Single nucleotide variation catalogue from clinical isolates mapped on tertiary and quaternary structures of ESX-1 related proteins reveals critical regions as putative Mtb therapeutic targets

Oren Tzfadia[1], Axel Siroy[2], Alexandra Vujkovic[1], Abril Gijsbers[3], Jihad Snobre[1], Roger Vargas[4,6], Wim Mulders[1], Conor J. Meehan[1, 5], Maha Farhat[6], Peter J. Peters[7], Bouke C. de Jong*[1], Raimond B.G. Ravelli*[7]*

* Corresponding authors. Emails:

Affiliations

1. Mycobacteriology Unit, Institute of Tropical Medicine, Antwerpen, Belgium
2. Structure and Function of Bacterial Nanomachines, UMR 5234, Univ. Bordeaux, CNRS, Institut Européen de Chimie et Biologie, France
3. Departamento de Química de Biomacromoléculas, Instituto de Química, Universidad Nacional Autónoma de México, Mexico City, Mexico
4. Roger, Moderna
5. Nottingham Trent University, UK
6. Department of Biomedical Informatics, Harvard Medical School, Boston, USA
7. Division of Nanoscopy, Maastricht Multimodal Imaging Institute (M4i), Maastricht University, The Netherlands

Abstract

Proteins encoded by the ESX-1 genes of interests are essential for full virulence in all *Mycobacterium tuberculosis* complex (MTBc) lineages, the pathogens with the highest mortality worldwide. Identifying critical regions in these ESX-1 related proteins could provide preventive or therapeutic targets for MTB infection, the game changer needed for tuberculosis control. We analysed a compendium of whole genome sequences of clinical MTB isolates from all lineages from >32,000 patients and identified single nucleotide variations (SNV). When mutations corresponding to all nonsynonymous SNPs were mapped on the surface of known and AlphaFold-predicted ternary protein structures, fully conserved regions emerged. Some could be assigned to known quaternary structures, whereas others could be predicted to be involved in yet-to-be-discovered interactions. Some mutants had clonally expanded (found in >1% of the isolates): these were mostly located at the surface of globular domains, remote from known intra- and inter-molecular protein–protein interactions. Fully conserved intrinsically disordered regions (IDRs) of proteins were found, suggesting that these are crucial for the pathogenicity of the MTBc. Altogether, our findings provide an evolutionary structural perspective on MTB virulence and highlight fully conserved regions of proteins as attractive vaccine antigens and drug targets. Extending this approach to other pathogens can provide a novel critical resource for the development of innovative tools for pathogen control.

Introduction

Mycobacterium tuberculosis causes tuberculosis (TB) in humans and other mammals. This remarkably monomorphic pathogen shares 99.9% nucleotide similarity and identical 16S rRNA in its larger *Mycobacterium tuberculosis* complex (MTBc), unlike diversity seen in other bacteria (Böddinghaus et al. 1990; Sreevatsan et al. 1997; Achtman and Wagner 2008; Wiens et al. 2018). In the past decades, extensive research has been done to clarify the precise virulence mechanisms of the MTBc. ESAT-6 (6-kDa early secretory antigenic target, also known as EsxA) was identified as the main virulence-determining secreted protein (Andersen et al. 1995; Brodin et al. 2004). The avirulence of the century-old vaccine strain *M. bovis* Calmette-Guérin (BCG) is explained by deletion of the chromosomal region (RD1) containing *esxA* (Mahairas et al. 1996). Deletion of RD1 from the MTBc caused decreased virulence similar to that of BCG *in vitro* (Lewis et al. 2003). *M. microtii* spontaneously lost a similar RD1 region. It was shown that ESAT-6 interacts with CFP-10 and that this secreted heterodimer is critical for MTBc virulence through its cytolytic activity (Wel et al. 2007; Xu et al. 2007; Houben et al. 2012; Tiwari et al. 2019).

RD1 contains genes of the ESAT-6 secretion system 1, ESX-1 (Mahairas et al. 1996; Tekaia et al. 1999). ESX-1 is a member of the type VII secretion systems (T7SS), and is essential for full virulence in all MTBc lineages (L1–L8) as well as in the closely related pathogenic *M. marinum*, *M. kansasi* (Jagielski et al. 2020), and *M.* leprae (PMID: 11597336). Some genes necessary for ESX-1 transport and its regulation have been found outside the RD1 locus. The ESX-1 genes of interest (GOI) include multiple genes across different loci, required for the building and the functioning of ESX-1, the transport of virulence factors, and their membrane lysis activity. The ESX-1 GOI encodes 34 proteins (ESX-1 related proteins) that can be divided into four functional categories: 7 substrates (products that are secreted during virulence), 6 core complex (proteins part of the secretion machinery), 8 regulators (transcription factors), and 14 peripherals (exact function yet to be determined). The structures of two related T7SS inner membrane core complexes of MTB have been elucidated: ESX-3 (Famelis et al. 2019) and ESX-5 (Bunduc et al. 2021). The 3D structure of a few of these proteins (in isolation or as part of a complex) have been elucidated and validated, whereas predictions of all individual 3D protein structures have become available through artificial intelligence (AI) techniques (Baek et al. 2021; Jumper et al. 2021). Using the predicted 3D structures of individual proteins and knowledge of homologous protein complexes and interacting interfaces one could propose models for the quaternary structures of known interactions within the set of 34 ESX-1 related proteins.

Despite their high genomic similarity, MTBc lineages differ meaningfully in the host immune response, host tropism, phenotypes, drug resistance, and transmissibility (Brosch et al. 2002; Brosch et al. 2007; Wirth et al. 2008; Peters et al. 2020). So far, most research on the ESX-1 machinery has focused on MTBc L2 and L4 because of their widespread geographic range and availability of laboratory-adapted reference strains such as H37Rv (L4) and HN878 (L2). Thus, our knowledge on the genomic differences and convergence in the ESX-1 genes across the MTBc is limited.

To obtain a deeper understanding of virulence across MTBc, we generated a single nucleotide variation catalogue (synonymous and non-synonymous SNPs) in the ESX-1 GOIs, for all MTBc lineages (L1–L8) of human importance, using whole genome sequencing (WGS) data from >32,000 publicly available clinical isolates. First, we provide evidence illustrating the variation tolerance of the ESX-1 GOI, confirming that it is the most SNV-dense group of genes within the MTB genome. Next, we to identified several genomic regions including variants that arose independently under positive selection as done by Vargas et al (2022). We then examined the amino-acid locations that bear abundant SNV-counts. In 34 ESX-1 genes, only 21 SNPs were found in more than 1% of the isolates: these are considered successful fully functional transmission events. None of these resulted from convergence but either due to opportunistic sampling of the dataset or occurring in ancestral lineages. The data also reveal which parts of the 34 proteins are fully conserved. We mapped all SNVs onto each of the known or predicted 3D structures and inspected its surfaces. Varying sizes of conserved regions were found, some proteins showed clear polarity (of SNPs distribution). We then zoomed in on some essential motifs (such as Walker motifs, secretion signals, SS-bonds), and found <0.001% of isolates bearing SNVs in those locations. Next, we analyzed known and predicted quaternary structures, correlated interaction interface with SNV distribution maps, and experimentally validated that certain mutations on the interaction interface which still permit complex formation. We highlight a diverse set of conserved protein surface regions and hypothesize new interaction partners for these. Finally, we scrutinized the intrinsically disordered regions (IDRs) within our set of proteins, found two long stretches that are fully conserved, and discuss their potential essential role in immunological recognition. Combined, our findings highlight new targets for interfering with MTBc virulence.

Results

Consolidating SNPs for ESX-1 GOI

A collection of 32,399 unique MTBc isolates, including clinical MTBc isolates (Vargas et al. 2022) L1–L6 (NCBI), L7 (Chiner-Oms et al. 2019), as well as *M. bovis* (unpublished data), was collated (Figure 1, Supplementary Figure 1). The clinical isolates from human TB patients were considered a ‘filter’ for fully functional virulence. For the ESX-1 GOI, we examined 34 genes encoding 34 proteins with a total of 11,167 amino acids, for which a total of 8616 had a SNV, including 2742 synonymous mutations (sSNPs) and 5874 non-synonymous mutations (nSNPs). Almost 40% of the encoded amino acids had at least one nSNP. Figure 1 (and Supplementary Figures 3-6) visualizes the SNVs mapped on the predicted AlphaFold structures for each of the 34 ESX-1 GOIs.

Converging evolution

From convergence analysis, two silent SNPs in *espI* in four independent lineages (L1, L4, L3, and L8), in a total of ten unrelated clinical isolates were identified and confirmed by Sanger sequencing (Supplementary Table 2). EspI represses ESX-1 activity under conditions of lowered cellular ATP levels in the MTBc and may play a role during latent tuberculosis infection and reactivation (Zhang *et al.*, 2015). In the isolates with two silent SNPs in espI, a repeat of 6 cytosines is formed. In cancer, this type of motif has been linked to changes in methylation (Dogan et al. 2017). Long PacBio reads could provide information about whether these mutations affect the methylation state of ESX-1. RNA-seq (gene expression profiling) on the mutant EspI strains followed by eQTL analysis will allow answering to what extent the SNPs in espI are biologically significant.

Conservation on 3D level

We mapped all SNPs onto each of the known or predicted 3D structures and inspected the surfaces. Varying sizes of conserved regions were found, as shown by dominant green color of the predicted protein structures and the small number of red hotspots (Figure 2 and

Supplementary Figures 3-6). Some proteins showed clear polarity, like WhiB6, PhoR, PhoP, DevR, MprA and MprB display SNPs on one side of their surface, while the opposite side showed no SNPs (Supplementary Figure 3).

Comparing the prevalence of SNPs across the 34 ESX-1 GOI, we found that regulators WhiB6 (65.2%) and PhoR (54.2%) as well as two members of the ESX-1 inner membrane core complex EccB1 (54.4%) and EccE1 (53.8%) had the highest percentage of amino acid changes. In comparison, the regulator DevR was most conserved, with only 13.5% of its amino acids showing any mutation, and 98% of its SNPs were found at one position (see next section). This trend of clustered SNPs to one amino acid was also observed for *esxB* (96%), *PE34* (99%), *espA* (91%), *mprB* (94%), *mprA* (83%), *espH* (78%), *espE* (77%), and *espL* (69%) (Supplementary Figures 3–6).

Hotspots / abundant SNPs in ESX-1 GOIs

After excluding polymorphisms due to convergent evolution, we identified 78% of the nSNPs in so-called hotspots (the same mutation found in >3000 clinical isolates, >1% of the isolates), reflecting a phylogenetically more ‘basal’ polymorphism with clonal expansion. We identified 21 hotspots (Figure 2), which involved 14 of the 34 ESX-1 GOI, involving all lineages (Supplementary Table 3). Only 4 proteins (PPE68, EccD1, EccE1, and EspK) showed more than 1 hotspot, while 9 (PhoR, MprA, MrpB, DevR, EspA, EccB1, EsxB, EspH, and EspB) each showed a single hotspot. Two other abundant SNPs were excluded as hotspots (T192I in EspA and E99* in PE34), since they included almost the entire dataset, suggesting that the H37Rv reference genome has a polymorphism, whereas one (L339H in MprB) included half the dataset. The hotspots occurred in all protein categories: we found them in 4 regulators, 3 peripheral, 4 substrates and 3 core components. Upon mapping the hotspots onto the protein structures, we found them predominantly on the surface of the protein, and to include both polar as well as apolar residues (Figure 2). Visual inspection of the experimental as well as predicted protein structures of these 14 proteins and 21 hotspots showed very few polar and H-bond interactions of the hotspot side chains within the protein itself (data not shown). Future identification of potential compensatory mutations may explain how these clonally expanded mutations are non-detrimental ~~nor with other subunits for those proteins where experimental complex structures were available (more below).~~ For example, in EsxB the most prevalent mutation was identified as E68K. Chimera predicted it to have minimal impact on the binding site between EsxB and EsxA. To further investigate this finding, we are currently cloning a mutant strain and assessing its protein interaction with EsxA.

Motifs and binding regions

Next, we analysed the SNPs found for the two motifs known to be essential for secretion in the 7 substrates: WxG and YxxxD/E. Overall, these secretion motifs were found to be highly conserved. Within EsxB, only one SNP was found in WxG and two in YxxxD/E. Within EpsJ, the YxxxD/E motif had 11 total SNPs. For YxxxD/E, and EspH had 6 SNPs EspF had 1. The YxxxD motif is fully conserved in EspA and EspC.

Some of the ESX-1 core components contain ATPase domains. We analysed the Walker A motifs in EccA1, EccCa1 and EccCb1. The Walker A motif, also called the "P-loop" for its phosphate binding, has the classical pattern (G/A)xxxxGK(T/S) (Allemand, Maier, and Smith 2012). Our data show a total of 4 SNPs within this motif for the total of five ATPase domains found in EccCa1 and EccCb1, no SNP found in EccA1 motif (Supplementary Table 4). The Walker B motif (hhhhD, where h is any hydrophobic residue) is fully conserved or has allowed changes in all three proteins. We next analysed the conservation of disulfide bonds. EccB1 has a disulfide bond within its periplasmic domain (CysXX-CysYY): it is fully conserved. Cys48 in substrate EspC has been described to play a role in EspC polymerisation and subsequent dimer and disulfide bond formation (Lou et al. 2017). No mutations were found for this residue. Finally, we evaluated the Fe-S cluster in regulator whiB6, to which NO can bind as part of the innate immune response to mycobacterial infection. This binding inhibits the secretion of ESAT-6 and CFP10. We found no SNPs in the binding pocket of this Fe-S cluster. As for the inner membrane core complex ESX-1 genes, it appears that the protein tails of EccCb1, EccD1, MycP1, are all SNP free. We modelled the ESX-1 core complex as a trimer of dimers and colored all its components according to the number of SNPs. Zooming out on the complex, it seems to contain mutation-free pockets. As of today, there is no experimental structure known of this complex. Our dataset reveals a large overall number of mutations within each of its components. These might include mutation-pairs, in which interactions are maintained using different pairs of amino-acids. Detailed analysis of those awaits better experimental insight in its structure. Together, these data demonstrate that functional regions within the set of 34 ESX-1 GOI are largely conserved.

Intrinsically disordered regions (IDRs)

The predicted AlphaFold structures contain regions of low and very low model confidence scores (respectively, <70 and <50). The percentage of low-confidence regions varies for different species, and is relatively small for bacterial proteomes such as *M. tuberculosis*

(13.29%) (Aderinwale et al. 2022). We analysed whether these regions of low confidence in the proteins encoded by the ESX-1 GOIs reflect intrinsically disordered regions (IDRs) (Piovesan, Monzon, and Tosatto 2022), which tend to be hydrophilic unstructured protein structures thought to be disproportionally involved in interactions. Thirteen of the 34 ESX-1 GOI proteins contained at least one IDR, as defined both by low pLDDT confidence score and IDR-analysis schemes; Supplementary Figure 3). In particular, all seven substrates have large stretches of IDRs: EspA, EspE, PPE68, EspI, EspJ, EspB and EspK. Few nSNPs and no hotspots were found within these regions. We analysed the percentage of mutated amino acids within the IDRs of each of these proteins and found a strikingly low percentage within IDRs of EspE, EspI, EspK and EspB (Figure 3). EspB even displayed a long, entirely nSNP-free IDRbetween the structured N-terminal domain (residues 1–262) and the MycP cleavage site (residue 345); a second IDR follows that maturation site and contains a few nSNPs. MycP1 cleaves off this second IDR within the periplasm, after the translocation of EspB through the ESX-1 inner membrane core complex. Structural studies, including macromolecular crystallography and cryo-electron microscopy, could not reveal a structural insight about the first IDR (residues 262–375).

Quaternary structures

Because the T7SS uses several protein complexes for the transport of virulence factors in pathogenic mycobacteria, we also inspected the nSNPs within the context of quaternary structures to obtain insight into the fidelity of these interfaces for interaction. Both known and putative complex interfaces were examined using the PISA interface tool (Krissinel and Henrick 2007). For each MTB complex interface, we determined the number of SNPs for each interacting residue. Our analysis revealed interesting variations in SNP counts among different protein-protein and protein-DNA interactions (Supplementary Tables 7 -12) (Figure 4). In addition to quaternary structure interactions, we determined the presence of SNPs in experimentally verified regions where several ESX-1 substrates interact.

The main substrates ejected from the phagosome by ESX-1 are EsxA and EsxB, known to form a dimer before being secreted. Both proteins have only 2 residues with 2 SNPs detected (Supplementary Table 7). Surprisingly, when analyzing the self-interaction of EspB and PhoR, where we count total of <50 SNPs (in >30k isolates), in <25% of interacting residues (Supplementary Tables 8-9). This observation suggests the presence of natural variations within the EspB and PhoR proteins, potentially indicating functional diversity among MTB strains. Further studies are needed to elucidate the impact of these variations on the overall virulence and pathogenicity of MTB.

Furthermore, in the interaction between EspK and EspB, we observed only 5 residues with SNPs count 1-15 SNPs. The other 10 residues known to interact had no SNPs detected in 32k isolates, suggesting a high degree of conservation in this interaction rendering stability and fidelity of the EspK-EspB complex (Supplementary Table 10).

Regarding the interaction between regulators such and EspR and PhoR with DNA, we see complete conservation in all 11 interaction sites (7 residues in PhoR – DNA and 4 residues in EspR and DNA; Supplementary Tables 11-12). The conservation of this interaction site indicates its functional importance for proper gene regulation and underscores its significance in the context of MTB's virulence mechanisms.

Discussion

These studies demonstrate that 34 genes encoding ESX-1 proteins were somewhat conserved. For example, *espF* had 44, *esxA* had 39, and *espC* had only 37 SNVs across their respective sequences. In cases where >50% of the amino acid positions had ≥1 SNP, we define it as *mutated protein*. Of the 34 ESX-1 GOIs analysed, only 6 mutated proteins were identified: EccB1, EccD1, EspJ, PhoR, Rv3613c and WhiB6.

When examining the distribution of SNPs across the ESX-1 GOI proteins, captivating patterns emerge, which can shed light on both known and unknown proteins' functions and interactions. Mapping these onto experimental and predicted 3D structures revealed biased SNV distribution patterns. Mutation hotspots seem to occur mostly at places on the protein surface for which no interactions have been reported so far. The observed clustering of SNPs to a single amino acid was not surprising (Supplementary Figures 2–3), as these are known to be prone to polymorphisms. The hotspots suggest that these regions can vary without functional consequence, and thus have no evolutionary pressure to remain stable.

Interestingly, IDRs also harbored relatively few SNPs (Figure 3). This sequence conservation across 34 genes from >32,000 isolates indicates that these protein regions play crucial roles in the virulence of *M. tuberculosis*.

Conservation plays a crucial role in maintaining the integrity and functionality of proteins involved in the ESX-1 virulence mechanism of MTB. One important aspect of conservation is the absence of SNVs in the interaction sites of these proteins.

Most proteins have signature sequences or motifs that are characteristic of protein families, these motifs represent an important feature in the protein structure or function. ESX-1 is a molecular motor that pumps proteins through mycobacterial membranes using the chemical energy of ATP hydrolysis. The Walker A and B motifs are motifs founds in ATPases and GTPases involved in the nucleotide binding and hydrolysis. Any variation in the sequence could potentially affect any of these two capacities, for that reason, it is not surprising the conservation level observed in ESX-1 ATPases. It is noteworthy that even ATPase 2 and 3 domains, which correspond to EccCb1, are conserved even though it has been suggested that these domains are not catalytically active (ref). This idea is based on residues orientation in EccC structure from Thermomonospora curvata and the fact that the crystal structure included ATP molecules in these domains. However, it is possible that those observations do not represent what happens in all EccCb's as we have evidence of ATPase activity in EccCb1 from *M. tuberculosis*). In all >32k samples, four variations in the Walker A glycines were found, either for an alanine or arginine. These residues are involved in the structural integrity of the motif or the coordination to the ATP β-phosphate, which could be affected with those changes; however, it is important to note that the actual impact of this mutation would depend on the specific protein, its three-dimensional structure, and its functional context. Experimental studies and structural analyses would be necessary to determine the precise consequences of such a mutation on the protein's activity.

Further investigation is necessary to determine the functional consequences of these variations and their potential implications for MTB's ability to manipulate host responses.

Overall, our analysis of SNP counts in specific interactions within the ESX-1 virulence mechanism of MTB provides valuable insights into the conservation and genetic diversity within these critical protein-protein and protein-DNA interfaces. These findings contribute to our understanding of the functional significance of these interactions and their potential implications for MTB's pathogenicity. Many proteins contain signature sequences (motifs) that are characteristics of a protein family. These signature sequences are part of important structural or functional domains.

Methods

Whole genome sequencing (WSG) and phylogenetic analysis of 971 clinical isolates

A total of 971 genomes of clinical isolates from the MTBc (*n*=965; L1–L8) and from *M. bovis* (*n*=6) were included, from Bangladesh and Gambia (Comas *et al.*, 2013; Lempens *et al.*, 2020; Ngabonziza *et al.*, 2020). We used this data set as the initial backbone phylogeny for the rest of the analysis. The semi-automated MTBseq pipeline was used for reads mapping and variant detection (Kohl *et al.*, 2018). The output of MTBseq was used to generate an SNP alignment using an in-house Python script (https://github.com/alxndravc/ESX-1-MS). Based on this SNP alignment, a maximum likelihood tree was built using RaxML-NG (Stamatakis, 2014) with a GTR + CAT model of evolution and 100 bootstraps; *M. canetti* was added as an outgroup. The different phylogenetic lineages were visualised using the online interactive Tree Of Life (iTOL) tool (Letunic and Bork, 2019).

Phylogenetic analysis of 31,428 publicly available MTBc isolates

We used a SNP barcode (Freschi *et al.*, 2020) to type a collection of 31,428 WGS MTBc isolates downloaded from NCBI into MTBc lineage and sub-lineage. We excluded isolates that were missing 10% of SNP sites, were not typed as belonging to MTBc L1–8 or were typed as L4 but were not typed further with an L4 sub-lineage. We split the 31,428 isolates into eight groups based on genetic similarity, five groups corresponding to global L1, L2, L3, L5, L6 and three groups for lineage 4 (i.e., L4.1.12). To generate phylogenies for each of these groups, we first merged VCF files of the isolates in each group with bcftools (Li *et al.*, 2009). We then removed repetitive, antibiotic resistance and low-coverage regions (Freschi *et al.*, 2020). We generated a multi-sequence FASTA alignment from the merged VCF file with vcf2phylip (version 1.5). Finally, we constructed the phylogenetic trees for each group with IQ-TREE 1.6.12 (Nguyen *et al.*, 2015). We used the *mset* option to restrict model selection to GTR + CAT models and selected the GTR+F+I+R model for the six isolate groups corresponding to L1–L4 and implemented the automatic model selection with ModelFinder Plus (Kalyaanamoorthy *et al.*, 2017) for the isolate groups corresponding to L5 and L6.

SNP catalog

An in-house Python script was used to count the unique SNPs present within each isolate group (stratified by global lineage) within the ESX-1 GOIs (https://github.com/alxndravc/ESX-1-MS). To calculate the SNP frequency per 1 kb, the number of unique SNP locations per gene were multiplied by 1000 bp and divided by their respective gene length. The same methodology was used on the 31,000 WGS isolates from NCBI. The genes were divided into 4 groups:

machinery, substrates, regulatory and peripheral (i.e. genes not belonging to any of the other three categories) (Supplementary Table 1). One-way ANOVA with posthoc Tukey was used to determine between which pairs of means there is a significant difference ($P < 0.05$), excluding the genes in the unknown group. Z-scores were determined as standard deviations from the mean for each of the genes, with a cut-off set at 1.5 to indicate genes with significant SNPs. To count the SNPs that represented at least 20% of the total isolates per lineage, SAMtools and IGV were used. The IGV coverage allele-fraction threshold was set at 0.2, i.e., if a nucleotide differs from the reference sequence in greater than 20% of reads, IGV colors the bar in the coverage bar chart in proportion to the read count of each base (A, C, G, T).

AlphaFold2-generated protein 3D models were collected from the EBI/AlphaFold collection of models built upon the UniProt database (version 4, last accessed in 20/7/2022, available at https://alphafold.ebi.ac.uk), using their corresponding UniProt reference. For each model, pLDDT scores per amino-acid position were extracted from the PDB files (b-factor column), and nSNP counts per amino acid were assigned into corresponding attribute files prior to their rendering with UCSF Chimera 1.15 and/or ChimeraX 1.4 (ref, websites). Singletons were filtered out by setting up the following color oding scheme: green for 0 or 1 nSNP, purple for a minimum of 2 nSNPs, red for the 1% outliers (>300 nSNPs).

nSNPs vs pLDDT vs IDR plots

Putative Intrinsic Disordered Regions (IDRs) of the 34 ESX-1 Proteins of Interest (POIs) were predicted in batch using the OdiNPred server (Prediction of Order and Disorder by evaluation of NMR data, https://st-protein.chem.au.dk/odinpred) without evolution and the Disorder Probability (DP) scores were retained for further processing. nSNP counts, pLDDT scores and DP scores were assigned to each amino acid position, per POI, and DP scores were normalized to % for plotting purpose. Figures were generated with SigmaPlot 12.5 (Systat Software): nSNPs counts per AA were displayed on a logarithmic scale (1 to 40000) to visually filter singletons out (left axis) while pLDDT and DP scores per AA were displayed on a 0-to-100% scale, (right axis).

SNPPar analysis homoplasy

The resulting phylogenetic tree from the 971 clinical isolates was used in SNPPar along with the SNP dataset to obtain the mutation events across all MTBc lineages. To screen for convergent SNP sites in the alignment, SNPPar was used. Based on the provided phylogenetic tree, SNPPar searches for SNPs that are the same mutation (e.g., C G) at the same position in two or more unrelated isolates or different mutations that result in the same base (e.g., C G, A G) on the same position. It also detects revertant mutation back to the ancestral state (e.g., C G C) (Edwards *et al.*, 2020). We used the default settings of SNPPar, which is a TreeTime for ancestral state reconstruction (ASR). As input, the phylogenies mentioned above were used together with the H37Rv reference genome in Genbank format (NC_000962.3) and an SNP position file. On the large dataset of 31,428 publicly available samples, SNPPar was run 8 times on 8 independent sets of isolates corresponding to 8 genetic backgrounds (L1–L6). Lineage sample counts are as follows: L1: 2,815; L2: 8,090; L3: 3,398; L4A: 5,839; L4B: 6,958; L4C: 4,134; L5: 98; and L6: 96.

Quarterly structure interfaces analysis

Understanding the molecular interactions and stability of complexes in Mycobacterium tuberculosis (MTB) is crucial for deciphering its pathogenesis and identifying potential therapeutic targets. In this study, we provide a detailed analysis of experimentally verified MTB complexes, focusing on their interface characteristics, energy stability, and the presence of single nucleotide polymorphisms (SNPs) (Supplementary Tables 7-12).

Sanger sequencing of espI

To validate the two *espI* silent SNPs (Pro134/135Pro), primers were designed for Sanger sequencing on the DNA extracts of available mutant isolates (i.e., containing the SNPs) and phylogenetically closely related wild-type (WT) isolates. A list of used genetic isolates and primer sequences can be found in Supplementary Table 2.

Figures

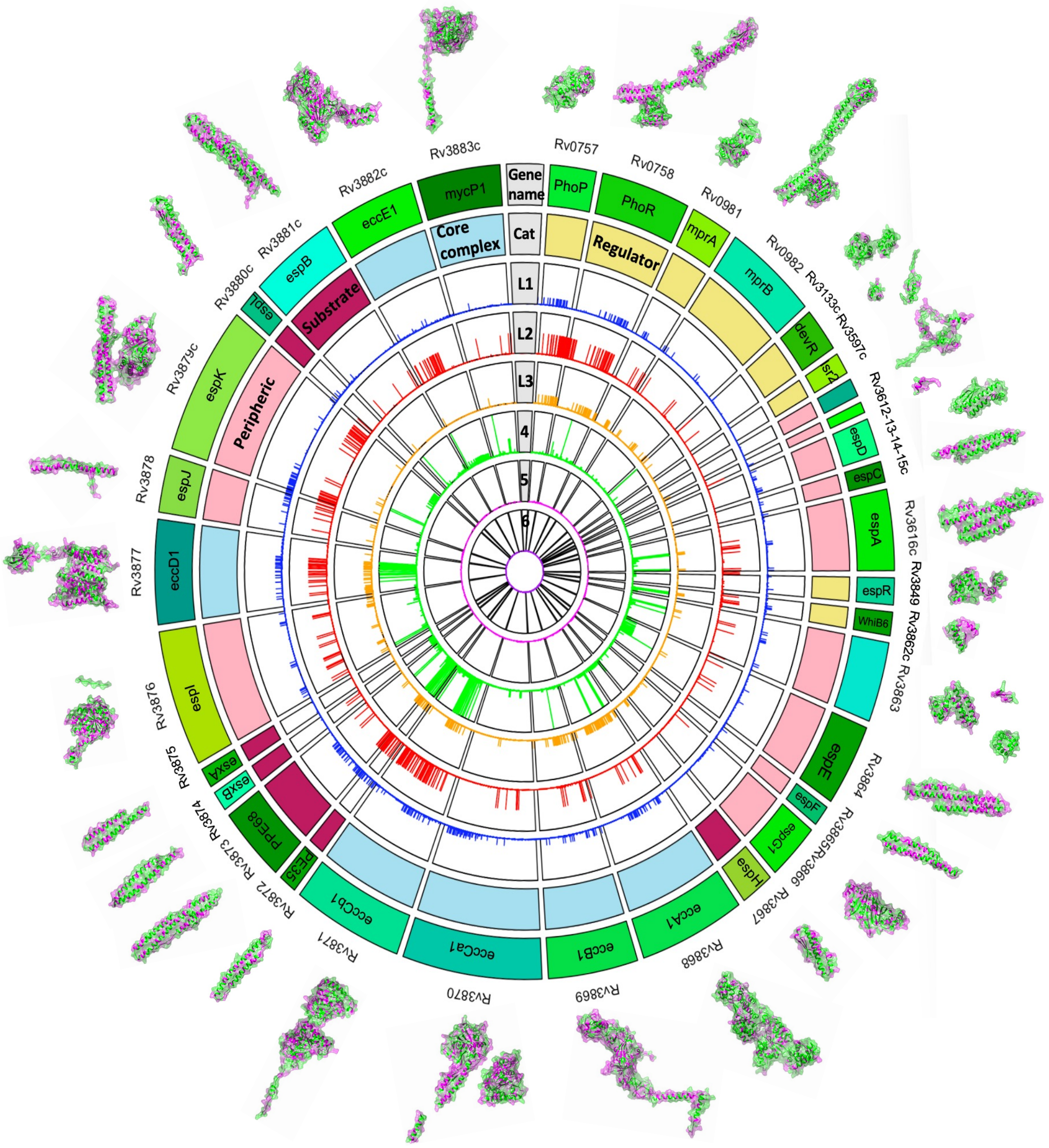

Figure 1. SNPs catalogue for 35 ESX-1 GOIs for each of the MTBc lineages. Circos plot visualization of distinctive nSNPs counted in the data was made from merged variant files (MTBseq pipeline) and then filtered by uniqueness per lineage (not shared with other lineages). The outer lane depicts gene names. Inner lane is color coded by functionality (purple – *substrates genes*, cream white for regulators, pale green for machinery coding genes, pale blue for peripheral coding genes. The next 7 inner lanes designated for lineage stratification (L1 – L6) (we exclude L9 due to small sample size (n=1)).

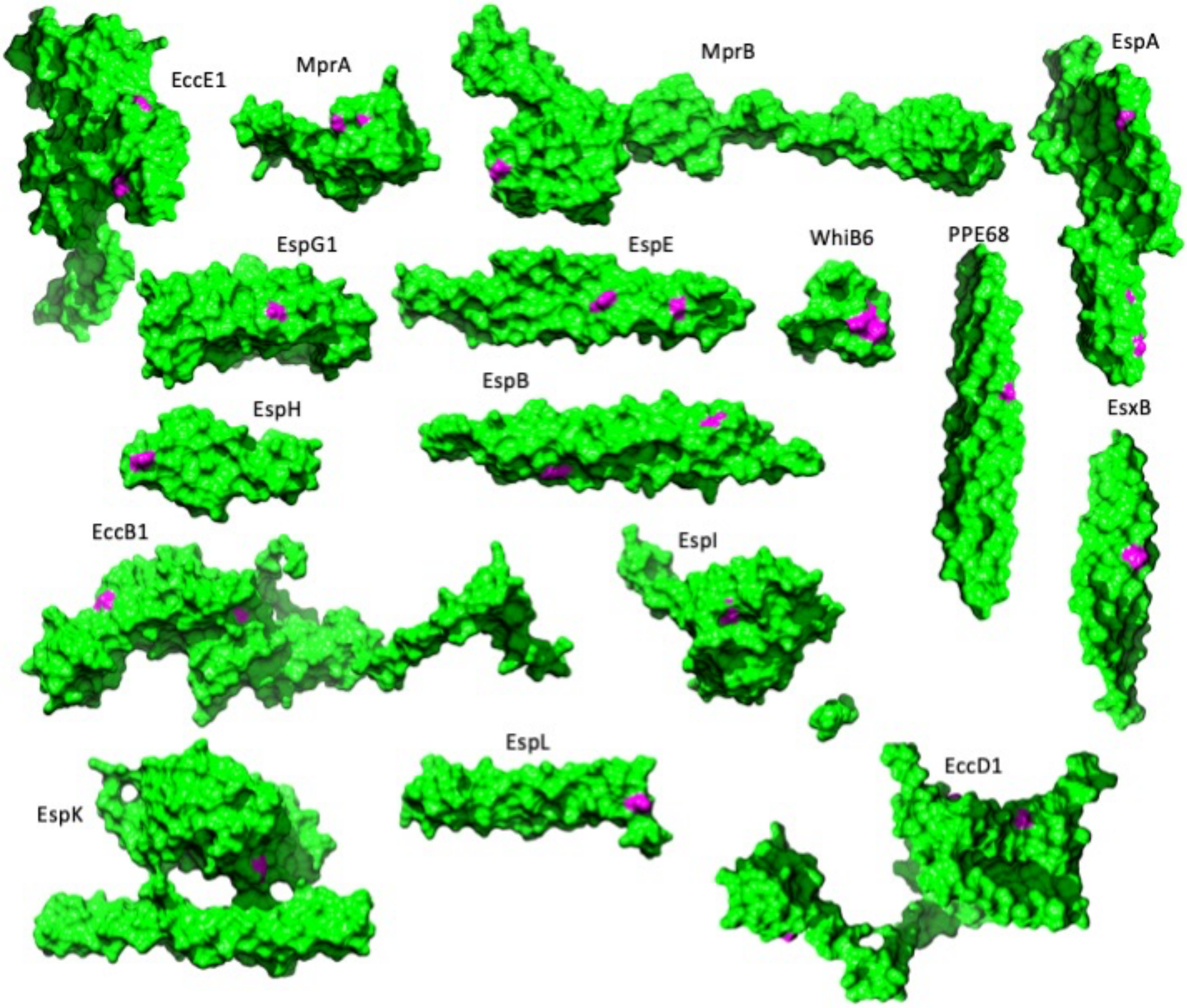

Figure 2. SNPs Hotspots. A hotspot is when the same mutation found in >300 clinical isolates (>1% of the isolates).

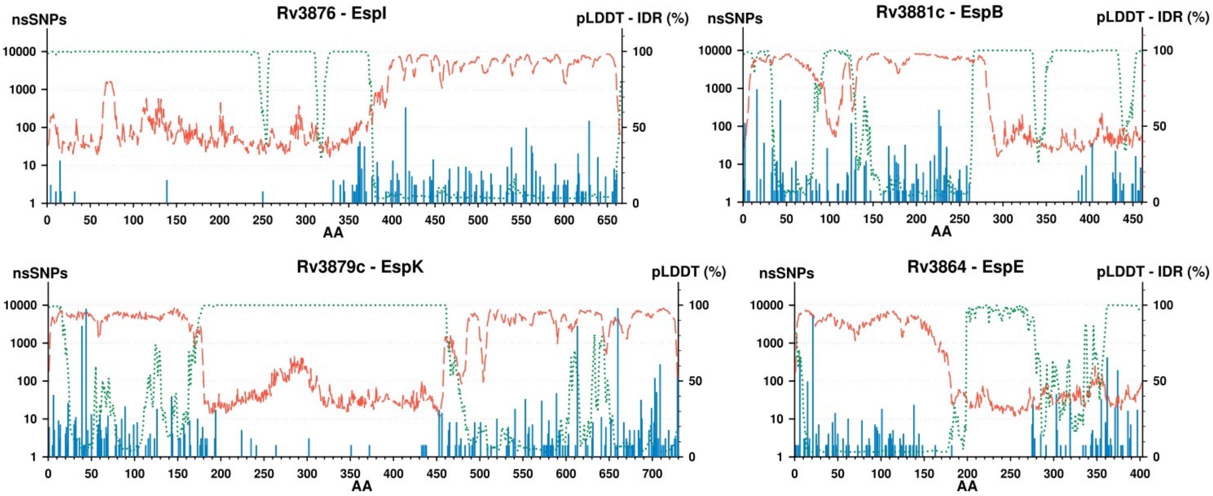

Figure 3: ESX-1 components with the most conserved predicted IDRs.
The amino-acid sequences of the ESX-1 proteins of interest were analyzed with the AlphaFold2 and ODiNPred algorithms. The proteins with the longest low-pLDDT, high-disorder scores (IDR), low nSNP frequency stretches were identified as EspE (>100 AA), EspB (<150 AA), EspK (~250 AA) and EspI (<350AA). Per amino-acid position: nSNPs counts, AlphaFold2 confidence and ODiNPred disorder scores were plotted with - respectively - vertical bars (blue, left axis, logarithmic scale), dashed red (pLDDT score) and green dots (IDR score) lines (right axis, 0-to-100% scale).

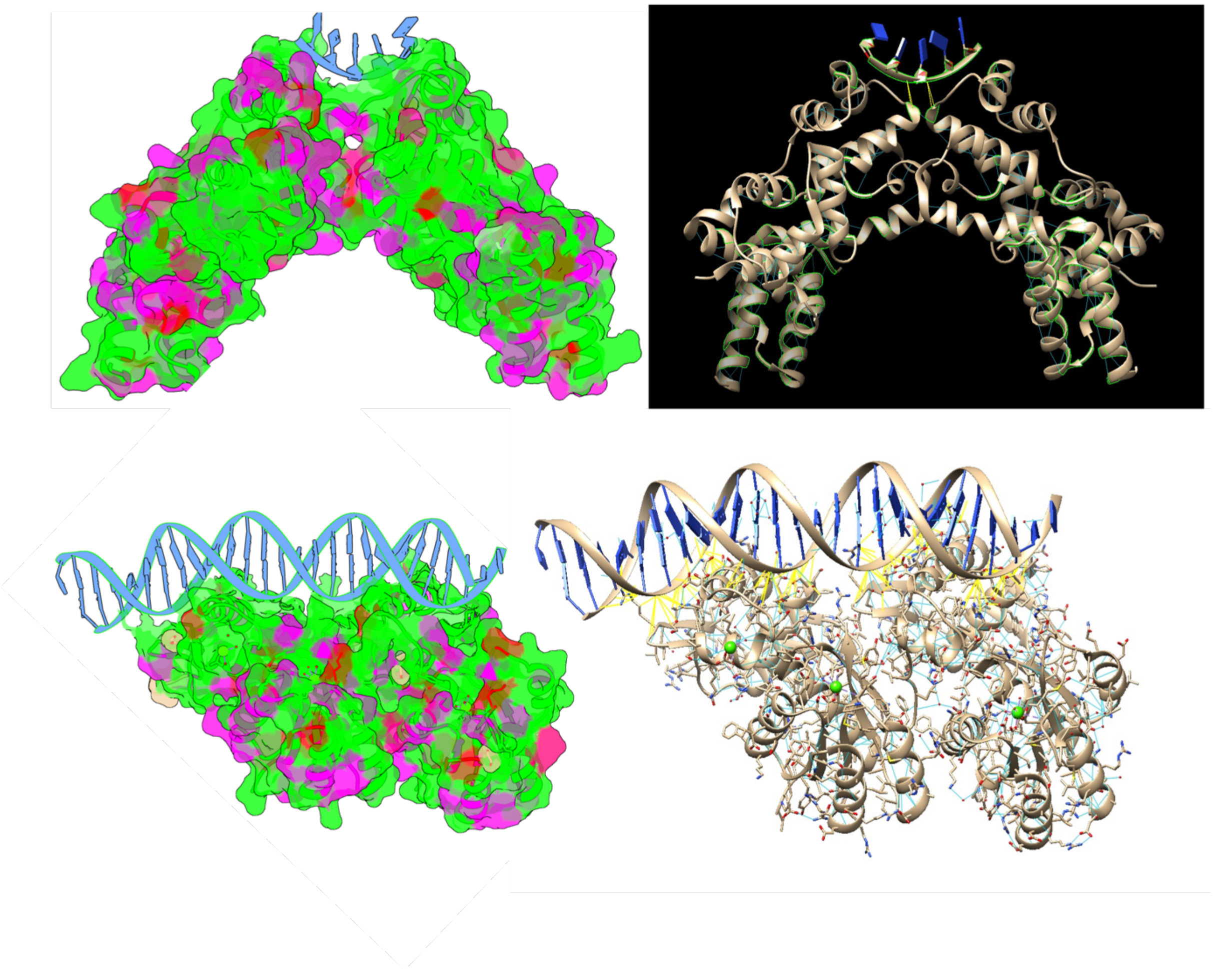

Figure 4. EspR (top) and PhoR (bottom) display interfaces with DNA clean of SNPs

References

Achtman, Mark, and Michael Wagner. 2008. 'Microbial diversity and the genetic nature of microbial species', *Nature Reviews Microbiology*, 6: 431-40.

Aderinwale, Tunde, Vijay Bharadwaj, Charles Christoffer, Genki Terashi, Zicong Zhang, Rashidedin Jahandideh, Yuki Kagaya, and Daisuke Kihara. 2022. 'Real-time structure search and structure classification for AlphaFold protein models', *Communications Biology*, 5: 316.

Allemand, Jean-François, Berenike Maier, and Douglas E. Smith. 2012. 'Molecular motors for DNA translocation in prokaryotes', *Current Opinion in Biotechnology*, 23: 503-09.

Andersen, P., A. B. Andersen, A. L. Sørensen, and S. Nagai. 1995. 'Recall of long-lived immunity to Mycobacterium tuberculosis infection in mice', *Journal of immunology (Baltimore, Md. : 1950)*, 154: 3349-72.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. 2021. 'Accurate prediction of protein structures and interactions using a three-track neural network', *Science*, 373: 871-76.

Böddinghaus, B., T. Rogall, T. Flohr, H. Blöcker, and E. C. Böttger. 1990. 'Detection and identification of mycobacteria by amplification of rRNA', *Journal of Clinical Microbiology*, 28: 1751-59.

Bosserman, Rachel E., Kathleen R. Nicholson, Matthew M. Champion, and Patricia A. Champion. 2019. 'A new ESX-1 substrate in M. marinum that is required for hemolysis but not host cell lysis', *Journal of Bacteriology*.

Brodin, Priscille, Ida Rosenkrands, Peter Andersen, Stewart T. Cole, and Roland Brosch. 2004. 'ESAT-6 proteins: protective antigens and virulence factors?', *Trends in Microbiology*, 12: 500-08.

Brosch, R., S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L. M. Parsons, A. S. Pym, S. Samper, D. van Soolingen, and S. T. Cole. 2002. 'A new evolutionary scenario for the Mycobacterium tuberculosis complex', *Proceedings of the National Academy of Sciences*, 99: 3684-89.

Brosch, Roland, Stephen V. Gordon, Thierry Garnier, Karin Eiglmeier, Wafa Frigui, Philippe Valenti, Sandrine Dos Santos, Stéphanie Duthoy, Céline Lacroix, Carmen Garcia-Pelayo, Jacqueline K. Inwald, Paul Golby, Javier Nuñez Garcia, R. Glyn Hewinson, Marcel A. Behr, Michael A. Quail, Carol Churcher, Bart G. Barrell, Julian Parkhill, and Stewart T. Cole. 2007. 'Genome plasticity of BCG and impact on vaccine efficacy', *Proceedings of the National Academy of Sciences*, 104: 5596-601.

Bunduc, Catalin M., Dirk Fahrenkamp, Jiri Wald, Roy Ummels, Wilbert Bitter, Edith N. G. Houben, and Thomas C. Marlovits. 2021. 'Structure and dynamics of a mycobacterial type VII secretion system', *Nature*, 593: 445-48.

Chiner-Oms, Álvaro, Michael Berney, Christine Boinett, Fernando González-Candelas, Douglas B. Young, Sebastien Gagneux, William R. Jacobs, Julian Parkhill, Teresa Cortes, and Iñaki Comas. 2019. 'Genome-wide mutational biases fuel transcriptional diversity in the Mycobacterium tuberculosis complex', *Nature Communications*, 10: 3994.

Dogan, Senol, Anis Cilic, Damir Marjanovic, and Amina Kurtovic-Kozaric. 2017. 'Detection of cytosine and CpG density in proto-oncogenes and tumor suppressor genes in promoter sequences of acute myeloid leukemia', *Nucleosides, Nucleotides and Nucleic Acids*, 36: 302-16.

Famelis, Nikolaos, Angel Rivera-Calzada, Gianluca Degliesposti, Maria Wingender, Nicole Mietrach, J. Mark Skehel, Rafael Fernandez-Leiro, Bettina Böttcher, Andreas Schlosser, Oscar Llorca, and Sebastian Geibel. 2019. 'Architecture of the mycobacterial type VII secretion system', *Nature*, 576: 321-25.

Forrellad, Marina A., Laura I. Klepp, Andrea Gioffré, Julia Sabio y García, Hector R. Morbidoni, María de la Paz Santangelo, Angel A. Cataldi, and Fabiana Bigi. 2013. 'Virulence factors of the Mycobacterium tuberculosis complex', *Virulence*, 4: 3-66.

Gijsbers, Abril, Vanesa Vinciauskaite, Axel Siroy, Ye Gao, Giancarlo Tria, Anjusha Mathew, Nuria Sánchez-Puig, Carmen López-Iglesias, Peter J. Peters, and Raimond B. G. Ravelli. 2021. 'Priming mycobacterial ESX-secreted protein B to form a channel-like structure', *Current Research in Structural Biology*, 3: 153-64.

Houben, Diane, Caroline Demangel, Jakko van Ingen, Jorge Perez, Lucy Baldeón, Abdallah M. Abdallah, Laxmee Caleechurn, Daria Bottai, Maaike van Zon, Karin de Punder, Tridia van der Laan, Arie Kant, Ruth Bossers-de Vries, Peter Willemsen, Wilbert Bitter, Dick van Soolingen, Roland Brosch, Nicole van der Wel, and Peter J. Peters. 2012. 'ESX-1-mediated translocation to the cytosol controls virulence of mycobacteria', *Cellular Microbiology*, 14: 1287-98.

Jagielski, Tomasz, Paulina Borówka, Zofia Bakuła, Jakub Lach, Błażej Marciniak, Anna Brzostek, Jarosław Dziadek, Mikołaj Dziurzyński, Lian Pennings, Jakko van Ingen, Manca Žolnir-Dovč, and Dominik Strapagiel. 2020. 'Genomic Insights Into the Mycobacterium kansasii Complex: An Update', *Frontiers in Microbiology*, 10: 2918.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596: 583-89.

Krissinel, Evgeny, and Kim Henrick. 2007. 'Inference of Macromolecular Assemblies from Crystalline State', *Journal of Molecular Biology*, 372: 774-97.

Lewis, Kaeryn N., Reiling Liao, Kristi M. Guinn, Mark J. Hickey, Sherilyn Smith, Marcel A. Behr, and David R. Sherman. 2003. 'Deletion of RD1 from Mycobacterium tuberculosis Mimics Bacille Calmette-Guérin Attenuation', *The Journal of Infectious Diseases*, 187: 117-23.

Lou, Ye, Jan Rybniker, Claudia Sala, and Stewart T. Cole. 2017. 'EspC forms a filamentous structure in the cell envelope of Mycobacterium tuberculosis and impacts ESX-1 secretion', *Molecular Microbiology*, 103: 26-38.

Mahairas, G. G., P. J. Sabo, M. J. Hickey, D. C. Singh, and C. K. Stover. 1996. 'Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. bovis', *Journal of Bacteriology*, 178: 1274-82.

Ngabonziza, Jean Claude Semuto, Chloé Loiseau, Michael Marceau, Agathe Jouet, Fabrizio Menardo, Oren Tzfadia, Rudy Antoine, Esdras Belamo Niyigena, Wim Mulders, Kristina Fissette, Maren Diels, Cyril Gaudin, Stéphanie Duthoy, Willy Ssengooba, Emmanuel André, Michel K. Kaswa, Yves Mucyo Habimana, Daniela Brites, Dissou

Affolabi, Jean Baptiste Mazarati, Bouke Catherine de Jong, Leen Rigouts, Sebastien Gagneux, Conor Joseph Meehan, and Philip Supply. 2020. 'A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region', *Nature Communications*, 11: 2917.

Peters, Julian S., Nabila Ismail, Anzaan Dippenaar, Shuyi Ma, David R. Sherman, Robin M. Warren, and Bavesh D. Kana. 2020. 'Genetic Diversity in Mycobacterium tuberculosis Clinical Isolates and Resulting Outcomes of Tuberculosis Infection and Disease', *Annual Review of Genetics*, 54: 1-27.

Piovesan, Damiano, Alexander Miguel Monzon, and Silvio C. E. Tosatto. 2022. 'Intrinsic protein disorder and conditional folding in AlphaFoldDB', *Protein Science*, 31: e4466.

Sreevatsan, Srinand, Xi Pan, Kathryn E. Stockbauer, Nancy D. Connell, Barry N. Kreiswirth, Thomas S. Whittam, and James M. Musser. 1997. 'Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination', *Proceedings of the National Academy of Sciences*, 94: 9869-74.

Tekaia, F., S. V. Gordon, T. Garnier, R. Brosch, B. G. Barrell, and S. T. Cole. 1999. 'Analysis of the proteome of Mycobacterium tuberculosis in silico', *Tubercle and Lung Disease*, 79: 329-42.

Tiwari, Sangeeta, Rosalyn Casey, Celia W. Goulding, Suzie Hingley-Wilson, and William R. Jacobs Jr. 2019. 'Infect and Inject: How Mycobacterium tuberculosis Exploits Its Major Virulence-Associated Type VII Secretion System, ESX-1', *Microbiology Spectrum*, 7.

Vargas, Roger, Michael J. Luna, Luca Freschi, Kenan C. Murphy, Thomas R. Ioerger, Christopher M. Sassetti, and Maha R. Farhat. 2022. 'Phase variation as a major mechanism of adaptation in Mycobacterium tuberculosis complex', *bioRxiv*: 2022.06.10.495637.

Wel, Nicole van der, David Hava, Diane Houben, Donna Fluitsma, Maaike van Zon, Jason Pierson, Michael Brenner, and Peter J. Peters. 2007. 'M. tuberculosis and M. leprae Translocate from the Phagolysosome to the Cytosol in Myeloid Cells', *Cell*, 129: 1287-98.

Wiens, Kirsten E., Lauren P. Woyczynski, Jorge R. Ledesma, Jennifer M. Ross, Roberto Zenteno-Cuevas, Amador Goodridge, Irfan Ullah, Barun Mathema, Joel Fleury Djoba Siawaya, Molly H. Biehl, Sarah E. Ray, Natalia V. Bhattacharjee, Nathaniel J. Henry, Robert C. Reiner, Hmwe H. Kyu, Christopher J. L. Murray, and Simon I. Hay. 2018. 'Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis', *BMC Medicine*, 16: 196.

Wirth, Thierry, Falk Hildebrand, Caroline Allix-Béguec, Florian Wölbeling, Tanja Kubica, Kristin Kremer, Dick van Soolingen, Sabine Rüsch-Gerdes, Camille Locht, Sylvain Brisse, Axel Meyer, Philip Supply, and Stefan Niemann. 2008. 'Origin, Spread and Demography of the Mycobacterium tuberculosis Complex', *PLoS Pathogens*, 4: e1000160.

Xu, Junjie, Olli Laine, Mark Masciocchi, Joanna Manoranjan, Jennifer Smith, Shao Jun Du, Nathan Edwards, Xiaoping Zhu, Catherine Fenselau, and Lian-Yong Gao. 2007. 'A unique Mycobacterium ESX-1 protein co-secretes with CFP-10/ESAT-6 and is necessary for inhibiting phagosome maturation', *Molecular Microbiology*, 66: 787-800.