resistance (AMR) showed that

they have similar IDR content as proteins that have never been associated with such mechanisms (Brunner Munzel test, p-value = 0.0845) (Fig. 2A, Additional file 1: Table S1). Conversely, virulence factors had a higher IDR fraction than proteins that do not contribute to the virulence of Mtb (Brunner Munzel test, p-value = 0.0030) (Fig. 2A, Additional file 1: Table S1). Some virulence factors function as kinases or in lipid metabolisms (the GO terms we identified above). We asked whether orthologs from other pathogenic bacteria shared IDR representation and location with Mtb proteins. We thus searched the proteomes of five pathogens in the order *Mycobacteriales* (*Dietzia cinnamea, Gordonia bronchialis,Corynebacterium diphtheriae, Nocardia abscessus or N. farcinica*, and *Prescottella equi*) for orthologs of ten Mtb kinases/lipid metabolism proteins involved in virulence. We detected at least one ortholog for six of them, although the paralogous mmpL8 and mmpL10 shared the same orthologs. Analysis of IDR content showed a variegated scenario: only pknG showed the same pattern across orthologs, all of them having an N-terminal IDR, whereas in the case of lipid transporters, only the Mtb and *D. cinammea* proteins had some unstructured regions. As for the mycolyltransferase (fbpC) only the orthologs from *P. equi*, *G. bronchialis*, and D.*cinnamea,* in addition to Mtb, had IDRs (Additional file 2: Fig. S2). Overall, these analyses suggest that structural disorder is volatile and poorly conserved even across orthologs of bacteria in the same order (see also below).

Finally, we investigated the contribution of IDRs to PPIs. We thus leveraged different sources (see Methods) to assemble a list of 70 Mtb proteins that participate in direct PPIs with human proteins (Additional file 3: Table S2). We found that bacterial proteins that are involved in host–pathogen PPIs have a significantly higher fraction of IDRs (Brunner Munzel test, p-value = 0.0036), in line with the GO results showing significant IDR enrichment in proteins that counteract immune responses and that are secreted or membrane-associated (Fig. 2B and C). Also, some of these proteins contribute to Mtb virulence (Fig. 2C). Conversely, when we analyzed data of physical interactions between Mtb proteins, we observed that proteins with IDRs have lower degree centrality (i.e., fewer interactions) than proteins without IDRs (Brunner Munzel test, p-value = 0.0016) (Additional file 1: Table S1). Overall, these results indicate that IDRs are likely to contribute to Mtb virulence and to modulate its interaction with the human host.

IDRs evolve at accelerated rate in the Mtb proteome

We next aimed to assess whether, as is the case for eukaryotes and some viruses [16, 48–53], IDRs in the Mtb proteome evolve faster than folded domains. To this aim, we exploited gene- and codon-level data from a previous analysis that used 10,209 Mtb genomes to estimate the non-synonymous/synonymous rate ratio (dN/dS) [54].

At the protein level, we found a significant, albeit weak, association between average dN/dS and IDR fraction (Kendall's rank correlation, tau = 0.11, p-value = 1.13×10^{-16}) (Additional file 1: Table S1). We thus moved to the codon level and we compared sites embedded in IDRs with those that are not in IDRs. A significantly higher dN/dS was observed for the former (Fig. 3A). Notably, this was true even when we considered proteins with IDRs only (Fig. 3A), indicating that this effect is not simply secondary to the functional characteristics of IDR-containing proteins (e.g., the fact that they interact with the host).

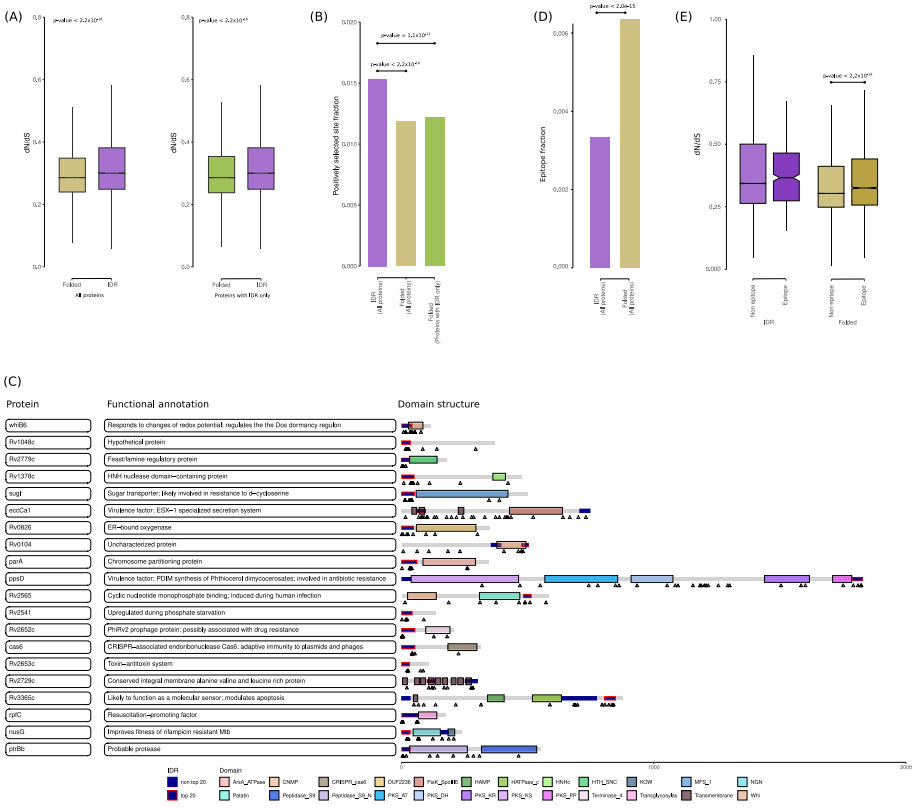

Fig. 3 IDR evolution and antigenicity. **A** Evolutionary rate in IDRs and folded domains was calculated using a site-wise measure of dN/dS. Higher dN/dS values are observed in IDRs compared with folded domains and this is the case for all proteins (left) and when only IDR-containing proteins (right) are analyzed. The p values derive from Brunner Munzel tests. **B** The fraction of positively selected sites (posterior probability of dN/dS ≥ 0.90) was calculated in IDRs and in folded domains of all proteins or IDR-containing proteins only. In both cases the fraction is significantly higher in IDRs (statistical analysis by Fisher exact tests). **C** Schematic representation of Mtb proteins carrying the 20 top IDRs in terms of positive selection signals (fraction of positively selected sites, represented as triangles). A functional annotation is also reported. Domain information was derived from the SMART database (https://smart.embl.de/). References are as follows: whiB6 [55], Rv2779c [56], sugI [11, 57, 58], Rv2565 [59], Rv2541 [60], Rv3365c [61], nusG [62]. **D** ATB and LTB epitopes for CD4 + T cells were mapped onto Mtb proteins. The fraction of epitope sites in IDRs is significantly lower than the one in folded domains (statistical analysis by Fisher exact test). **E** Comparison of dN/dS in epitopes and non-epitopes (statistical analysis by Brunner Munzel tests)

On one hand, high dN/dS values can result from either positive selection or from relaxation of functional constraints. On the other, when only a minority of sites are targeted, positive selection can occur even in genes showing, on average, low dN/dS [63]. We thus set out to quantify the amount of positive selection in IDR and non-IDR regions. To this aim, we defined positively selected sites as those having a posterior probability ≥ 0.90 of having dN/dS > 1, as suggested [54]. We found that IDRs have a significantly higher fraction of positively selected sites compared to non-IDR regions (Fisher's Exact test, OR = 1.30, $p < 2.2 \times 10^{-16}$) (Fig. 3B). Again, the same conclusion was confirmed when only IDR-containing proteins were analyzed (Fisher's Exact test, OR = 1.26, $p = 1.13 \times 10^{-13}$) (Fig. 3B). Overall, these analyses indicated that IDRs are fast evolving and represent preferential targets of positive selection in the Mtb proteome.

Inspection of the IDRs having the highest fraction (top 20) of positively selected sites (Additional file 4: Table S3) indicated that several of them are located within proteins that modulate interaction of the bacilli with their environment, including nutrient sensing, drug resistance, dormancy and resuscitation, as well as immunity to invading DNA

elements [11, 55–58, 60, 62, 64] (Fig. 3C). The enrichment of positively selected sites is clearly evident in these IDRs (Fig. 3C).

Mtb IDRs are poorly antigenic

Immune responses mediated by CD4 + T cells and CD8 + T cells are essential to control Mtb infection [4]. Previous analyses based on epitope prediction methods suggested that, due to biased amino acid composition, peptides presented by HLA class I and HLA class II molecules are less likely to derive from IDRs than from folded domains [65]. However, recent evidence indicated that IDR-containing human endogenous proteins are preferentially degraded by the proteasome [66], and non-canonical proteins, which account for about 25% of the HLA class I immunopeptidome source proteins, have high IDR fractions [67]. These contrasting observations raise the question of whether IDRs can be an active source of epitopes during Mtb infection. We thus set out to investigate the relationships between antigenicity and protein disorder, as well among these parameters and evolutionary rates. To this aim, we retrieved information on antigenic peptides deriving from two studies that used high-throughput assays to define the CD4 + T cell responses in LTB and ATB [68, 69]. A total of 692 epitopes were mapped onto 318 Mtb proteins (Additional file 5: Table S4). Antigenic proteins were found to have similar IDR fraction and similar dN/dS as the non-antigenic ones (Brunner Munzel test, both p values > 0.05).

We next moved at the codon level. We first asked whether antigenic peptides have the same probability to derive from IDRs or from folded domains. We found that epitopes are significantly and strongly depleted within IDRs either when all epitopes are analyzed together (Fig. 3D) or when ATB and LTB epitopes were separated (Fisher's Exact test, all p values $< 2.7 \times 10^{-13}$) (Additional file 2: Fig. S3). Similar results were obtained when only proteins with at least one mapped epitope were considered (Fisher's Exact test, p value $< 2.2 \times 10^{-16}$).

Previous studies showed that Mtb T-cell epitopes are hyperconserved (i.e., have lower dN/dS than non-epitopes) [70, 71]. Because we observed that IDRs evolve faster than folded domains and are depleted of epitopes, we reasoned that dN/dS comparisons should take structural information into account. Results indicated that epitopes in folded regions have, on average, higher dN/dS than non-epitopes (Fig. 3E, Brunner Munzel test, $p < 2.2 \times 10^{-16}$). In IDRs, epitopes were also found to evolve faster than non-epitopes, although the difference was not statistically significant (possibly also because of the small sample size) (Fig. 3E). These same patterns were observed for both ATB and LTB epitopes (Additional file 2: Fig. S3). As above, we thus wondered whether epitopes were more common targets of positive selection than non-epitopes. This was not the case, as no statistically significant enrichment of positively selected sites was detected in epitopes, either from IDRs or from folded domains (Fisher's Exact tests, p values > 0.05). Overall, these data indicated that IDRs in the Mtb proteome are a poor source of antigenic peptides. Contrary to previous findings [70, 71], we failed to detect hyperconservation of epitope sequences. Rather, we observed that, on average, epitopes evolve faster than non-epitopes, although this does not seem to be directly related to a stronger action of positive selection. Concerning the discrepancy with previous data, we should add that we used a dN/dS estimate deriving from more than 10,000 Mtb sequences, which increases the likelihood to include alternative amino acid alleles, most of which

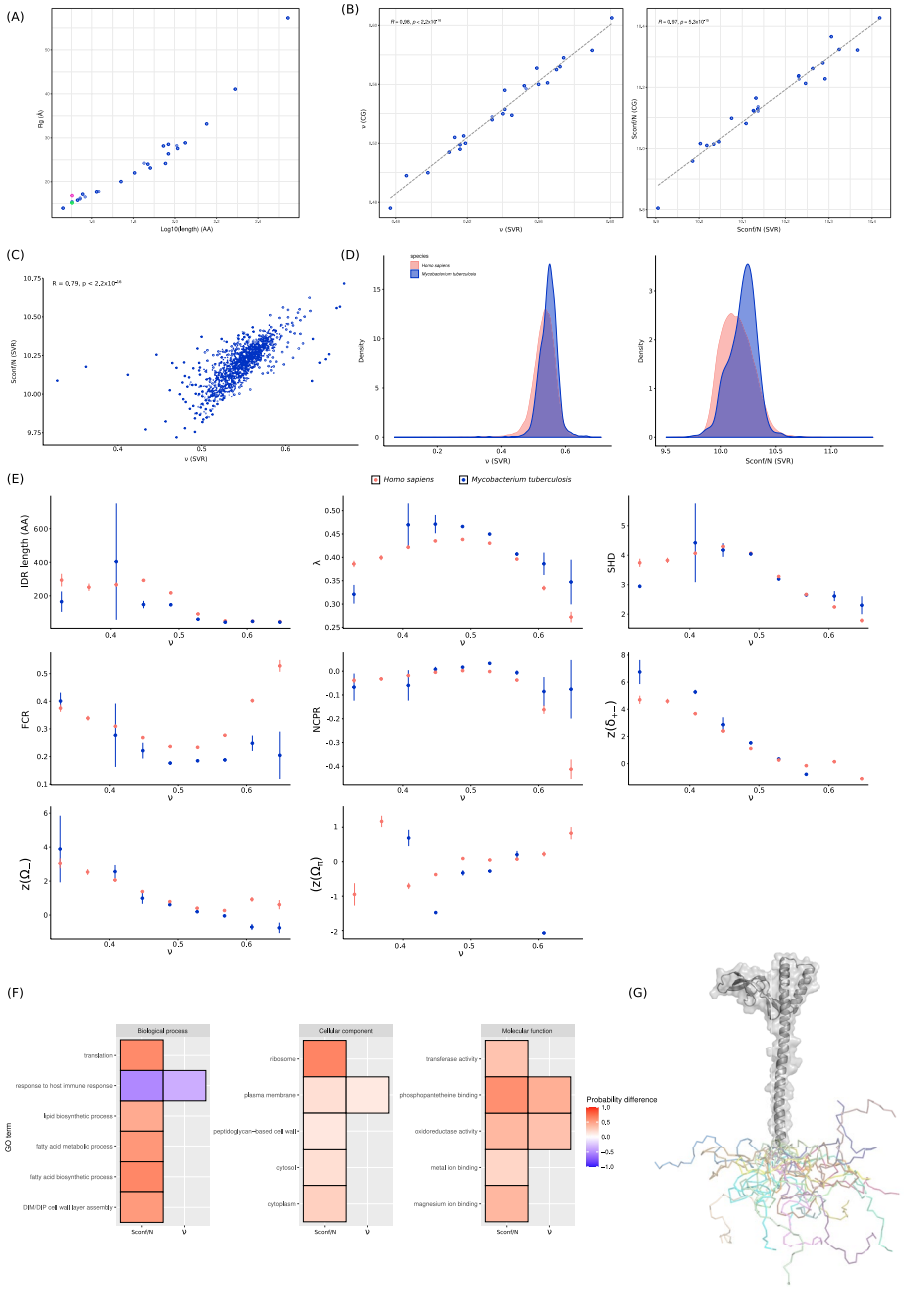

Fig. 4 (See legend on next page.)

are rare [54]. Earlier studies were instead based on estimates deriving from fewer than 25 Mtb sequences [70, 71].

Sequence-ensemble-function relationships of Mtb IDRs

We next aimed to investigate the biophysical properties of IDRs in the Mtb proteome. The difficulty in predicting structural features of IDRs has previously hampered efforts to understand their functional roles. However, some ensemble properties are quantifiable and can provide information on 3D features and IDR function. These include the conformational entropy per residue (*S*conf/*N*) and the Flory scaling exponent (*ν*), a measure of chain compactness. Recently, a predictor of Sconf/N and ν from sequence

(See figure on previous page.)

Fig. 4 Sequence-ensemble-function relationships of Mtb IDRs. **A** Relationship between radius of gyration (R_g) and length for 25 selected IDRs. The experimentally determined R_g for the FhaA IDR [74] is shown in magenta, the value derived from coarse-grained (CG) molecular dynamics simulations is in green. **B** Correlations between Sconf/N and ν values obtained with the SVR predictor and using CG simulations for 25 IDRs. Statistical significance was evaluated by Spearman's rank correlations. **C** Correlation between Sconf/N and ν values obtained with the SVR predictor for all IDRs in the Mtb proteome. **D** Comparison of conformational parameters between human and Mtb IDRs. Sconf/N and ν values for the human proteome were derived from a previous work [34]. **E** Sequence-ensemble relationships for Mtb and human IDRs. The following parameters are plotted as a function of ν: IDR length, average residue stickiness (λ), sequence hydropathy decoration (SHD), fraction of charged residues (FCR), net charge per residue (NCPR), NARDINI z-scores for the patterning of positively and negatively charged residues (δ + −), negatively charged residues (Ω −) and aromatic residues (Ω*π*). Positive z-scores indicate that the distribution of the residues/residue types is clustered along the IDR, whereas negative z-scores suggest an evenly spaced distribution. All values are plotted as mean and standard error calculated in bins of ν values. **F** GO analysis of IDR Sconf/N and ν. Only significant terms are shown (*p* value < 0.01 after FDR correction). The scale represents the probability difference from Brunner Munzel tests. **G** AlphaFold2 model of the GrpE protein (grey cartoon and surface) together with a qualitative representation of the conformational ensemble of its N-terminal IDR (colored sticks). IDR conformers have been extracted from a CG simulation, by selecting one frame every 50 (for a total of 20 conformers). They have been then structurally aligned with respect to their last three residues (A45, D46, A47) which, in turn, have been superimposed to the main chain of the corresponding residues of the GrpE AlphaFold2 model. In this latter, the N-terminal (residues 1–44) and C-terminal (192–235) IDRs have been omitted

information, based on a support vector regression (SVR) model, was developed [34]. The predictor was trained on simulations of all the IDRs in the human proteome performed using CALVADOS (Coarse-graining Approach to Liquid–liquid phase separation Via an Automated Data-driven Optimisation Scheme) [72]. To assess whether the SVR model generates reliable results in the case of bacterial proteins, as well, we first inspected the Protein Ensemble Database (PED, https://proteinensemble.org) for experimenta lly measured structural ensembles of Mtb proteins [73]. Only one entry was available (PED00509), corresponding to the N-terminal region of FhaA (Rv0020c). In a solution NMR analysis [74], the 32 N-terminal residues of FhaA were shown to be disordered, a finding that perfectly matches with our prediction of a 32 AA IDR at the N-terminus of the protein. The radius of gyration (R_g), a length-dependent measure of compaction, for this IDR was experimentally determined to be of 16.86 Å [74] (https://proteinensemble.org/entries/PED00509). We next used coarse-grained (CG) force field simulations to estimate the R_g of the FhaA IDR, plus 24 Mtb IDRs of variable length. For the FhaA IDR, the R_g we obtained resulted to be very similar to the experimental measure (Fig. 4A), indicating that the CG simulations can reliably capture ensemble properties. We thus used the simulations to calculate Sconf/N and ν [34, 72] and compare them with values obtained from the SVR model. The results were very highly correlated for both parameters (Spearman's rank correlation coefficients > 0.95, p values < 1×10^{-10}) (Fig. 4B), indicating that the SVR predictor is robust to sequence origin and can reliably estimate the conformational parameters for Mtb IDRs.

We thus used the predictor to calculate these parameters for all IDRs in the Mtb proteome ($n = 1351$) (Additional file 4: Table S3). Results indicated that the average values of ν and Sconf/N amount to 0.55 (SD: 0.027) and 10.20 (SD: 0.124) and that, as expected, IDRs with low Sconf/N tend to be more compact [34] (Fig. 4C). Comparison with the same parameters calculated for IDRs from the human proteome, indicated that the distribution of ν and Sconf/N for Mtb IDRs are narrower than those of human IDRs (Fig. 4D). Also, bacterial IDRs have significantly higher conformational entropy and adopt less compact conformations compared to the human ones (Brunner Munzel test, both p values < 2.2×10^{-16}) (Fig. 4D).

Previous analysis of human IDRs indicated that sequence features relate to conformational properties. For instance, compared to extended IDRs, the compact ones were found to be longer, and to display higher average residue stickiness [34]. We thus sought to investigate whether similar sequence-ensemble relationships characterize Mtb IDRs. To this aim, we calculated the following parameters: average residue stickiness (λ, a measurement of the strength of attractive intra-chain interactions) [72], sequence hydropathy decoration (SHD, a measure of the patterning of hydrophobic residues [75], fraction of charged residues (FCR), and net charge per residue (NCPR) (Additional file 4: Table S3). For the calculation of other sequence patterning parameters, we instead adopted the approach implemented in NARDINI (Non-random Arrangement of Residues in Disordered Regions Inferred using Numerical Intermixing) to account for sequence composition and length [76]. Specifically, NARDINI quantifies the extent of mixing or segregation of different pairs of amino acid types and computes z-scores that derive from the comparison of a given sequence with the distribution obtained from scrambled sequences of the same length and composition. Using this approach, for each IDR we calculated the patterning of positively and negatively charged residues (z(δ + −)), as well as the clustering of negatively charged (z(Ω −)) and aromatic (z(Ωπ)) residues (Additional file 4: Table S3).

In analogy to observations in the human proteome, we found that compact Mtb IDRs tend to be longer than those that adopt an extended conformation [34] (Fig. 4E). Comparison of sequence features with human IDRs indicated similar trends, whereby λ peaks at values of compaction between 0.4 and 0.5, and z(δ + −) and z(Ω −) decrease with increasing ν, suggesting an important role for electrostatic interactions in driving compaction (Fig. 4E). The clustering of hydrophobic residues also has a similar trend as in human IDRs, whereas z(Ωπ) is more difficult to evaluate, due to few available data points (Fig. 4E). The most notable difference between human and Mtb IDRs is their overall fraction of charged residues, which is lower in the bacterial sequences, especially in the extended ones (Fig. 4E). Mtb IDRs with high ν are also considerably less acidic than those adopting a similarly extended conformation in the human proteome (Fig. 4E).

Finally, we aimed to explore ensemble-function relationships. In eukaryotes, IDR compaction was shown to relate to specific protein functional classes or localization [34, 77–79]. For instance, compact IDRs were reported to be associated with chromatin binding and to be more likely to promote phase separation (PS) [79]. Although PS is still poorly explored in bacteria, an ATP-binding cassette (ABC) transporter of Mtb (encoded by *Rv1747)* undergoes phosphorylation-dependent PS and forms higher-order nanoclusters within the cell membrane [80]. The ABC transporter is the only Mtb protein known to undergo PS and our estimates of conformational parameters indicate that it contains a compact IDR with low conformational entropy (Sconf/N = 10.015, 8th percentile; ν = 0.531, 22th percentile). Thus, to investigate the association of IDR conformational properties with the biological function and localization of the corresponding proteins, we again resolved to GO analysis. We thus tested whether in proteins associated to a given GO term, the probability of finding compact IDRs or IDRs with low Sconf/N is greater or smaller than it is in proteins that are not associated to the same term. Analyses indicated a significant enrichment of compact IDRs with low entropy only in Mtb proteins that counteract the host immune response (Fig. 4F). The low-entropy fraction comprised several PE/PPE family members, including some PE-PGRS proteins

(Additional file 4: Table S3). Conversely, we found that expanded IDRs or IDRs with high conformational entropy are preferentially located in ribosomal proteins that participate in translation, in proteins that have a role in fatty acids metabolism, as well as in those that contribute to cell wall assembly (Fig. 4F). Interestingly, two of the IDRs having highest conformational entropy are located at the N- and C-terminus (85.1th and 99.9th percentiles) of the GrpE protein. As a cofactor of the DnaK chaperone, GrpE functions in the response to hyperosmotic and heat shock [81]. A recent cryo-EM study showed that the DnaK – GrpE complex is dynamic and undergoes a series of defined motions, and that the N-terminal IDR is essential for substrate release from DnaK. The authors proposed that the IDR drives the movements of DnaK to allow its function [81]. We thus used coarse grained simulations to represent the conformational ensemble of the GrpE N-terminal IDR. The simulations confirmed a wide variety of conformers, in line with high Sconf/N (Fig. 4G).

Human proteins that are engaged in PPIs with Mtb tend to be slow evolving

Host proteins that interact with microbial and viral components are often engaged in host–pathogen conflicts and tend to be fast evolving [63, 82, 83]. Also, one report showed that human-proteins that establish PPIs with viral proteins have a higher IDR fraction than those that do not [84]. We thus aimed to study the evolution and IDR composition of the 238 human proteins that we found to directly interact with Mtb proteins (Fig. 2C, Additional file 3: Table S2). Using the same AlphaFold2-based approach we applied to predict Mtb IDRs, we thus analyzed 19,852 human proteins and, in line with previous estimates, we found 31.6% of the proteome to be disordered [34] (Additional file 6: Table S5).

We first asked whether human proteins that interact with Mtb have a higher IDR fraction. This was not the case and, instead, the opposite situation was observed, although the difference did not reach the conventional cutoff for significance (Brunner Munzel test, $p = 0.088$) (Fig. 5A, Fig. 2C, Additional file 6: Table S5).

We next compared the relative evolutionary rates of Mtb-interacting proteins with those of proteins that do not interact with Mtb using two approaches. (1) We obtained

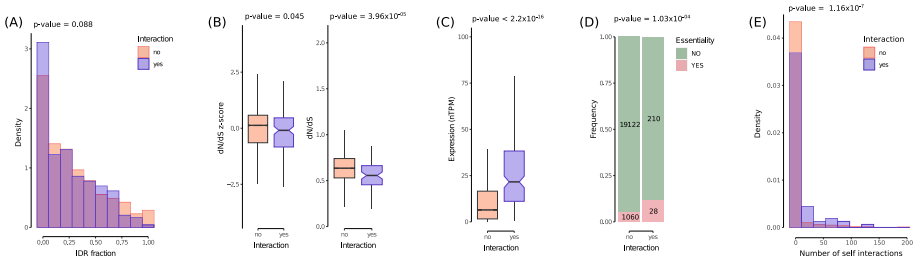

Fig. 5 Characterization of human proteins that interact with Mtb. **A** Comparison of the IDR fraction in proteins that do or do not interact with Mtb proteins (*p* value from Brunner Munzel test). **B** Evolutionary rates of proteins that interact with Mtb vs those that do not interact. Two measures of dN/dS are shown: on the left, *Z*-scores of dN/dS that quantifies the divergence of human genes relative to the ancestral primate genome; on the right, a measure of dN/dS obtained from of human variation data (missense and synonymous variant counts in ExAC). Both p values derive from Brunner Munzel tests. **C** Average expression level in 50 human tissues for human genes that participate or do not participate to PPIs with Mtb proteins. Levels are expressed as average consensus normalized expression (nTPM, normalized transcripts per million) (p value from Brunner Munzel test). **D** Comparison of the fraction of essential proteins among those that interact or do not interact with Mtb (p value from Fisher Exact Test). **E** Comparison of the number of human–human interactions for proteins that interact or do not interact with Mtb (p value from Brunner Munzel test)

adjusted dN/dS ratios from a previous work that analyzed 11,667 orthologs in primates [85]. In particular, the authors calculated a measure that represents the *Z*-score (corrected for GC content) that quantifies the divergence of human genes relative to the ancestral primate genome. (2) We retrieved a dN/dS measure of human variation from an analysis of 7506 genes [86] (Additional file 6: Table S5). Specifically, the estimate is based on missense and synonymous variant counts from the ExAc database. The two measures resulted to be correlated although not very strongly (Spearman's rank correlation, $R = 0.30$, $p = 2.2 \times 10^{-16}$). This is expected as they capture evolutionary rates over very different time periods and the estimate based on human variation data most likely incorporates several mildly deleterious mutations that are maintained in the population because selection had no time to eliminate them. Using both measures, we observed that Mtb interacting proteins are significantly more conserved than the ones not interacting with Mtb (Brunner Munzel test, $p = 0.045$ and 3.96×10^{-5}) (Fig. 5B).

Next, we studied another factor known to be associated with conservation, gene expression level [87] (Additional file 6: Table S5). By calculating the average transcript expression level across 50 human tissues using the Human Protein Atlas (HPA) consensus tissue data [88], we observed that genes encoding Mtb-interacting proteins have significantly higher expression than genes encoding proteins that are not reported to interact with Mtb (Brunner Munzel test, p-value $< 2.2 \times 10^{-16}$) (Fig. 5C).

We next focused on gene essentiality measures using a dataset of 1,103 human essential genes [89] which were identified using several in vitro and in vivo assays (Additional file 6: Table S5). We found that a significantly higher fraction of Mtb-binding proteins are encoded by essential genes (28 proteins, 11.7%) compared with genes that do not encode Mtb-interacting proteins (1060 proteins, 5.3%) (Fisher's exact test, odds ratio = 2.41, p value $= 1.03 \times 10^{-4}$) (Fig. 5D).

Finally, we analyzed human interactome data from the STRING database [90]. We restricted our analysis to high-confidence physical interactions ($n = 41{,}764$) and we observed that human proteins that interact with Mtb have significantly higher degree centrality (i.e., more interactions in the human interactome) than proteins that are not known to bind Mtb proteins (Brunner Munzel test, p-value $= 1.16 \times 10^{-7}$) (Fig. 5E, Additional file 6: Table S5).

Overall, these data indicate that human proteins engaged by Mtb in PPIs are evolutionary constrained, often essential and highly expressed, and interact with many other human proteins.

Comparison with other mycobacteria and human-infecting bacteria

The analyses above were conducted using the proteome of the reference Mtb strain, which has a long history of laboratory usage [91]. We thus decided to investigate whether IDR content is similar in clinical isolates. We selected 5 isolates from different continents, obtained their proteomes and used AlphaFold2 to obtain structural models of orthologous proteins that were found in all isolates ($n = 2235$) and in the H37Rv proteome. We then used the AlphaFold2-based method to identify IDRs. We found that all clinical isolates have a similar fraction of residues embedded in IDRs as the reference strain (Additional file 2: Fig. S4). At the individual protein level, very strong correlations between the IDR fraction in the H37Rv proteome and in either of the clinical isolate

proteomes were detected (Kendall's correlation coefficients higher than 0.88, all p values $< 2.2 \times 10^{-16}$) (Additional file 2: Fig. S4).

We next sought to compare the fraction of residues in IDRs among different mycobacteria. From the Mycobrowser portal (https://mycobrowser.epfl.ch/) we selected four mycobacteria for which reference proteomes are available [92]. One of them, *M. bovis*, is part of the *Mycobacterium tuberculosis* complex and causes TB in humans and other animals. Two other mycobacteria, *M. leprae* (the causative agent of leprosy) and *M. marinum*, are also pathogenic, whereas *M. smegmatis* is usually considered a non pathogenic bacterium [93–95]. Due to different evolutionary histories and ecological characteristics, these bacteria have proteomes of variable sizes, ranging from 1603 proteins in *M. leprae* to more than 6600 in *M. smegmatis*. Calculation of the IDR fraction in the proteomes of these mycobacteria indicated a similar IDR fraction as in Mtb (Fig. 6A, Additional file 7: Table S6).

To perform a more direct comparison, we identified one-to-one orthologs in the genomes of these mycobacteria. A total of 608 protein-coding genes present in all five mycobacteria were found (hereafter referred to as core genes) (Additional file 8: Table S7). In line with the results obtained above for essential/non essential genes, the IDR fraction was lower in the proteins encoded by these core genes compared to the full proteomes, and very similar among the five species (Fig. 6A). Consistently, in all species with the exclusion of *M. smegmatis*, non core proteins had significantly higher IDR fractions than the core ones (Fig. 6B).

Previous studies of the IDR fraction in prokaryotic proteins provided average estimates ranging from 4 to 10%, depending on the study, the methodology, and the analyzed species [96–98]. We thus aimed to put our results in a wider context by comparing the IDR fraction in mycobacteria proteomes with other bacteria that infect humans. We thus obtained a list of human-infecting bacterial pathogens [99] and we selected the ones for which a reference proteome was available. Of these we retained those that had at least 90% of proteins available in the AlphaFold Structure database. A total of 1,712,841 proteins from 540 proteomes were analyzed (Additional file 9: Table S8). We observed a wide variation in the average fraction of residues embedded in IDRs for these bacterial proteomes, ranging from 2.4% in *Campylobacter jejuni* (subsp. *Jejuni*) to 35.5% in *Mycobacteroides abscessus* (subsp. *abscessus*) (Fig. 6C). The IDR fraction of Mtb corresponds to the 71th percentile in this distribution. The overall IDR fraction was unrelated to proteome size (Spearman's rank correlation, $R = -0.03$, $p = 0.46$) and only weakly correlated with the genomic GC content (Spearman's rank correlation, $R = 0.17$, $p = 8.1 \times 10^{-5}$).

We next asked whether the high IDR fraction in some proteomes was due to more proteins with IDRs or to longer IDRs. Depending on the proteome, both cases were observed and the ones having a very high IDR fraction had both IDRs longer than the average and more IDR-containing proteins (Fig. 6C, Additional file 9: Table S8). As expected, IDR content was found to vary with taxonomic classification: clear differences were evident among phyla, although considerable intra-phylum variance was evident. Bacteria in the *Spirochetota*, *Mycoplasmatota*, *Chlamydiota*, *Actinomycetota,* and *Thermodesulfobacteriota* showed, on average, the highest IDR content (Fig. 6D, Additional file 9: Table S8). Moreover, several proteomes were outliers, including the above-mentioned *M. abscessus* and *Rothia mucilaginosa* (Fig. 6D).

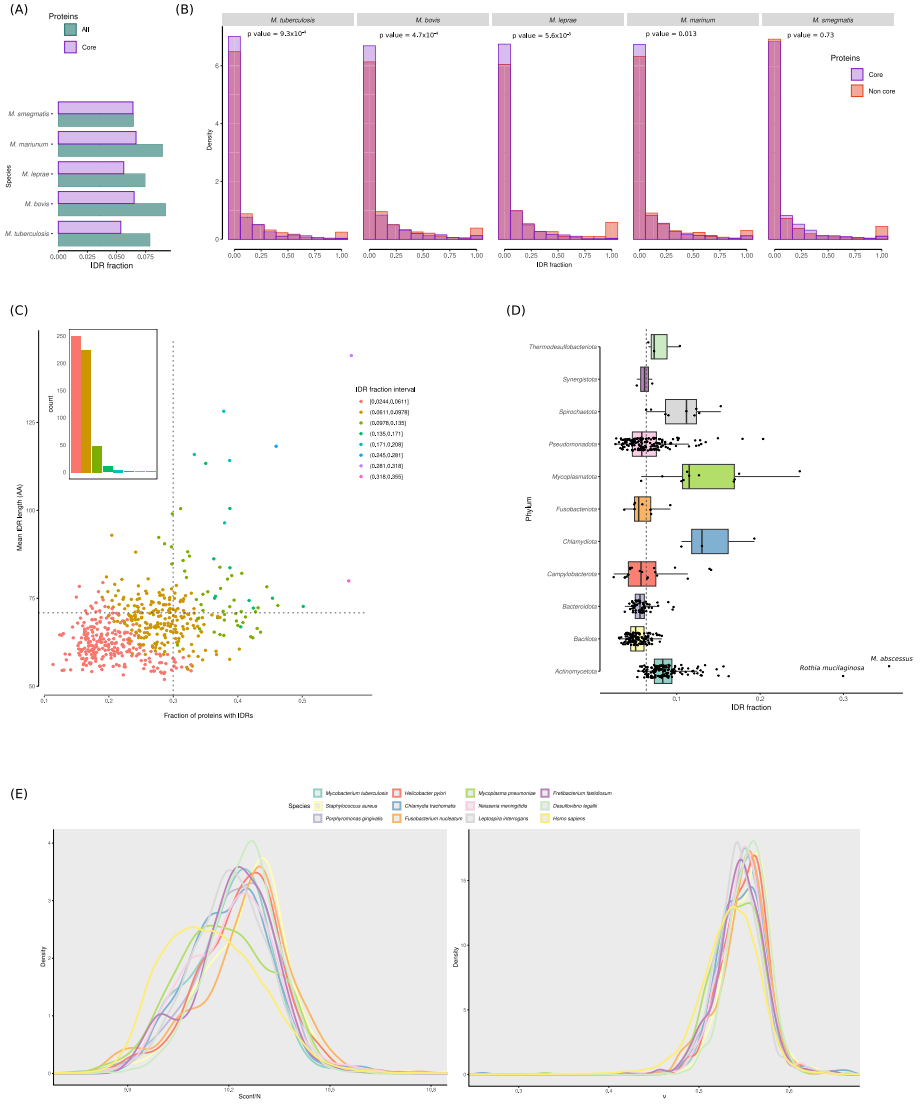

Fig. 6 Comparison with other mycobacteria and bacteria in different phyla. **A** The average IDR fraction in the proteomes of five mycobacteria is shown together with the average fraction calculated for 608 core genes. **B** Comparison of the IDR fraction between core and non core proteins in the five mycobacteria. All p values derive from Brunner Munzel tests. **C** The distribution of the average IDR fraction in the proteomes of 540 bacterial species that infect humans is shown in the inset. For the same proteomes, the fraction of proteins with at least one IDR is plotted against average IDR length (in amino acids). The hatched lines correspond to the 75.th percentiles of both distributions. **D** Comparison of IDR fraction among bacteria in different phyla. The hatched line corresponds to the median calculated for all proteomes. **E** Conformational parameters (Sconf/N on the left, ν on the right) calculated for all IDRs in the proteomes of 11 bacterial species selected to be representative of different phyla. The same distribution for IDRs in the human proteome is also reported [34]. To increase the resolution, extreme values (lower than the 0.001 or higher than the 0.999 percentiles in the distribution of all values) were removed. Curves are colored to correspond to the phylum each species belongs to (as per D panel)

Finally, we investigated how IDR ensemble properties vary within and among bacterial proteomes and in comparison to the human proteome. We thus selected 10 bacterial species to be representative of different phyla (*Actinomycetota* is represented by Mtb) and we used the SVR predictor [34] to calculate ν and Sconf/N (Additional file 10: Table S9). Results indicated that the distributions of Sconf/N are wider than those of ν and display more variability among proteomes (Fig. 6E). In general, bacterial IDRs from all species have higher conformational entropy and are less compact than human IDRs.

Overall, these results underscore a wide diversity in the representation of IDRs in bacterial proteomes and only partial association between IDR fraction and taxonomic classification. Bacterial IDRs also differ in terms of conformational features from human IDRs, specifically in terms of higher conformational entropy.

Discussion

Possibly because they are less abundant than in eukaryotes (and in some viruses), surprisingly little is known about the distribution, function and evolution of IDRs in prokaryotes, including human-infecting bacteria. Whereas a few studies have investigated the biochemical and functional features of specific IDRs [19–31, 81], systematic analyses are missing and a number of questions remain unanswered. After benchmarking different disorder prediction tools, we dissected the distribution of IDRs in the Mtb proteome. Our data clearly point to a role of these protein regions in contributing to bacterial virulence, interaction with the human host, and control of immune responses. Consistently, IDRs are particularly abundant in membrane and secretory proteins. Mtb infects macrophages and establishes a niche as an intracellular pathogen by thwarting immune responses. Within macrophages, Mtb persists and replicates in the phagosomes, which the bacilli permeabilize so that secreted bacterial proteins leak into the cytosol. Also, Mtb can arrest phagosome maturation and translocate into the cytosol, where its proteins are directly secreted [4]. Several secretory proteins have specific eukaryotic signatures or can interact with host proteins to modulate host immune functions [100–102]. Like in many other bacteria, protein secretion in Mtb relies on different secretion systems (e.g., ESX-1 to ESX-5) that transport proteins, including PE–PPE/PGRS family members, beyond the cytoplasmic membrane and/or beyond the cell wall [103]. Secretion systems include membrane proteins, several of which represent virulence factors [102, 103]. For instance, most of the core components of ESX/type VII secretion systems (EccB, EccD, and EccE) are membrane-associated virulence factors with variable IDR fraction [102–104].

Another class of proteins associated with a high fraction of IDRs is that involved in lipid transport and biosynthesis. This is particularly interesting in light of Mtb biology. In fact, the bacterium is characterized by a unique cell wall structure that includes an outer membrane composed of various lipopolysaccharides and very long-chain fatty acids (mycolic acids). This lipid-rich cell wall forms a low-permeability barrier that makes the bacterium naturally resistant to several antibiotics, but also dependent on specialized transport molecules that scavenge lipids and other nutrients from the host cell [105–108]. Among lipid transporters, members of the mce (mammalian cell entry) family have been implicated in Mtb virulence [105, 109] and we found that most mce proteins contain IDRs. Related to this, it is worth noting that IDRs within proteins with different functions were found to have distinct conformational properties. Specifically, those involved in lipid binding and biosynthesis tend to have IDRs with high conformational entropy. Most studies of interaction between lipids and IDRs have focused on membrane proteins [110–114]. Very recently, an analysis of prenylated IDRs in small GTPases showed that conformational plasticity is a key determinant of lipid sorting and that lipid interactions are conformation-dependent [115]. Although these results derive from a very different system, it will be extremely relevant to determine whether the ensemble properties of Mtb IDRs affect the extent and repertoire of lipid binding and if

this is related to changes in lipid availability or other environmental challenges. At the opposite side of the spectrum, we observed that compact IDRs with low conformational entropy are primarily found in proteins that counteract the host immune response. As mentioned above, chain compactness is a feature of proteins that promote PS [34, 79]. Little is known on PS in bacteria, however recent evidence indicated that the lactoferrin-binding protein B secreted by *Moraxella bovis* uses an IDR to promote the formation of biomolecular condensates that sequester antimicrobial peptides [30]. Thus, an intriguing possibility is that some Mtb proteins hijack host immune responses by forming molecular condensates, for instance to segregate effector molecules.

We were also highly interested in determining whether, as previously shown for eukaryotes and viruses [16, 48–53], IDRs evolve at a faster rate than folded domain. We found it to be the case for Mtb and our data show that this effect is not simply related to protein function. A common explanation for the rapid evolution of IDRs is that, because they experience limited structural constraints, they are more tolerant to change [16]. Whereas this may partially explain our results, it is worth noting that up to 25% of pathogenic missense mutation in humans are estimated to occur within IDRs [116] and amino acid replacements in IDRs were found to increase fitness and promote host adaptation in viruses [117, 118]. This clearly implies that amino acid changes within these regions can have important functional consequences and that IDRs do not simply evolve by relaxation of selective constraints. Indeed, we report that IDRs represent a major target of positive selection in Mtb. We hypothesize that, on one hand Mtb uses fast-evolving IDRs to interact with host molecules, but also that IDRs might serve the function of environmental sensors that promote bacterial adaptation and response to different stresses. This is in line with the idea that IDRs, due to their dynamic conformations and solvent accessibility, are perfectly suited to function in physicochemical sensing [119]. Conversely, positive selection in Mtb IDRs is unlikely to be mediated by immune selection. In fact, we show that these regions are a poor source of peptides for antigen presentation. Contrasting data are available in the literature concerning the ability of IDRs to generate peptides that can be loaded onto HLA class II molecules. On one hand, immunopeptidome data from endogenous proteins showed that disorder is associated with a smaller number of presented epitopes [120]. Likewise, epitope prediction analyses of both endogenous and pathogen-derived proteins indicated that IDRs contain relatively fewer HLA class II binding peptides than folded domains [65, 121]. On the other hand, immunopeptidome analyses of human-infecting viruses showed that IDRs are an active source of peptides [122, 123]. For instance, the nucleocapsid protein of SARS-CoV-2 has three IDRs [124] that, based on recent immunopeptidome data, generate epitopes with similar efficiency as folded domains [123]. An interesting possibility, which will require additional investigation, is that specific IDR properties modulate antigenicity. As an important corollary, our analysis of Mtb epitopes indicated that these regions are not hyperconserved. Whereas the possible reasons for discrepancy with previous data [70, 71] are reported above, this observation represents a shift in the paradigm whereby Mtb might benefit from recognition by human CD4 + T cells [71]. However, our failure to detect a significant enrichment of positive selection signals within epitopes suggests that the immuno-mediated selective pressure is very weak.

The observation that Mtb uses fast evolving IDRs to interact with its host is in principle consistent with a genetic conflict scenario. Typically, this results in cyclical adaptation

and counter-adaptation that drives the rapid evolution of interacting pathogen and host proteins [63]. Host–pathogen genetic conflicts were described in a number of instances in the context of bacterial, viral and parasitic infections [63]. However, our analysis of human proteins that interact with Mtb indicated just the opposite: these proteins are more conserved than the average in the human proteome, and this was observed using evolutionary rate estimates measured along different time frames. Indeed, we also found that Mtb-interacting proteins are often essential, are widely expressed, and engage in several PPIs with endogenous proteins. Their essentiality, possible pleiotropic function in the multiple tissues where they are expressed, and the necessity to maintain multiple interactions are the likely determinants of their constrained evolution. Whereas these data are not in line with the genetic conflict scenario, they resonate with recent findings in virus-host interactions, whereby rapidly evolving viral motifs, often embedded in IDRs, preferentially target conserved and highly expressed essential host proteins [125]. This confers a selective advantage to the virus, as the host is locked in its ability to evolve in response to deleterious interactions with viral proteins. Apparently, Mtb has adopted a similar strategy and mutations in human proteins that prevent interaction with Mtb are unlikely to reach high frequency as they may have a strong deleterious effect on host fitness. Considering that Mtb is a human-adapted pathogen with no other natural hosts, these observations raise the question of what is the selective pressure driving the evolution of IDRs in the bacterial proteome. Growing evidence indicates that the genetic diversity of Mtb, which is structured into ten human-adapted lineages (L1–L10), has phenotypic and clinical relevance [126, 127]. Indeed, differences among lineages, and even within lineages, were described in terms of virulence, transmissibility, intracellular survival, induction of host responses, disease presentation, drug resistance, and other traits [126, 127]. The ten lineages also show distinct geographic distributions, with some occurring globally and others being geographically restricted [126, 127]. It has been suggested that the epidemiological success of distinct lineages results from different factors, including human migrations, but also by the adoption of different evolutionary strategies [126]. For instance, compared to L2, L3, and L4, L1 is more often associated with extrapulmonary disease and it is more likely to cause asymptomatic TB [128–130]. L1 has a reduced transmission rate per unit of time, but it is associated with a longer duration of the infectious period [131]. While this probably explains why L1 was not outcompeted by the more virulent lineages, it suggests that lineage-specific interactions with the host occur. Amino acid changes in IDRs can alter conformational properties or create/disrupt embedded motifs, which can facilitate rewiring of interactions with host factors (e.g. by strengthening/weakening existing interactions, or by evolving a motif to form a new interaction) [132–135]. In addition to genetic factors, human hosts differ in a number of relevant parameters, including population density, nutritional health, climatic environment, and co-infection status. The latter is particularly relevant, as TB is common in people living with HIV and previous studies have indicated that specific lineages are preferentially associated with seropositivity [127]. It is thus possible that the altered immune responses of HIV-infected patients exerts a selective pressure on Mtb, as the bacilli are dependent on the host response to promote granuloma formation. Finally, evidence of Mtb genetic diversity at the intra-host level, which results from either bacterial microevolution or mixed infections, sets the basis for competition and, consequently, natural selection [126]. While these considerations may account for our observations,

they by no means imply that human genetic diversity has no effect on Mtb infection, as documented by several associations [126]. Overall, these observations open a new scenario in our understanding of bacterial interactions with their hosts and indicate that the classical arms race paradigm is not universal.

We should add that a previous analysis on a subset of the Mtb interacting proteins we studied herein indicated a higher frequency of positive selection in recent human evolutionary history compared with proteins that do not interact with the bacterium [136]. Being based on the integrated haplotype score (iHS), the approach used by Penn and coworkers is very different from the one we applied herein. Because iHS identifies recent selection signatures by searching for unusually long haplotypes, it equally detects selection driven by coding and noncoding variants [137]. Indeed, estimates in humans indicated that selective sweeps driven by regulatory noncoding polymorphisms were much more abundant than those resulting from nonsynonymous changes [138]. Conversely, we were interested in capturing the evolutionary rate within coding sequences, as this translates into amino acid differences that may affect human-Mtb PPIs.

Our study has limitations. Whereas we applied and combined different methods to identify IDRs in Mtb proteins, all of them rely on predictions. Thus, as was the case with Metapredict and SPOT-Disorder, we cannot exclude that the AlphaFold2-based approach we applied has biased our IDR estimates in some proteins/protein families. Likewise, conformational parameters were derived from an SVR predictor. Although we performed coarse-grained simulations to verify the performance of the predictor with bacterial proteins, some discrepancies might exist in the wider repertoire of analyzed proteins. The most severe limitation of our study is probably the availability of a small number of known human-Mtb PPIs. The majority of such PPIs derive from high-throughput methods that focused on secreted and membrane proteins [136, 139]. Others derived from databases or recent reviews of the literature that collected information from different sources [4, 100]. Thus, some of these PPIs might be biased by scientific interest or by technical simplicity of working with specific proteins. Another potential caveat is that information about LTB and ATB epitopes was derived from a relatively small number of infected subjects [68, 69]. This is a reason for concern because HLA class I and class II genes are highly polymorphic in human populations and the analyzed subjects are clearly not representative of the overall genetic diversity. Whereas this may bias the identification of recognized epitopes, we decided to limit our analysis to studies that used high-throughput peptide assays to avoid biases related to the choice of specific proteins or antigens.

Conclusions

Our data underscore wide variations in IDR representation and conformational properties among bacterial proteomes, possibly suggesting that structural disorder performs different functions in distinct bacteria, and that it has a diverse impact on proteome features. In Mtb, IDRs contribute to the interaction with the human host, they are fast evolving and poorly antigenic. Based on the distribution of structural disorder, we also suggest that Mtb also uses IDRs to sense and interact with its environment, a hypothesis that will need experimental validation. From a more methodological perspective, our data underscore the need for benchmarking disorder prediction tools, especially when proteomes that contain unique protein families or domains are analyzed. Conversely,

the estimation of IDR conformational properties using methods that were trained on eukaryotic proteomes is robust to use on bacterial proteins.

Methods

Proteomes and structure models

The reference proteome of Mtb strain ATCC 25618/H37Rv (UP000001584) was downloaded from UniProt. Of 3,995 proteins, 3,950 were available in the AlphaFold structure database and had the same length as in Uniprot. For the human proteome (UP000005640), we downloaded the list of reviewed (Swiss-Prot) canonical proteins (20,420). Of these, 19,852 were available in the AlphaFold structure database. The list of other bacterial proteomes, their protein numbers and matches in the AlphaFold structure database are available in Tables S6 and S8.

The structure of PE-PGRS16 (Uniprot: Q79FU3) was downloaded from the AlphaFold structure database together with pLDDT scores and visualized using Pymol (PyMOL(TM) Molecular Graphics System, Version 2.4.0, Schrödinger, LLC).

For clinical isolates, five were selected to be representative of different continents: INS_MDR (UP000022296, Peru), BTB07-246 (UP000025451, Sweden), A70376 (UP000043524, Uganda), TBMENG-03 (UP000251908, India), SBH70 (UP000438333, Malaysia).

Model predictions

When structural models were not available in the AlphaFold database (this was the case of the Mtb clinical isolate proteomes and of several proteins from *Dietzia cinnamea, Gordonia bronchialis, Corynebacterium diphtheriae, Nocardia abscessus or N. farcinica*, and *Prescottella equi*), we predicted them using AlphaFold2 through the LocalColabFold tool [140]. Each protein was modeled as a monomer, we estimated 5 models, ran 3 prediction recycles, and selected the one ranked best. All other parameters were set as default.

Disorder prediction

We used different tools to predict disordered residues and IDRs. The Metapredict tool [32, 33] defines IDRs by applying a deep-learning algorithm based on a consensus score calculated from eight different disorder predictors [32]. Metapredict V2 was run using default parameters and IDRs were defined as consecutive disordered stretches longer than 30 residues.

AlphaFold2-based predictions [35] were generated as recently suggested by a study that analyzed the human proteome [34]. Briefly, we used the protti R package [141] to derive pLDDT scores from the AlphaFold structure database (https://AlphaFold.ebi.ac.uk/) [35]. We next generated window-averaged pLDDT scores using a window size of 15 AA [142]. Residues with $\langle \text{pLDDT} \rangle > 0.8$, were considered as folded, and those with $\langle \text{pLDDT} \rangle < 0.7$ were labeled as disordered. Residues with $0.7 \leq \langle \text{pLDDT} \rangle \leq 0.8$ were initially defined as gap regions. Next, folded and disordered regions shorter than ten residues were reclassified as gaps. Gap regions were then reassigned to the disordered fraction if they were flanked by disordered regions on both edges or on one single edge when N- or C-terminal. All other gap regions were instead relabelled as folded. IDRs shorter than 30 residues were discarded.

IDRs in PE-PGRS proteins were also predicted using SPOT-DISORDER-Single [36], which takes as input a protein sequence and, using an ensemble of deep neural networks, estimates short and long disordered regions. The program was run using default parameters and we only retained IDRs longer than 30 amino acids.

The sequences of all identified IDRs in the proteomes of *Mycobacterium tuberculosis*, *Chlamydia trachomatis, Desulfovibrio legallii*, *Fretibacterium fastidiosum, Fusobacterium nucleatum*, *Helicobacter pylori, Leptospira interrogans, Mycoplasma pneumoniae, Neisseria meningitidis, Porphyromonas gingivalis,* and *Staphylococcus aureus* are available as Additional data (Additional file 11: Datasets S1 to S11).

Comparison with experimentally solved protein structures

To evaluate the performance of the three disorder predictors, we leveraged experimentally determined structures, where regions that fail to be resolved (missing residues) are likely to represent IDRs [43]. The structures of all available Mtb (strain H37Rv) structures were downloaded from the Protein Data Bank (PDB) [42]. A total of 1753 structures were available, corresponding to 594 unique proteins. For computational tractability, we limited analysis to X-ray diffraction experiments and we discarded heteromultimers. For each protein, we then retained the structure covering the full protein and with the best resolution ($n = 229$) (Additional file 1: Table S1). These PDB files were parsed to derive missing residues (gaps). Gaps longer than 29 residues were considered IDRs. We next compared the gap locations from the PDB files with the IDR predictions obtained from AlphaFold2, Metapredict, and SPOT-Disorder. Specifically, we calculated the Matthews Correlation Coefficient (MCC) and the F1 score. These metrics are appropriate for binary predictors and MCC is reliable even when the classes are imbalanced (as in this case where IDRs account for a much lower fraction of residues than folded regions) [44–47].

Functional characterization of Mtb proteins

Data on protein essentiality were retrieved from a previous screen based on transposon mutagenesis [143]. We considered in the essential category proteins encoded from genes classified as "required for optimal growth" and "slow growth when mutated". Non-essential genes were classified as in the original publication. Information on essentiality was available for 3156 proteins present in our dataset (Additional file 1: Table S1).

A list of 58 genes involved in antibiotic resistance was derived from the Comprehensive Antibiotic Resistance Database (CARD) database (https://card.mcmaster.ca/, version 3.2.9, 2024–02–13 release) [144], whereas 251 proteins involved in Mtb virulence were derived from the virulence factor database (VFDB, http://www.mgc.ac.cn/VFs/) [145] (Additional file 1: Table S1). GO terms were derived from QuickGO (https://www.ebi.ac.uk/QuickGO/) using the protti R package [141].

Protein–protein interaction analysis

We obtained the list of interactions between Mtb proteins and between human proteins from the STRING database [90] (version 12.0). Specifically, we included only physical interactions that are based on experiments. We only retained high-confidence PPIs (interaction score > 0.70). Using these criteria, we obtained 2835 Mtb PPIs and 41,764 human PPIs.

To identify Mtb proteins that physically interact with human proteins, we leveraged different sources. Thus, the majority of interactions were derived from two relatively-large scale analyses that used a high-confidence mass spectrometry approach [136] or a yeast-2-hybrid system (only validated interactions were included) [139]. We also included PPIs deriving from two recent reviews [4, 100], as well as from host–pathogen interaction databases: PHISTO (Pathogen-Host Interaction Search Tool, https://www.phisto.org/), and Host–Pathogen Interaction Database (HPIDB 3.0, https://hpidb.igbb.msstate.edu/) [146, 147] (Additional file 3: Table S2). The human-Mtb interaction network was visualized with Cytoscape v 3.9.1 (https://cytoscape.org/) [148].

Mtb protein evolutionary rates and positive selection analysis

Gene-wise and codon-wise estimates of dN/dS were obtained from a previous work that used the GenomegaMap method to analyze 3,979 genes from 10,209 Mtb genomes from the CRyPTIC Consortium [54, 149]. Of these genes, 3778 were present in our dataset, amounting to 1,267,180 codons. For both genes and codons, we used the merged estimates from the site and window models. Positively selected codons were defined as having a probability ≥ 0.90 of $dN/dS > 1$, as in the original publication [54].

Human protein evolutionary rate analysis

To study the relative evolutionary rate of protein coding sequences we used two separate datasets. 1) Normalized dN/dS for 11,667 1:1 orthologs in primates, as estimated by Dumas and coworkers [85]. The authors adjusted the dN/dS ratios for biases induced by variations of mutation rate with the GC content of codons and renormalized the values obtained for each taxon across the whole genome. The final values represent the *Z*-score corrected for GC content that quantifies the divergence of human genes relative to the ancestral primate genome [85]. 2) A dN/dS measure of human variation, which was obtained by correcting the observed unique missense and synonymous variant counts in a gene (as derived from ExAC) for the total possible missense and synonymous variants in that gene based on the codon table [86].

Human gene expression levels and gene essentiality analysis

Consensus transcript expression levels summarized per gene in 50 tissues based on transcriptomics data from the Human protein Atlas (HPA) and GTEx were downloaded from the HPA website (https://www.proteinatlas.org/about/download) [88] (Additional file 6: Table S5). The consensus normalized expression (nTPM, normalized transcripts per million) value is calculated as the maximum nTPM value for each gene in the two data sources. Values were averaged across tissues.

We used a dataset of 1,103 essential genes, assembled by Bartha et al. [89]. For 1088 of these genes, the corresponding protein was present in our database. We considered a gene to be essential if it was found to be essential in at least one of the three main screens used in this work.

Epitope mapping and analysis

Epitope sequences were derived from two studies that defined the T cell responses in LTB and ATB [68, 69]. Epitopes were mapped to Mtb proteins using protein BLAST. When a particular epitope had more than one equally likely hit (in the case of epitopes

deriving from paralogous proteins) it was assigned to all of them. Epitopes mapping to proteins for which IDR information was not available (i.e., not present in the AlphaFold structure database) were discarded. Because of extensive epitope overlapping (also between ATB and LTB), epitopes covered the following numbers of codons: all ($n = 7615$), ATB ($n = 2081$), LTB ($n = 6052$).

Analysis of IDR conformational properties and sequence patterns

The conformational entropy per residue ($Sconf/N$) and the Flory scaling exponent (ν) were calculated for all bacterial IDRs identified using the AlphaFold2-based approach described above. $Sconf/N$ is a measure of the landscape of different structures accessible to an IDR. ν derives from the scaling laws of polymers that describe how chain dimensions vary as a function of chain length [150]. Both parameters were estimated using a Colab notebook (https://colab.research.google.com/github/KULL-Centre/_2023_Tesei_IDRome/blob/main/IDR_SVR_predictor.ipynb), which uses a support vector regression model trained on simulations performed using the CALVADOS model [72, 151]. The same SVR predictor was used to derive other measures: λ (average residue stickiness) [72], SHD (sequence hydropathy decoration) [75], FCR (fraction of charged residues), and NCPR (net charge per residue).

Sequence patterns were calculated using NARDINI (Non-random Arrangement of Residues in Disordered Regions Inferred using Numerical Intermixing) [76]. This tool combines residues in eight different groups based on their properties: polar, hydrophobic, positive, negative, aromatic, alanine, proline, and glycine and then it assesses their distribution along the sequence with respect to each other. By calculating z-scores derived from 10^5 shuffled sequences of the same length, nonrandom segregation between two types of groups are evaluated: positive z-scores indicate that the distribution of the two groups is clustered along the IDR, whereas negative z-scores suggest an uniform distribution along the sequence or a non-random mixing between the two residue groups. We then retrieved z-scores for patterning of positively and negatively charged residues ($z(\delta+-)$), negatively charged residues ($z(\Omega-)$) and aromatic ($z(\Omega\pi)$) residues. Biophysical properties were analyzed as a function of ν values. In particular, we created groups of IDR sequences based on ν values, ranging from the minimum and maximum values calculated for all Mtb IDRS, with a $\Delta\nu = 0.04$. Within each bin, mean and standard error for all other parameters were calculated. As a comparison, the same procedure was applied for human data from a previous work [34].

Molecular simulations

The coarse-grained force field CALVADOS 2 was used to perform molecular simulations and to generate IDR conformational ensembles [34, 72]. Molecular dynamics simulations were performed on a set of 24 IDR of variable length going from 30 to 350 residues, in the NVT ensemble with the Langevin integrator and a time step of 10 fs. Each simulation was carried out at 310 K, setting an ionic strength of 0.150 M, modelling both N- and C-terminals in charged form and histidine residues in neutral form. The simulation time was automatically set on the basis of sequence length, as suggested [34]. More in detail, the number of saved frames was set to 7000 for sequences < 150 residues, while for longer sequences it was determined using the quadratic function $3e^{-4}$*(number of residues)^2*1000. The number of steps has then been set to 1010*(number of saved frames),

and converted into nanoseconds according to ns = (number of steps)*0.01/1000. For each sequence, four independent replicas were carried out in order to improve statistics so that the properties of interests (R_g, as well as ν and Sconf/N, calculated as described in [34] could be averaged over the four trajectories.

Identification of orthologs

Orthologous mycobacterial genes were identified through the OrthologueDB, a database of orthologs for bacteria and archaea [152]. We included only one-to-one orthologs with the following classifications: SSD (Supporting species divergence) or RBB (Reciprocal Best-BLAST) (Additional file 8: Table S7).

For clinical isolates, orthologs were identified using the easy-cluster workflow of the MMseqs2 program [153]. Only one-to-one orthologs were retained.

Statistical analysis

All statistical analyses were performed in the R v.4.0.5 environment. Kendall's and Spearman's rank correlations and Fisher's Exact tests were performed using the stats package. Brunner Munzel tests were performed using the brunnermunzel R package. This test was chosen because it is not influenced by tied values and it does not require the assumption of equal variances between groups [154]. Statistical details about each comparison can be found in the figure legends.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03854-6.

Additional file 1: Table S1. List of proteins in the Mtb proteome with annotations relative to IDRs, essentiality, host interaction, dN/dS, virulence, AMR, within proteome interactions, and PDB accessions

Additional file 2: Fig. S1. Schematic representation of Mtb proteins carrying IDRs and belonging to significant GO categories. Fig. S2. Schematic representation of Mtb proteins, as well as orthologs from other bacteria in the ***Mycobacteriales*** order, belonging to significant GO categories and associated with virulence. Fig. S3. Analysis of IDRantigenicity with data split on the basis epitope type. Fig. S4. IDR content of the proteomes of five Mtb clinical isolates.

Additional file 3: Table S2. List of human-Mtb PPIs

Additional file 4: Table S3. List of Mtb IDRs with information relative to conformational and sequence pattern parameters, number of positively selected sites

Additional file 5: Table S4. List of epitopes with information relative to sequence and type

Additional file 6: Table S5. List of human proteins with annotations relative to IDRs, essentiality, dN/dS, gene expression, and within proteome interactions

Additional file 7: Table S6. List of proteins from other mycobacteria with annotations relative to IDR fraction, size, and number

Additional file 8: Table S7. List of mycobacterial core proteins

Additional file 9: Table S8. List of bacterial proteomes with information relative to the number of proteins in AlphaFold Structure Database, IDR fraction, number of proteins with IDRs, average IDR length

Additional file 10: Table S9. List of IDRs from 10 selected bacterial proteomes, with information relative to conformational parameters

Additional file 11: Dataset S1. Sequences of IDRs from the Mtb proteome. Dataset S2. Sequences of IDRs from the ***Chlamydia trachomatis*** proteome. Dataset S3. Sequences of IDRs from the ***Desulfovibrio legallii*** proteome. Dataset S4. Sequences of IDRs from the ***Fretibacterium fastidiosum*** proteome. Dataset S5. Sequences of IDRs from the ***Fusobacterium nucleatum*** proteome. Dataset S6. Sequences of IDRs from the ***Helicobacter pylori*** proteome. Dataset S7. Sequences of IDRs from the ***Leptospira interrogans*** proteome. Dataset S8. Sequences of IDRs from the ***Mycoplasma pneumoniae*** proteome. Dataset S9. Sequences of IDRs from the ***Neisseria meningitidis*** proteome. Dataset S10. Sequences of IDRs from the ***Porphyromonas gingivalis*** proteome. Dataset S11. Sequences of IDRs from the ***Staphylococcus aureus*** proteome.

Peer review information
Leonard Foster and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
MS and UP conceived the study and designed the experiment; RC, DF, UP, and FA performed the experiments and generated the data, with assistance from MS and LDG; RC, DF, UP, and FA conducted the data analysis with support from MS and LDG; MS and UP drafted the manuscript with input from all authors; MS, DF and UP organized the data and finalized the manuscript.

Funding
The work was supported by the Italian Ministry of Health—"Ricerca Corrente" program (to RC) and by the European Union—NextGenerationEU through the Italian Ministry of University and Research under PNRR—M4C2-I1.3 Project PE_00000019 "HEAL ITALIA" (to LDG).

Data availability
All the source data and result files in this paper are provided in the Additional Data Sets, and these data can be accessed and obtained from Zenodo (https://doi.org/10.5281/zenodo.17395676) [155]. Proteomes were downloaded from the Uniprot database (https://www.uniprot.org/) [156], whereas protein models were obtained through the AlphaFold Protein Structure database (https://alphafold.ebi.ac.uk/) [35]. Experimentally solved Mtb protein structures were downloaded from the Protein Data Bank [42]. Experimentally determined conformational parameters were derived from the Protein Ensemble Database (PED, https://proteinensemble.org) [73]. Data on antibiotic resistance and virulence were derived from the Comprehensive Antibiotic Resistance Database (CARD) database (https://card.mcmaster.ca/) [144] and the Virulence factor database (VFDB, http://www.mgc.ac.cn/VFs/) [145], respectively. Mtb and human gene essentiality were obtained from previous works [89, 143]. The same applied to dN/dS data for human and Mtb proteins [54, 85, 86, 149]. Host-pathogen PPIs were obtained from previous analyses [4, 100, 136, 139] and by interrogation of the Pathogen-Host Interaction Search Tool (PHISTO) (https://www.phisto.org/) [147] and Host-Pathogen Interaction Database (HPIDB) (https://hpidb.igbb.msstate.edu) [146], whereas intraspecies PPIs were retrieved from the STRING database (https://string-db.org) [90]. ATB and LTB epitopes were obtained from previous analyses [68, 69]. Gene ontology information was obtained from QuickGO (https://www.ebi.ac.uk/QuickGO/) [157]. Orthologous mycobacterial genes were identified through the OrthologueDB (https://ortholugedb.ca/) [152] and from the Mycobrowser portal (https://mycobrowser.epfl.ch/) [92]. Expression data of human genes were derived from the Human protein atlas (https://www.proteinatlas.org/) [158].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 9 January 2025 / Accepted: 30 October 2025
Published online: 24 November 2025

References

1. Philips JA, Ernst JD. Tuberculosis pathogenesis and immunity. Annu Rev Pathol Mech Dis. 2012;7:353–84. https://doi.org/10.1146/annurev-pathol-011811-132458.
2. Pai M, Behr MA, Dowdy D, Dheda K, Divangahi M, Boehme CC, et al. Tuberculosis. Nat Rev Dis Primers. 2016;2:16076. https://doi.org/10.1038/nrdp.2016.76.
3. Roy CJ, Milton DK. Airborne transmission of communicable infection — the elusive pathway. N Engl J Med. 2004;350:1710–2. https://doi.org/10.1056/NEJMp048051.
4. Chandra P, Grigsby SJ, Philips JA. Immune evasion and provocation by *Mycobacterium tuberculosis*. Nat Rev Microbiol. 2022;20:750–66. https://doi.org/10.1038/s41579-022-00763-4.
5. Colangeli R, Gupta A, Vinhas SA, Chippada Venkata UD, Kim S, Grady C, et al. *Mycobacterium tuberculosis* progresses through two phases of latent infection in humans. Nat Commun. 2020;11:4870. https://doi.org/10.1038/s41467-020-18699-9.
6. Blaser MJ, Kirschner D. The equilibria that allow bacterial persistence in human hosts. Nature. 2007;449:843–9. https://doi.org/10.1038/nature06198.
7. Rodrigo T, Caylà JA, García de Olalla P, Galdós-Tangüis H, Jansà JM, Miranda P, et al. Characteristics of tuberculosis patients who generate secondary cases. Int J Tuberc Lung Dis. 1997;1:352–7.
8. Brites D, Gagneux S. Co-evolution of *Mycobacterium tuberculosis* and *Homo sapiens*. Immunol Rev. 2015;264:6–24. https://doi.org/10.1111/imr.12264.
9. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet. 2013;45:1176–82. https://doi.org/10.1038/ng.2744.
10. Kohli S, Singh Y, Sharma K, Mittal A, Ehtesham NZ, Hasnain SE. Comparative genomic and proteomic analyses of PE/PPE multigene family of *Mycobacterium tuberculosis* H37Rv and H37Ra reveal novel and interesting differences with implications in virulence. Nucleic Acids Res. 2012;40:7113–22. https://doi.org/10.1093/nar/gks465.

11. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998;393:537–44. https://doi.org/10.1038/31159.
12. Brennan MJ, Delogu G, Chen Y, Bardarov S, Kriakov J, Alavi M, et al. Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. Infect Immun. 2001;69:7326–33. https://doi.org/10.1128/IAI.69.12.7326-7333.2001.
13. Veyrier FJ, Dufort A, Behr MA. The rise and fall of the *Mycobacterium tuberculosis* genome. Trends Microbiol. 2011;19:156–61. https://doi.org/10.1016/j.tim.2010.12.008.
14. Ahmad J, Khubaib M, Sheikh JA, Pancsa R, Kumar S, Srinivasan A, et al. Disorder-to-order transition in PE–PPE proteins of *Mycobacterium tuberculosis* augments the pro-pathogen immune response. FEBS Open Bio. 2020;10:70–85. https://doi.org/10.1002/2211-5463.12749.
15. Boradia V, Frando A, Grundner C, Waldor MK, editor. The *Mycobacterium tuberculosis* PE15/PPE20 complex transports calcium across the outer membrane. PLoS Biol. 2022;20:e3001906. https://doi.org/10.1371/journal.pbio.3001906.
16. Holehouse AS, Kragelund BB. The molecular basis for cellular function of intrinsically disordered protein regions. Nat Rev Mol Cell Biol. 2024;25:187–211. https://doi.org/10.1038/s41580-023-00673-0.
17. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015;16:18–29. https://doi.org/10.1038/nrm3920.
18. Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014;114:6589–631. https://doi.org/10.1021/cr400525m.
19. Ladouceur A-M, Parmar BS, Biedzinski S, Wall J, Tope SG, Cohn D, et al. Clusters of bacterial RNA polymerase are biomolecular condensates that assemble through liquid–liquid phase separation. Proc Natl Acad Sci U S A. 2020;117:18540–9. https://doi.org/10.1073/pnas.2005019117.
20. Holmes JA, Follett SE, Wang H, Meadows CP, Varga K, Bowman GR. *Caulobacter* PopZ forms an intrinsically disordered hub in organizing bacterial cell poles. Proc Natl Acad Sci USA. 2016;113:12490–5. https://doi.org/10.1073/pnas.1602380113.
21. Lasker K, Von Diezmann L, Zhou X, Ahrens DG, Mann TH, Moerner WE, et al. Selective sequestration of signalling proteins in a membraneless organelle reinforces the spatial regulation of asymmetry in *Caulobacter crescentus*. Nat Microbiol. 2020;5:418–29. https://doi.org/10.1038/s41564-019-0647-7.
22. Harami GM, Kovács ZJ, Pancsa R, Pálinkás J, Baráth V, Tárnok K, et al. Phase separation by ssDNA binding protein controlled via protein–protein and protein–DNA interactions. Proc Natl Acad Sci USA. 2020;117:26206–17. https://doi.org/10.1073/pnas.2000761117.
23. Hausmann S, Geiser J, Allen GE, Geslain SAM, Valentini M. Intrinsically disordered regions regulate RhlE RNA helicase functions in bacteria. Nucleic Acids Res. 2024;52:7809–24. https://doi.org/10.1093/nar/gkae511.
24. Shinn MK, Cohan MC, Bullock JL, Ruff KM, Levin PA, Pappu RV. Connecting sequence features within the disordered C-terminal linker of *Bacillus subtilis* FtsZ to functions and bacterial cell division. Proc Natl Acad Sci U S A. 2022;119:e2211178119. https://doi.org/10.1073/pnas.2211178119.
25. Oguri T, Kwon Y, Woo JKK, Prehna G, Lee H, Ning M, et al. A family of small intrinsically disordered proteins involved in flagellum-dependent motility in Salmonella enterica. O'Toole G, editor. J Bacteriol. 2019;201. https://doi.org/10.1128/JB.00415-18.
26. Brunet YR, Habib C, Brogan AP, Artzi L, Rudner DZ. Intrinsically disordered protein regions are required for cell wall homeostasis in *Bacillus subtilis*. Genes Dev. 2022;genesdev;gad.349895.122v1. https://doi.org/10.1101/gad.349895.122.
27. Mets T, Kurata T, Ernits K, Johansson MJO, Craig SZ, Evora GM, et al. Mechanism of phage sensing and restriction by toxin-antitoxin-chaperone systems. Cell Host Microbe. 2024;32:1059-1073.e8. https://doi.org/10.1016/j.chom.2024.05.003.
28. Hadži S, Živič Z, Kovačič M, Zavrtanik U, Haesaerts S, Charlier D, et al. Fuzzy recognition by the prokaryotic transcription factor HigA2 from *Vibrio cholerae*. Nat Commun. 2024;15:3105. https://doi.org/10.1038/s41467-024-47296-3.
29. Busby JN, Trevelyan S, Pegg CL, Kerr ED, Schulz BL, Chassagnon I, et al. The ABC toxin complex from *Yersinia entomophaga* can package three different cytotoxic components expressed from distinct genetic loci in an unfolded state: the structures of both shell and cargo. IUCrJ. 2024;11:299–308. https://doi.org/10.1107/S2052252524001969.
30. Ostan NKH, Cole GB, Wang FZ, Reichheld SE, Moore G, Pan C, et al. A secreted bacterial protein protects bacteria from cationic antimicrobial peptides by entrapment in phase-separated droplets. Hummer G, editor. PNAS Nexus. 2024;3:pgae139. https://doi.org/10.1093/pnasnexus/pgae139.
31. Byeon C, Hansen KH, Jeffrey J, Saricayir H, Andreasen M, Akbey Ü. Intrinsically disordered *Pseudomonas* chaperone FapA slows down the fibrillation of major biofilm-forming functional amyloid FapC. FEBS J. 2024;291(9):1925–43. https://doi.org/10.1111/febs.17084.
32. Emenecker RJ, Griffith D, Holehouse AS. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. Biophys J. 2021;120:4312–9. https://doi.org/10.1016/j.bpj.2021.08.039.
33. Emenecker RJ, Griffith D, Holehouse AS. Metapredict V2: An update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. 2022. https://doi.org/10.1101/2022.06.06.494887.
34. Tesei G, Trolle AI, Jonsson N, Betz J, Knudsen FE, Pesce F, et al. Conformational ensembles of the human intrinsically disordered proteome. Nature. 2024;626:897–904. https://doi.org/10.1038/s41586-023-07004-5.
35. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.
36. Hanson J, Paliwal K, Zhou Y. Accurate single-sequence prediction of protein intrinsic disorder by an ensemble of deep recurrent and convolutional architectures. J Chem Inf Model. 2018;58:2369–76. https://doi.org/10.1021/acs.jcim.8b00636.
37. Shi Z, Chen K, Liu Z, Ng A, Bracken WC, Kallenbach NR. Polyproline II propensities from GGXGG peptides reveal an anticorrelation with β-sheet scales. Proc Natl Acad Sci USA. 2005;102:17964–8. https://doi.org/10.1073/pnas.0507124102.
38. De Maio F, Berisio R, Manganelli R, Delogu G. PE_PGRS proteins of *Mycobacterium tuberculosis*: a specialized molecular task force at the forefront of host–pathogen interaction. Virulence. 2020;11:898–915. https://doi.org/10.1080/21505594.2020.1785815.
39. Warkentin E, Weidenweber S, Schühle K, Demmer U, Heider J, Ermler U. A rare polyglycine type II-like helix motif in naturally occurring proteins. Proteins. 2017;85:2017–23. https://doi.org/10.1002/prot.25355.
40. Pentelute BL, Gates ZP, Tereshko V, Dashnau JL, Vanderkooi JM, Kossiakoff AA, et al. X-ray structure of snow flea antifreeze protein determined by racemic crystallization of synthetic protein enantiomers. J Am Chem Soc. 2008;130:9695–701. https://doi.org/10.1021/ja8013538.

41. Dunne M, Denyes JM, Arndt H, Loessner MJ, Leiman PG, Klumpp J. *Salmonella* phage S16 tail fiber adhesin features a rare polyglycine rich domain for host recognition. Structure. 2018;26:1573-1582.e4. https://doi.org/10.1016/j.str.2018.07.017.
42. Berman H, Henrick K, Nakamura H. Announcing the worldwide protein data bank. Nat Struct Mol Biol. 2003;10:980–980. https://doi.org/10.1038/nsb1203-980.
43. Maiti S, Singh A, Maji T, Saibo NV, De S. Experimental methods to study the structure and dynamics of intrinsically disordered regions in proteins. Curr Res Struct Biol. 2024;7:100138. https://doi.org/10.1016/j.crstbi.2024.100138.
44. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 1975;405:442–51. https://doi.org/10.1016/0005-2795(75)90109-9.
45. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-Score and ROC: a family of discriminant measures for performance evaluation. In: Sattar A, Kang B, editors. AI 2006: Advances in Artificial Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 1015–21. https://doi.org/10.1007/11941439_114.
46. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21:6. https://doi.org/10.1186/s12864-019-6413-7.
47. Cao C, Chicco D, Hoffman MM. The MCC-F1 curve: a performance evaluation technique for binary classification. arXiv; 2020. https://doi.org/10.48550/ARXIV.2006.11278.
48. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. Evolution and disorder. Curr Opin Struct Biol. 2011;21:441–6. https://doi.org/10.1016/j.sbi.2011.02.005.
49. Afanasyeva A, Bockwoldt M, Cooney CR, Heiland I, Gossmann TI. Human long intrinsically disordered protein regions are frequent targets of positive selection. Genome Res. 2018;28:975–82. https://doi.org/10.1101/gr.232645.117.
50. Zarin T, Strome B, Nguyen Ba AN, Alberti S, Forman-Kay JD, Moses AM. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. Elife. 2019;8:e46883. https://doi.org/10.7554/eLife.46883.
51. Mozzi A, Forni D, Cagliani R, Clerici M, Pozzoli U, Sironi M. Intrinsically disordered regions are abundant in simplexvirus proteomes and display signatures of positive selection. Virus Evol. 2020;6:veaa028. https://doi.org/10.1093/ve/veaa028.
52. Molteni C, Forni D, Cagliani R, Bravo IG, Sironi M. Evolution and diversity of nucleotide and dinucleotide composition in poxviruses. J Gen Virol. 2023;104. https://doi.org/10.1099/jgv.0.001897.
53. Cagliani R, Forni D, Mozzi A, Fuchs R, Tussia-Cohen D, Arrigoni F, et al. Evolution of virus-like features and intrinsically disordered regions in retrotransposon-derived mammalian genes. Mol Biol Evol. 2024;41:msae154. https://doi.org/10.1093/molbev/msae154.
54. Wilson DJ, The CRyPTIC Consortium, Crook DW, Peto TEA, Walker AS, Hoosdally SJ, et al. GenomegaMap: Within-Species Genome-Wide dN/dS Estimation from over 10,000 Genomes. Rosenberg M, editor. Mol Biol Evol. 2020;37:2450–60. https://doi.org/10.1093/molbev/msaa069.
55. Chen Z, Hu Y, Cumming BM, Lu P, Feng L, Deng J, et al. Mycobacterial WhiB6 differentially regulates ESX-1 and the Dos regulon to modulate granuloma formation and virulence in zebrafish. Cell Rep. 2016;16:2512–24. https://doi.org/10.1016/j.celrep.2016.07.080.
56. Betts JC, Lukey PT, Robb LC, McAdam RA, Duncan K. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. Mol Microbiol. 2002;43:717–31. https://doi.org/10.1046/j.1365-2958.2002.02779.x.
57. Anthony RM, Molemans M, Akkerman O, Sturkenboom MGG, Mulder A, De Zwaan R, et al. The appearance of *sugI* mixed loci in three individuals during treatment for MDR-TB, supports the involvement of *sugI* in *Mycobacterium tuberculosis* d-cycloserine resistance *in vivo*. 2023. https://doi.org/10.1101/2023.05.30.542839.
58. Chen J, Zhang S, Cui P, Shi W, Zhang W, Zhang Y. Identification of novel mutations associated with cycloserine resistance in *Mycobacterium tuberculosis*. J Antimicrob Chemother. 2017;72:3272–6. https://doi.org/10.1093/jac/dkx316.
59. Kumar M, Khan FG, Sharma S, Kumar R, Faujdar J, Sharma R, et al. Identification of *Mycobacterium tuberculosis* genes preferentially expressed during human infection. Microb Pathog. 2011;50:31–8. https://doi.org/10.1016/j.micpath.2010.10.003.
60. Rifat D, Bishai WR, Karakousis PC. Phosphate depletion: a novel trigger for *Mycobacterium tuberculosis* persistence. J Infect Dis. 2009;200:1126–35. https://doi.org/10.1086/605700.
61. Danelishvili L, Everman JL, McNamara MJ, Bermudez LE. Inhibition of the Plasma-Membrane-Associated Serine Protease Cathepsin G by Mycobacterium tuberculosis Rv3364c Suppresses Caspase-1 and Pyroptosis in Macrophages. Front Microbio. 2012;2. https://doi.org/10.3389/fmicb.2011.00281 .
62. Eckartt KA, Delbeau M, Munsamy-Govender V, DeJesus MA, Azadian ZA, Reddy AK, et al. Compensatory evolution in NusG improves fitness of drug-resistant *M. tuberculosis*. Nature. 2024;628:186–94. https://doi.org/10.1038/s41586-024-07206-5.
63. Sironi M, Cagliani R, Forni D, Clerici M. Evolutionary insights into host-pathogen interactions from mammalian sequence data. Nat Rev Genet. 2015;16:224–36. https://doi.org/10.1038/nrg3905.
64. Bhalla N, Nanda RK. Pangenome-wide association study reveals the selective absence of CRISPR genes (Rv2816c-19c) in drug-resistant *Mycobacterium tuberculosis*. Microbiol Spectr. 2024;e00527-24. https://doi.org/10.1128/spectrum.00527-24.
65. Mitić NS, Pavlović MD, Jandrlić DR. Epitope distribution in ordered and disordered protein regions — part A. T-cell epitope frequency, affinity and hydropathy. J Immunol Methods. 2014;406:83–103. https://doi.org/10.1016/j.jim.2014.02.012.
66. Myers N, Olender T, Savidor A, Levin Y, Reuven N, Shaul Y. The disordered landscape of the 20S proteasome substrates reveals tight association with phase separated granules. Proteomics. 2018;18:1800076. https://doi.org/10.1002/pmic.201800076.
67. Ruiz Cuevas MV, Hardy M-P, Hollý J, Bonneil É, Durette C, Courcelles M, et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep. 2021;34:108815. https://doi.org/10.1016/j.celrep.2021.108815.
68. Lindestam Arlehamn CS, Gerasimova A, Mele F, Henderson R, Swann J, Greenbaum JA, et al. Memory T Cells in Latent Mycobacterium tuberculosis Infection Are Directed against Three Antigenic Islands and Largely Contained in a CXCR3+CCR6+ Th1 Subset. Salgame P, editor. PLoS Pathog. 2013;9:e1003130. https://doi.org/10.1371/journal.ppat.1003130.
69. Panda S, Morgan J, Cheng C, Saito M, Gilman RH, Ciobanu N, et al. Identification of differentially recognized T cell epitopes in the spectrum of tuberculosis infection. Nat Commun. 2024;15:765. https://doi.org/10.1038/s41467-024-45058-9.
70. Arlehamn CSL, Paul S, Mele F, Huang C, Greenbaum JA, Vita R, et al. Immunological consequences of intragenus conservation of *Mycobacterium tuberculosis* T-cell epitopes. Proc Natl Acad Sci U S A. 2015;112:E147-55. https://doi.org/10.1073/pnas.1416537112.

71. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat Genet. 2010;42:498–503. https://doi.org/10.1038/ng.590.
72. Tesei G, Lindorff-Larsen K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. Open Res Eur. 2023;2:94. https://doi.org/10.12688/openreseurope.14967.2.
73. Ghafouri H, Lazar T, Del Conte A, Tenorio Ku LG, PED Consortium, Aspromonte MC, et al. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. Nucleic Acids Research. 2024;52:D536–44. https://doi.org/10.1093/nar/gkad947.
74. Roumestand C, Leiba J, Galophe N, Margeat E, Padilla A, Bessin Y, et al. Structural insight into the *Mycobacterium tuberculosis* Rv0020c protein and its interaction with the PknB kinase. Structure. 2011;19:1525–34. https://doi.org/10.1016/j.str.2011.07.011.
75. Zheng W, Dignon G, Brown M, Kim YC, Mittal J. Hydropathy patterning complements charge patterning to describe conformational preferences of disordered proteins. J Phys Chem Lett. 2020;11:3408–15. https://doi.org/10.1021/acs.jpclett.0c00288.
76. Cohan MC, Shinn MK, Lalmansingh JM, Pappu RV. Uncovering non-random binary patterns within sequences of intrinsically disordered proteins. J Mol Biol. 2022;434:167373. https://doi.org/10.1016/j.jmb.2021.167373.
77. González-Foutel NS, Glavina J, Borcherds WM, Safranchik M, Barrera-Vilarmau S, Sagar A, et al. Conformational buffering underlies functional selection in intrinsically disordered protein regions. Nat Struct Mol Biol. 2022;29:781–90. https://doi.org/10.1038/s41594-022-00811-w.
78. Martin EW, Holehouse AS, Peran I, Farag M, Incicco JJ, Bremer A, et al. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science. 2020;367:694–9. https://doi.org/10.1126/science.aaw8653.
79. Ibrahim AY, Khaodeuanepheng NP, Amarasekara DL, Correia JJ, Lewis KA, Fitzkee NC, et al. Intrinsically disordered regions that drive phase separation form a robustly distinct protein class. J Biol Chem. 2023;299:102801. https://doi.org/10.1016/j.jbc.2022.102801.
80. Heinkel F, Abraham L, Ko M, Chao J, Bach H, Hui LT, et al. Phase separation and clustering of an ABC transporter in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A. 2019;116:16326–31. https://doi.org/10.1073/pnas.1820683116.
81. Xiao X, Fay A, Molina PS, Kovach A, Glickman MS, Li H. Structure of the M. tuberculosis DnaK–GrpE complex reveals how key DnaK roles are controlled. Nat Commun. 2024;15:660. https://doi.org/10.1038/s41467-024-44933-9.
82. Meyerson NR, Sawyer SL. Two-stepping through time: mammals and viruses. Trends Microbiol. 2011;19:286–94. https://doi.org/10.1016/j.tim.2011.03.006.
83. Tenthorey JL, Emerman M, Malik HS. Evolutionary landscapes of host-virus arms races. Annu Rev Immunol. 2022;40:271–94. https://doi.org/10.1146/annurev-immunol-072621-084422.
84. Lou DI, Kim ET, Meyerson NR, Pancholi NJ, Mohni KN, Enard D, et al. An intrinsically disordered region of the DNA repair protein Nbs1 is a species-specific barrier to Herpes simplex virus 1 in primates. Cell Host Microbe. 2016;20:178–88. https://doi.org/10.1016/j.chom.2016.07.003.
85. Dumas G, Malesys S, Bourgeron T. Systematic detection of brain protein-coding genes under positive selection during primate evolution and their roles in cognition. Genome Res. 2021;31:484–96. https://doi.org/10.1101/gr.262113.120.
86. van der Lee R, Wiel L, van Dam TJP, Huynen MA. Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. Nucleic Acids Res. 2017;45:10634–48. https://doi.org/10.1093/nar/gkx704.
87. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA. 2005;102:14338–43. https://doi.org/10.1073/pnas.0504070102.
88. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347:1260419. https://doi.org/10.1126/science.1260419.
89. Bartha I, Di Iulio J, Venter JC, Telenti A. Human gene essentiality. Nat Rev Genet. 2018;19:51–62. https://doi.org/10.1038/nrg.2017.75.
90. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49:D605–12. https://doi.org/10.1093/nar/gkaa1074.
91. Chitale P, Lemenze AD, Fogarty EC, Shah A, Grady C, Odom-Mabey AR, et al. A comprehensive update to the *Mycobacterium tuberculosis* H37Rv reference genome. Nat Commun. 2022;13:7068. https://doi.org/10.1038/s41467-022-34853-x.
92. Kapopoulou A, Lew JM, Cole ST. The mycobrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis. 2011;91:8–13. https://doi.org/10.1016/j.tube.2010.09.006.
93. Reyrat J-M, Kahn D.*Mycobacterium smegmatis*: an absurd model for tuberculosis? Trends Microbiol. 2001;9:472–3. https://doi.org/10.1016/S0966-842X(01)02168-0.
94. Aubry A, Mougari F, Reibel F, Cambau E.*Mycobacterium marinum*. Microbiol Spectr. 2017. https://doi.org/10.1128/microbiolspec.TNMI7-0038-2016.
95. Taye H, Alemu K, Mihret A, Wood JLN, Shkedy Z, Berg S, et al. Global prevalence of *Mycobacterium bovis* infections among human tuberculosis cases: systematic review and meta-analysis. Zoonoses Public Health. 2021;68:704–18. https://doi.org/10.1111/zph.12868.
96. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol. 2004;337:635–45. https://doi.org/10.1016/j.jmb.2004.02.002.
97. Uversky VN. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. Front Phys. 2019;7:10. https://doi.org/10.3389/fphy.2019.00010.
98. Basile W, Salvatore M, Bassot C, Elofsson A. Why do eukaryotic proteins contain more intrinsically disordered regions? PLoS Comput Biol. 2019;15:e1007186. https://doi.org/10.1371/journal.pcbi.1007186.
99. Bartlett A, Padfield D, Lear L, Bendall R, Vos M. A comprehensive list of bacterial pathogens infecting humans. Microbiology. 2022;168. https://doi.org/10.1099/mic.0.001269.
100. Pal R, Bisht MK, Mukhopadhyay S. Secretory proteins of *Mycobacterium tuberculosis* and their roles in modulation of host immune responses: focus on therapeutic targets. FEBS J. 2022;289:4146–71. https://doi.org/10.1111/febs.16369.
101. Prisic S, Husson RN. *Mycobacterium tuberculosis* serine/threonine protein kinases. Microbiol Spectr. 2014. https://doi.org/10.1128/microbiolspec.MGM2-0006-2013.
102. Gröschel MI, Sayes F, Simeone R, Majlessi L, Brosch R. ESX secretion systems: mycobacterial evolution to counter host immunity. Nat Rev Microbiol. 2016;14:677–91. https://doi.org/10.1038/nrmicro.2016.131.

103. Bitter W, Houben ENG, Bottai D, Brodin P, Brown EJ, Cox JS, et al. Systematic genetic nomenclature for type VII secretion systems. PLoS Pathog. 2009;5:e1000507. https://doi.org/10.1371/journal.ppat.1000507.
104. Poweleit N, Czudnochowski N, Nakagawa R, Trinidad DD, Murphy KC, Sassetti CM, et al. The structure of the endogenous ESX-3 secretion system. Elife. 2019;8:e52983. https://doi.org/10.7554/eLife.52983.
105. Pandey AK, Sassetti CM. Mycobacterial persistence requires the utilization of host cholesterol. Proc Natl Acad Sci USA. 2008;105:4376–80. https://doi.org/10.1073/pnas.0711159105.
106. Lee W, VanderVen BC, Fahey RJ, Russell DG. Intracellular *Mycobacterium tuberculosis* exploits host-derived fatty acids to limit metabolic stress. J Biol Chem. 2013;288:6788–800. https://doi.org/10.1074/jbc.M112.445056.
107. Rempel S, Gati C, Nijland M, Thangaratnarajah C, Karyolaimos A, De Gier JW, et al. A mycobacterial ABC transporter mediates the uptake of hydrophilic compounds. Nature. 2020;580:409–12. https://doi.org/10.1038/s41586-020-2072-8.
108. Arnold FM, Weber MS, Gonda I, Gallenito MJ, Adenau S, Egloff P, et al. The ABC exporter IrtAB imports and reduces mycobacterial siderophores. Nature. 2020;580:413–7. https://doi.org/10.1038/s41586-020-2136-9.
109. Gioffré A, Infante E, Aguilar D, Santangelo MDLP, Klepp L, Amadio A, et al. Mutation in mce operons attenuates *Mycobacterium tuberculosis* virulence. Microbes Infect. 2005;7:325–34. https://doi.org/10.1016/j.micinf.2004.11.007.
110. Cornish J, Chamberlain SG, Owen D, Mott HR. Intrinsically disordered proteins and membranes: a marriage of convenience for cell signalling? Biochem Soc Trans. 2020;48:2669–89. https://doi.org/10.1042/BST20200467.
111. Sigrist SJ, Haucke V. Orchestrating vesicular and nonvesicular membrane dynamics by intrinsically disordered proteins. EMBO Rep. 2023;24:e57758. https://doi.org/10.15252/embr.202357758.
112. Zeno WF, Baul U, Snead WT, DeGroot ACM, Wang L, Lafer EM, et al. Synergy between intrinsically disordered domains and structured proteins amplifies membrane curvature sensing. Nat Commun. 2018;9:4152. https://doi.org/10.1038/s41467-018-06532-3.
113. Hicks A, Escobar CA, Cross TA, Zhou H-X. Fuzzy association of an intrinsically disordered protein with acidic membranes. JACS Au. 2021;1:66–78. https://doi.org/10.1021/jacsau.0c00039.
114. Yuan F, Lee CT, Sangani A, Houser JR, Wang L, Lafer EM, et al. The ins and outs of membrane bending by intrinsically disordered proteins. Sci Adv. 2023;9:eadg3485. https://doi.org/10.1126/sciadv.adg3485.
115. Araya MK, Gorfe AA. Conformational ensemble-dependent lipid recognition and segregation by prenylated intrinsically disordered regions in small GTPases. Commun Biol. 2023;6:1111. https://doi.org/10.1038/s42003-023-05487-6.
116. Tsang B, Pritišanac I, Scherer SW, Moses AM, Forman-Kay JD. Phase separation as a missing mechanism for interpretation of disease mutations. Cell. 2020;183:1742–56. https://doi.org/10.1016/j.cell.2020.11.050.
117. Dolan PT, Taguwa S, Rangel MA, Acevedo A, Hagai T, Andino R, et al. Principles of dengue virus evolvability derived from genotype-fitness maps in human and mosquito cells. Elife. 2021;10:e61921. https://doi.org/10.7554/eLife.61921.
118. Charon J, Barra A, Walter J, Millot P, Hebrard E, Moury B, et al. First experimental assessment of protein intrinsic disorder involvement in an RNA virus natural adaptive process. Mol Biol Evol. 2018;35:38–49. https://doi.org/10.1093/molbev/msx249.
119. Moses D, Ginell GM, Holehouse AS, Sukenik S. Intrinsically disordered regions are poised to act as sensors of cellular chemistry. Trends Biochem Sci. 2023;48:1019–34. https://doi.org/10.1016/j.tibs.2023.08.001.
120. Stražar M, Park J, Abelin JG, Taylor HB, Pedersen TK, Plichta DR, et al. Hla-II immunopeptidome profiling and deep learning reveal features of antigenicity to inform antigen discovery. Immunity. 2023;56:1681-1698.e13. https://doi.org/10.1016/j.immuni.2023.05.009.
121. Guy AJ, Irani V, MacRaild CA, Anders RF, Norton RS, Beeson JG, et al. Insights into the immunological properties of intrinsically disordered malaria proteins using proteome scale predictions. PLoS One. 2015;10:e0141729. https://doi.org/10.1371/journal.pone.0141729.
122. Brito-Sierra CA, Lannan MB, Siegel RW, Malherbe LP. The HLA class-II immunopeptidomes of AAV capsids proteins. Front Immunol. 2022;13:1067399. https://doi.org/10.3389/fimmu.2022.1067399.
123. Weingarten-Gabbay S, Chen D-Y, Sarkizova S, Taylor HB, Gentili M, Hernandez GM, et al. The HLA-II immunopeptidome of SARS-CoV-2. Cell Rep. 2024;43:113596. https://doi.org/10.1016/j.celrep.2023.113596.
124. Cubuk J, Alston JJ, Incicco JJ, Singh S, Stuchell-Brereton MD, Ward MD, et al. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. Nat Commun. 2021;12:1936. https://doi.org/10.1038/s41467-021-21953-3.
125. Shuler G, Hagai T. Rapidly evolving viral motifs mostly target biophysically constrained binding pockets of host proteins. Cell Rep. 2022;40:111212. https://doi.org/10.1016/j.celrep.2022.111212.
126. Goig GA, Windels EM, Loiseau C, Stritt C, Biru L, Borrell S, et al. Ecology, global diversity and evolutionary mechanisms in the *Mycobacterium tuberculosis* complex. Nat Rev Microbiol. 2025. https://doi.org/10.1038/s41579-025-01159-w.
127. Galagan JE. Genomic insights into tuberculosis. Nat Rev Genet. 2014;15:307–20. https://doi.org/10.1038/nrg3664.
128. Du DH, Geskus RB, Zhao Y, Codecasa LR, Cirillo DM, Van Crevel R, et al. The effect of *M. tuberculosis* lineage on clinical phenotype. PLOS Glob Public Health. 2023;3:e0001788. https://doi.org/10.1371/journal.pgph.0001788.
129. Freschi L, Vargas R, Husain A, Kamal SMM, Skrahina A, Tahseen S, et al. Population structure, biogeography and transmissibility of *Mycobacterium tuberculosis*. Nat Commun. 2021;12:6099. https://doi.org/10.1038/s41467-021-26248-1.
130. Gröschel MI, Pérez-Llanos FJ, Diel R, Vargas R, Escuyer V, Musser K, et al. Differential rates of *Mycobacterium tuberculosis* transmission associate with host–pathogen sympatry. Nat Microbiol. 2024;9:2113–27. https://doi.org/10.1038/s41564-024-01758-y.
131. Zwyer M, Rutaihwa LK, Windels E, Hella J, Menardo F, Sasamalo M, et al. Back-to-Africa introductions of *Mycobacterium tuberculosis* as the main cause of tuberculosis in Dar es Salaam, Tanzania. PLoS Pathog. 2023;19:e1010893. https://doi.org/10.1371/journal.ppat.1010893.
132. Glavina J, Rodríguez De La Vega RC, Risso VA, Leonetti CO, Chemes LB, Sánchez IE. Host diversification is concurrent with linear motif evolution in a Mastadenovirus hub protein. J Mol Biol. 2022;434:167563. https://doi.org/10.1016/j.jmb.2022.167563.
133. Chemes LB, Glavina J, Faivovich J, De Prat-Gay G, Sánchez IE. Evolution of linear motifs within the Papillomavirus E7 oncoprotein. J Mol Biol. 2012;422:336–46. https://doi.org/10.1016/j.jmb.2012.05.036.
134. Gitlin L, Hagai T, LaBarbera A, Solovey M, Andino R. Rapid evolution of virus sequences in intrinsically disordered protein regions. PLoS Pathog. 2014;10:e1004529. https://doi.org/10.1371/journal.ppat.1004529.

135. Ortiz JF, MacDonald ML, Masterson P, Uversky VN, Siltberg-Liberles J. Rapid evolutionary dynamics of structural disorder as a potential driving force for biological divergence in flaviviruses. Genome Biol Evol. 2013;5:504–13. https://doi.org/10.1093/gbe/evt026.
136. Penn BH, Netter Z, Johnson JR, Von Dollen J, Jang GM, Johnson T, et al. An Mtb-human protein-protein interaction map identifies a switch between host antiviral and antibacterial responses. Mol Cell. 2018;71:637-648.e5. https://doi.org/10.1016/j.molcel.2018.07.010.
137. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. Hurst L, editor. PLoS Biol. 2006;4:e72. https://doi.org/10.1371/journal.pbio.0040072.
138. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, et al. Identifying recent adaptations in large-scale genomic data. Cell. 2013;152:703–13. https://doi.org/10.1016/j.cell.2013.01.035.
139. Yang F, Lei Y, Zhou M, Yao Q, Han Y, Wu X, et al. Development and application of a recombination-based library versus library high- throughput yeast two-hybrid (RLL-Y2H) screening system. Nucleic Acids Res. 2018;46:e17–e17. https://doi.org/10.1093/nar/gkx1173.
140. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. Colabfold: making protein folding accessible to all. Nat Methods. 2022;19:679–82. https://doi.org/10.1038/s41592-022-01488-1.
141. Quast J-P, Schuster D, Picotti P. protti: an R package for comprehensive data analysis of peptide- and protein-centric bottom-up proteomics data. Arighi C, editor. Bioinformatics Advances. 2022;2:vbab041. https://doi.org/10.1093/bioadv/vbab041.
142. Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, et al. A structural biology community assessment of AlphaFold2 applications. Nat Struct Mol Biol. 2022;29:1056–67. https://doi.org/10.1038/s41594-022-00849-w.
143. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol. 2003;48:77–84. https://doi.org/10.1046/j.1365-2958.2003.03425.x.
144. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, Wlodarski MA, et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res. 2023;51:D690–9. https://doi.org/10.1093/nar/gkac920.
145. Liu B, Zheng D, Zhou S, Chen L, Yang J. VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res. 2022;50:D912–7. https://doi.org/10.1093/nar/gkab1107.
146. Kumar R, Nanduri B. HPIDB - a unified resource for host-pathogen interactions. BMC Bioinform. 2010;11:S16. https://doi.org/10.1186/1471-2105-11-S6-S16.
147. Durmuş Tekir S, Çakır T, Ardıç E, Sayılırbaş AS, Konuk G, Konuk M, et al. PHISTO: pathogen–host interaction search tool. Bioinformatics. 2013;29:1357–8. https://doi.org/10.1093/bioinformatics/btt137.
148. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
149. The CRyPTIC Consortium and the 100,000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N Engl J Med. 2018;379:1403–15. https://doi.org/10.1056/NEJMoa1800474.
150. Flory PaulJ, Volkenstein M. Statistical mechanics of chain molecules. Biopolymers. 1969;8:699–700. https://doi.org/10.1002/bip.1969.360080514.
151. Tesei G, Schulze TK, Crehuet R, Lindorff-Larsen K. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. Proc Natl Acad Sci U S A. 2021;118:e2111696118. https://doi.org/10.1073/pnas.2111696118.
152. Whiteside MD, Winsor GL, Laird MR, Brinkman FSL. OrthologeDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis. Nucleic Acids Res. 2013;41:D366–76. https://doi.org/10.1093/nar/gks1241.
153. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–8. https://doi.org/10.1038/nbt.3988.
154. Neubert K, Brunner E. A studentized permutation test for the non-parametric Behrens-Fisher problem. Comput Stat Data Anal. 2007;51:5192–204. https://doi.org/10.1016/j.csda.2006.05.024.
155. Pozzoli U, Forni D, Arrigoni F, Cagliani R, De Gioia L, Sironi M. Mycobacterium tuberculosis uses intrinsically disordered, fast evolving proteins to interact with conserved host factors. Datasets. Zenodo; 2025. https://doi.org/10.5281/ZENODO.17395676.
156. Consortium TU. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D158–69. https://doi.org/10.1093/nar/gkw1099.
157. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for gene ontology searching. Bioinformatics. 2009;25:3045–6. https://doi.org/10.1093/bioinformatics/btp536.
158. Sjöstedt E, Zhong W, Fagerberg L, Karlsson M, Mitsios N, Adori C, et al. An atlas of the protein-coding genes in the human, pig, and mouse brain. Science. 2020;367:eaay5947. https://doi.org/10.1126/science.aay5947.

Publisher's Note