bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Resolving the hypotheticome: annotating *M. tuberculosis* gene function through bibliomic reconciliation and structural modeling

Samuel J. Modlin[1], Deepika Gunasekaran[1†], Alyssa M. Zlotnicki[1†], Afif Elghraoui[1], Norman Kuo[1], Carmela K. Chan[1], Faramarz Valafar[1*]

[†]These authors contributed equally to this work
*Corresponding author: faramarz@sdsu.edu

[1]Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence, Biological and Medical Informatics Research Center, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182

Abstract

Each decade, billions are invested in Tuberculosis (TB) research to further characterize *M. tuberculosis* pathogenesis. Despite this investment, nearly half of the 4,031 *M. tuberculosis* protein-coding genes lack descriptive annotation in community databases, due largely to incomplete reconciliation with the literature and a lack of structure-based methods for functional inference. We coin the term "*hypotheticome*" as the set of genes in an organism without known function. For *M. tuberculosis'* hypotheticome, we compiled the set of genes lacking functional assignment in the most frequently used *Mycobacteria* annotation database through systematic, exhaustive manual literature curation and 3D-protein structure-based inference, and reconciled these annotations with frequented functional databases, creating a comprehensive *M. tuberculosis* functional knowledge-base. In doing so, we also introduce standard usage of qualifying adjectives based on quantitative measures of certainty with the hope that this approach is adopted in choosing qualifiers for future functional assignments.

Through these methods we functionally annotated 41.3% of the *M. tuberculosis* hypotheticome, and provide insight into its pathogenesis, antibiotic-resistance, and virulence. Processes implicated in the unique lifestyle of *M. tuberculosis* of long-term persistence and obligate pathogenesis in genotoxic host microenvironments – lipid metabolism, polyketide biosynthesis, and membrane transport and efflux – were overrepresented in our annotation. Our structural similarity approach unturned proteins that appear critical in host-interaction through apparent host mimicry, particularly involving the phagosome and vesicle-mediated transport, as well as putative structural analogs for highly mutable protein classes, including dozens of PE/PPE family proteins which are major players at the host-pathogen interface, and sixteen potential efflux pumps which are integral to *M. tuberculosis* drug tolerance. Hypotheses drawn from these proteins' function may help characterize the onset of latency and identify therapeutic targets. A unified annotation is essential for clear communication about *M. tuberculosis*. These improvements provide the most comprehensive *M. tuberculosis* genome annotation to date, and the approach presented can be applied to systematically annotate the genome of other organisms. We provide our novel annotations in General Feature Format with Enzyme Commission and Gene Ontology terms for integration into existing annotation frameworks.

Keywords: genome annotation, hypothetical genes, H37Rv, structural homology, functional genomics, tuberculosis, host mimicry, convergent evolution

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Basic *Mycobacterium tuberculosis* (*M. tuberculosis*) biology research underlies Tuberculosis (TB) eradication efforts. As genome sequencing improves and becomes more accessible, we have unprecedented capabilities to understand how chromosomal alterations affect mycobacterial physiology and steer evolution of pathogenicity, virulence, antibiotic resistance, and other phenotypic characteristics that challenge effective TB treatment. This opportunity, however, is limited by the quality, quantity, and recency of genome annotation.

Outdated annotations disconnect discoveries about *M. tuberculosis* from what is readily accessible, impeding research progress. Without an up-to-date annotation, laboratories must either maintain an annotation *in villa*, periodically perform time-intensive, exhaustive literature searches, or draw conclusions from incomplete information. While other databases have emerged in recent years[1–5], TubercuList remains the primary source for annotation information[6].

Most studies are performed on *M. tuberculosis* reference strain H37Rv, a descendant of strain H37, isolated from a pulmonary tuberculosis patient in 1905 and kept viable through repeated subculturing[7]. Following sequencing of the H37Rv genome, function was assigned to 40% of its 3,924 open reading frames (ORFs)[8], and in 2002 H37Rv was re-annotated, with 52% of its, then, 4,006 ORFs [9]. H37Rv annotations had continued to be added by TubercuList until March 2013. Despite being nearly five years outdated, TubercuList remains the primary resource for gene annotation for many TB researchers[6].

Research that infers phenotypic features from genomic data is challenged by the quarter of the genome (1,057 genes) completely lacking annotation on TubercuList, listed in “conserved hypotheticals” or “unknown” functional categories. Genes are classified as hypothetical when one or more open reading frames (ORFs) are identified through *in silico* methods, indicating the possible presence of a protein-coding region[10], but whether the gene encodes a functional protein is uncertain[11]. In addition to the 1,057 hypothetical and unknown genes on TubercuList, hundreds of others have product annotations that convey little or no meaning, such as “possible membrane protein”. We refer to the set of these genes collectively as the “hypotheticome”. Several attempts have been made to predict annotations for these genes, drawing from inferential techniques such as protein homology[12,13], protein fold similarity[14], metabolic pathway gap-filling[15], and STRING interactions[16]. However, these predictions require tenuous assumptions, and though useful for hypothesis generation in the absence of experimental evidence, can be incorrect, and produce potentially spurious conclusions.

In addition to *M. tuberculosis* databases, useful annotations can be found in global databases. Most notable of these is UniProt, which contains annotations for many proteins encoded by genes of the hypotheticome[17]. UniProt employs field experts who manually confirm and quality-check experimental characterization of previously unannotated proteins (through SwissProt) according to standardized protocols. UniProt also has a inferential branch, TrEMBL, to give *in silico* predictions of lower-confidence, based primarily on amino acid (AA) sequence identity[17]. Though UniProt contains many useful annotations, their integration into TubercuList and other TB databases is inconsistent and does not distinguish between low-confidence, unreviewed annotations predicted through AA sequence identity and high-quality, manually curated, reviewed annotations. To facilitate a measured interpretation of gene annotations and invoke skepticism where merited, the TB community needs a resource that provides annotations explicit in their source and reliability of evidence.

While current resources regularly incorporate computational annotation predictions via domain homology and sequence similarity, none, to our knowledge, implement large-scale annotation through structural similarity, leaving latent annotations undiscovered. Increasingly sophisticated structural

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

threading algorithms leverage structural and functional conservation to enable annotation of proteins where sequence has diverged. One such tool, Iterative Threading ASSEmbly Refinement (I-TASSER), predicts three-dimensional protein structure from AA sequence by building protein structure models through multiple threading alignment of Protein Data Bank (PDB)[18] templates, followed by iterative fragment assembly simulations[19]. I-TASSER provides superior structural prediction capabilities compared to similar programs[20–24], outputs well-defined estimation of model quality[25] (C-score) and pairwise structural similarity[26] (TM-score), and integrates function and structure prediction tools[27] (COACH and COFACTOR). Integrated functional predictions consist of Gene Ontology (GO) terms[28], Enzyme Commission (EC) numbers[29], and Ligand Binding Sites (LBS)[30].These schemes are widely adopted by enzymologists and bioinformaticians and incorporated into numerous other classification schemes, providing valuable integrability with other representations of biochemical knowledge and genome annotation frameworks.

To infer hypotheticome protein annotations, we classified 1,725 as the *M. tuberculosis* hypotheticome and generated high quality structure models using I-TASSER. These genes comprised 668 uninformatively annotated genes and the 1,057 unknown/hypothetical genes. To annotate function from these structural models, we transferred EC numbers and GO terms from structurally similar proteins of known function in the Protein Data Bank (PDB), which enabled us to name gene products systematically and identify pathways, subsystems, and processes enriched among novel annotations. We then supplemented these annotations with product names derived from manual comparison where putative analogs and homologs in PDB were not given EC and GO assignments and additional structural and ligand-binding site annotations.

We guided this effort with the following aims and philosophies:

1. Provide a comprehensive annotation resource that reconciles TubercuList's last update (March 2013) with all knowledge in the literature (File S1)
2. Use structural similarity as a means of annotation orthogonal to manual curation and sequence similarity to maximize the number of genes annotated with strong predictions while minimizing "overannotation", a problematic phenomenon that perpetuates incorrect annotations[31,32].
3. Use structural similarity to identify proteins most challenging to find through experiment and sequence similarity, such as transport proteins and structural analogs.
4. Assess potential functional implications of newly annotated gene products and functional notes.
5. Highlight the genes that remain uncharacterized, and discuss the common features prohibiting their characterization.

We refer to manual annotations as "*mannotations*" throughout the manuscript to differentiate them from "annotations" in the general sense. For simplicity, we refer to hypotheticome protein structure models that are structurally similar to known structures on PDB as "matches" when it is not clear if they descended from a common ancestor ("homologs") or converged upon similar structure independently ("analogs").

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Approach and scope

To fulfill the study aims, we designed annotation procedures and inclusion criteria to maximize true annotations while minimizing false annotations. Accordingly, we incorporated annotations hierarchically, prioritizing more reliable sources and methods. For resolving candidate annotations within each computational method, we consulted benchmarks of likelihood of correctness conducted in this work and from previous studies. The resulting annotation is, to our knowledge, substantially more complete and transparent than available elsewhere. These annotations were furnished through several thousand person-hours of manual literature curation, quality-assurance, consolidation of annotations from veritable databases, and functional inference through structural similarity according to precision benchmarks, incorporating only confident annotations.

We focused the scope of our annotation on 1,725 genes of unknown function ("GUF") that lacked informative product annotation on TubercuList. These GUF included all genes categorized as "conserved hypothetical" or "unknown" on TubercuList and all remaining ambiguous gene product annotations in other categories (S1 Table). Annotations were considered ambiguous if they were qualified by an adjective connoting low confidence, such as "predicted" or "possible", or if their annotation described only location (e.g. "membrane protein") or responses to a particular stimulus (e.g. "isoniazid-inducible protein A"), but not its immediate function. We also included the sparsely annotated PE_PPE genes as GUFs since most are not functionally characterized.

We first exhaustively searched the literature to include all experimentally proven annotations and identified functional annotations for the GUF, which annotated many GUF, but left the majority (1,448) without annotation. Next, we inferred function from predicted structural similarity with sufficient evidence, but prioritized manual annotations in cases of conflict. Annotations describing purely structural (CATH topology[39], see Methods) or local binding (COFACTOR ligand binding site predictions, described below) properties were included irrespective of gene product annotation since they describe attributes orthogonal to primary protein product function.

An expanded *Mycobacterium tuberculosis* reference genome annotation

Through mannotation and structural inference we annotated 713 GUF with products, shrinking the hypotheticome by 41.3% to 1,012 GUF (Table 1). Between these 713 annotated GUFs and those we annotated with CATH topologies or functional notes, 987 GUF (57.2%) received original annotation, providing more than half of *M. tuberculosis* hypotheticome with novel information with respect to the widely accessed TubercuList[38] (Table 1 and Fig 1). While manual curation presents the strongest evidence and notably expanded the set of annotated genes (283 added), our pipeline for annotation through structural inference permitted an additional 430 GUF to be assigned putative or probable functions. Functional hints were recorded in the form of notes for an additional 274 GUF, derived both from sub-threshold structural similarities, and literature-curated evidence insufficient to assign a product.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Hypotheticome annotation summary.

Annotation method	Genes	Novel	Cumulative	Hypotheticome fraction (%)
Global function				
Manual literature curation	283	283	283	16.4
Structural similarity (EC)	226	174	457	26.5
Structural similarity (GO)	207	207	664	38.5
Structural similarity (transporters & channels)	60	19	683	39.6
Structural similarity (PDB)	42	30	713	41.3
Notes (structural similarity)	25	24	737	42.7
Functional notes (literature)	605	211	948	55.0
Structure				
CATH Topology	373	39	987	57.2
Local function				
Ligand-binding sites	126	8	995	57.7

Product annotations from each method were incorporated only if the gene had not yet had one assigned by a method higher in the hierarchy, except for EC-derived gene product annotations, which were reconciled with manual annotations. Ligand-binding, notes, and CATH structural annotations do not conflict with gene product annotations and thus are included irrespective of gene product annotation. Cumulative gene counts refer to the union of all genes annotated with at least one gene product, CATH topology, Ligand binding site, or functional note. Hypotheticome fractions refers to cumulative genes/1,725 GUF.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

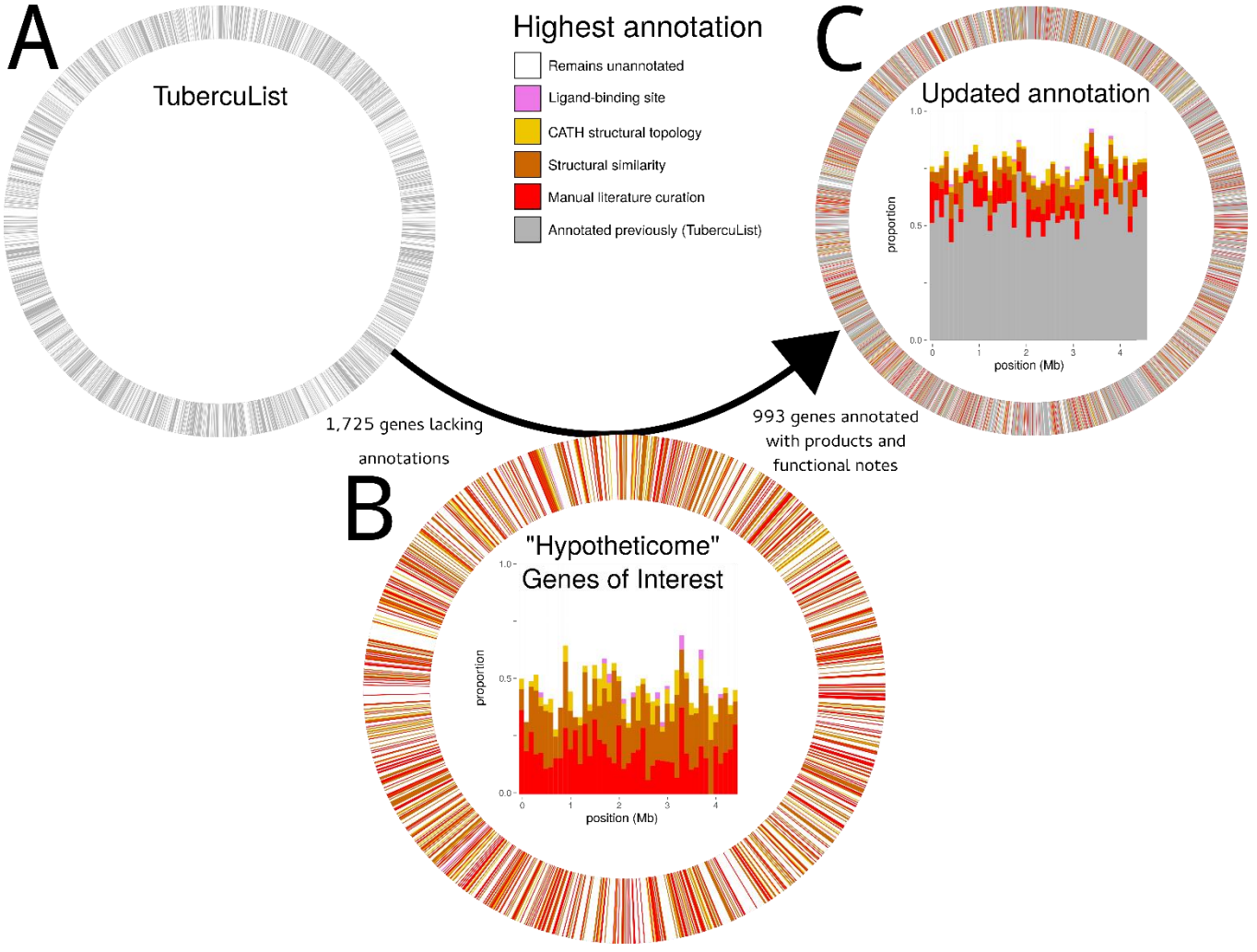


Fig 1. Updated annotations from structure and literature reduce the *M. tuberculosis* hypotheticome.

Circos plots illustrating annotation coverage (A) prior to the annotation effort and following it (C), colored according to annotation status. In plots A and C, all 4,031 CDSs are represented as segments of equal width while (B) segments the ring into only the 1,725 genes of the hypotheticome. Plot A reflects only what is on TubercuList, with the gray genes indicating they were considered "annotated" and are mutually exclusive from the 1,725 genes of the hypotheticome, colored in white. The circos plot of B shows only the 1,725 genes of the hypotheticome, while the circos plot in C merges A and B and includes all 4,031 original CDS. The plots inside of the circos rings in B and C are stacked bar charts with CDSs were split into 100kb bins, according to start position. Height of each color in a bin represents the proportion of genes annotated to that level out of total genes in the bin, and the total height of non-white bar represents total proportion annotated in that bin.

Of the 738 GUF remaining without product annotations or notes (S2A Table), 135 have quality structure models (C-score > -1.5)[19], but could not have annotations inferred through our methods (S2C Table). Meanwhile, 182 of those remaining have annotations conveying product function in UniProt (S2D Table) and/or *Mtb* Network portal (S2E Table), and Remaining still, however, are 427 GUF with no hint of function (S2B Table). Many of these genes cluster consecutively along the genome (105 genes across 15 clusters, S3 Table), indicating potential operons of unknown function.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We classified PDB template hits into candidate annotation categories according to the regression of precision against TM-score and AA% (Methods & Materials). More PDB templates qualified for transfer of lower confidence and specificity thresholds than for higher tiers (S1 Fig). Many templates with high TM-score but low AA% qualified for CATH annotation transfer, which underscores the utility of structure-based annotation in the absence of sequence homology. We then combed through genes with high structural similarity to pull in similar structures not annotated correctly in PDB, as well as proteins disfavored under the combined AA% and structure-based inclusion criteria due to low AA%—transporter proteins[42] and analogs[43].

Several PDB templates were 100% identical to query proteins, representing protein sequences of *M. tuberculosis* or closely related mycobacteria (Supplementary Note). Overall, model quality was high for annotations that passed inclusion criteria; PDB:query relationships meeting criteria for EC, GO or CATH inclusion had a mean C-score of 0.634 and distributed according to their confidence and specificity (S2D Fig). Similarly, model quality of relations not meeting inclusion criteria (gray and black) were lower than the relations meeting any of the inclusion criteria (red, greens, and blues, S2 Fig), demonstrating annotations were derived from high quality structural models, rather than false similarity from noisy structural predictions.

Systematic literature curation increases annotation over existing databases

We compared annotation distributions for common frameworks between frequently cited databases for *M. tuberculosis* to determine which GUF lacked annotation globally. Of these databases, BioCyc and UniProt are the most comprehensive for GO term annotations, while UniProt and Mtb Network Portal have the fewest hypothetical proteins (Table 2).

Table 2. Whole Proteome annotation comparison to the databases commonly referenced for *M. tuberculosis* annotation.

Metric	TubercuList	PATRIC	RefSeq	Mtb Network Portal	UniProt	Kegg	BioCyc
Coding Sequences (CDS)	4038	4367	3989	4038	3997	3906	4031
Proteins with functional assignments	2815	3007	2341	2853	2906	1750	2571
Hypothetical Proteins	1223	1360	1648	1185	1091	2156	1460
Proteins annotated with at least one GO term	2629	969	0	2460	3305	0	3557
Proteins annotated with EC numbers	1293	1074	1081	1003	1138	1050	1018

"Functional assignments" refer to annotations that describe protein function, and excluded hypothetical, unknown/uncharacterized, and PE/PPE family proteins. Counts are current as of May 17, 2017 for RefSeq[34], PATRIC[2] and Mtb Network Portal[5] and current as of June 23, 2017 for KEGG[36] and UniProt[17]. The number of CDSs in KEGG is reported as 3906 as they include only protein coding genes. The annotation for *M. tuberculosis* in KEGG is referenced from TubercuList[38].

Our mannotation produced annotations and functional notes novel compared to TubercuList, as well as the other most frequented databases for *M. tuberculosis* annotations. With respect to the last TubercuList update (March 2013), we mannotated 283 genes with new products: 138 definite, 105 probable, and 47 putative (one gene can have multiple products). Among the products annotated are 122 enzymes, 81 antigens, 28 regulatory proteins, 12 binding proteins, and 55 in various other categories (File S1 and S4 Table). Additionally, 316 genes were annotated with at least one functional note (S5 Table), amassing 599 (34.7% of the GUF) in total. Functional notes include information implicating drug resistance, pathogenesis, virulence, and more.

These genes' functions were diverse, but a few functions were particularly well-represented. Fourteen gene products that reduce oxidative stress were curated, which are critical for the bacillus to withstand

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the oxidative species generated by the host. Also prevalent among mannotation and crucial for cellular and genomic integrity and were proteins mediating RNA and DNA function, tallying 22 in total. Many serine hydrolases were mannotated, and mostly derived from a single study[44] that identified serine hydrolases in non-replicating hypoxic culture, a valuable addition considering their inaccessibility using traditional methods. Also of note are eight transporters/efflux pumps curated through our mannotation, which are crucial to drug tolerance and homeostasis.

We compared these 283 product mannotations to four other frequently cited databases. We initially compared our mannotations to UniProt's because their team of curators follow a standardized manual protein annotation curation protocol, and regularly update annotations[45] (S6 Table). This comparison revealed 139 GUF with mannotations absent in UniProt and an additional 33 GUF with mannotations more complete than in UniProt (S6 Table). We then compared these 172 GUF to Entrez, Mtb Network Portal, and PATRIC, revealing 118 GUF more thoroughly annotated in our mannotation than in any of the databases, and 54 not solely annotated as an antigen (Table 3). In total, 135 GUF received some level of new annotation (68 not solely as antigens).We distinguished genes annotated solely as antigens from those with other annotations because while antigenic properties are the primary function of some proteins, many have different primary functions and are simply recognized by the host.

Table 3. Comparison of manually curated GUF products with annotations novel to UniProt to three other databases.

Gene annotation status	All Annotations	Excluding Antigen Annotations
Annotated genes lacking prior annotation	85	29
Genes annotated more thoroughly than in other databases	33	25
Subtotal: new products	118	54
Genes annotated with additional products compared to other databases	17	14
Total	135	68

The set of 172 manually curated GUF products novel with respect to UniProt were compared to Mtb Network Portal[5], PATRIC[2], and RefSeq[34] and categorized as follows: (1) **Annotated genes lacking prior annotation**: GUF given a new mannotation when its previous annotation was "hypothetical/uncharacterized/unknown protein" or an annotation not descriptive of the protein's function (e.g. PE family protein, which describes a motif, not function). (2) **Genes annotated more thoroughly than in other databases**: Our mannotation contained a more informative product than that in all other databases: e.g. an annotation of "flippase" for a particular GUF in our annotation was considered more specific than annotations of "membrane protein" or "cell division protein" in other databases. (3) **Subtotal: new products**: the union of the first two categories. The set of GUF with more thorough annotation than the databases in the comparison. (4) **Genes annotated with additional products compared to other databases**: GUF with a valid mannotation distinct from annotation in at least one other database. (5) **Total**: the union of all three categorizations as GUF with annotations absent from all databases in the comparison. "All annotations" is inclusive of proteins annotated as antigens: due to the inconsistent nature of the inclusion of antigen annotations or how regularly they are updated, a separate count was produced for each of the five above classifications that excluded these annotations.

Until now, research into the functional implications of these genes would have been uninformed of these functions, regardless of which database the researchers used, missing potentially critical information.
The novel genes annotated in our curation effort are highlighted in Table 4 and can be explored in more detail in S7 Table.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 4. Novel annotations with respect to major databases

Gene	product	PMID	TubercuList	UniProt	Mtb Net Portal
			Novel		
Rv0309	**Adhesin/Putative L,D-transpeptidase**	23922800; 23889607; 26201501	Possible conserved exported protein	Possible conserved exported protein	
Rv0394c	**Hyaluronidase/ Chondrosulfatase**	23465892	Possible secreted protein	Possible secreted protein	Possible membrane protein
Rv0431	**probable vesiculogenesis/immune response regulator**	24248369; 21170273; 17436267; 26324094; 27765619	Putative tuberculin related peptide	Putative tuberculin related peptide	Tuberculin related peptide
Rv1430*	**Esterase**	23383323	PE family protein PE16	PE family protein (PE family protein PE16)	PE family protein
Rv1993c	**putative chaperone**	21925112	Conserved protein	Uncharacterized protein Rv1993c	
Rv2345	**Probable phosphatase**	25782739	Possible conserved transmembrane protein	UPF0603 protein Rv2345	Possible membrane protein
Rv2923c	**probable Osmotically induced bacterial protein C (OsmC, a probable peroxide reductase)**	22088319	Conserved protein	Uncharacterized protein Rv2923c	
Rv2954c	**Probable Methyltransferase**	23536839	Hypothetical protein	Uncharacterized protein	
Rv2969c	**Periplasmic disulfide-bond forming (Dsb) Enzyme**	24100317; 18539140	Possible conserved membrane or secreted protein	Membrane protein (Possible conserved membrane or secreted protein)	POSSIBLE CONSERVED MEMBRANE OR SECRETED PROTEIN
Rv3528c	**probable serine hydrolase**	26853625	Unknown protein	Uncharacterized protein	
			Greater specificity		
Rv0059	**Probable toxin DarT/Probable DNA ADP-ribosyl transferase**	27939941	Hypothetical protein	Uncharacterized protein	
Rv0060	**Probable antitoxin DarG/Probable DNA ADP-ribosyl glycohydrolase**	27939941	Conserved hypothetical protein	Uncharacterized protein	ADP-ribose 1-phosphate phophatase related protein
Rv1337	**Probable rhomboid protease/integral membrane protein**	19165721; 23029216	Probable integral membrane protein	Uncharacterized protein Rv1337	Rhomboid membrane family protein
Rv1357c	**cyclic diguanylate phosphodiesterase**	21151497	Conserved hypothetical protein	Uncharacterized protein Rv1357c	Sensory box/GGDEF family protein
Rv1566c	**probable non-catalytic peptidoglycan binding RipD protein/probable antigen**	24107184; 26481294	Possible Inv protein	Possible Inv protein	Invasion protein
Rv2024c	**Restriction enzyme/m-6-adenine DNA methyltransferase (Mycobacterial adenine methyltransferase B "MamB")**	26704977	Conserved hypothetical protein	Uncharacterized protein	putative helicase
Rv2695	**Probable serine hydrolase**	26853625	Conserved hypothetical alanine rich protein	Conserved hypothetical alanine rich protein	
Rv2991	**probable flavin/deazaflavin oxidoreductase**	26434506	Conserved protein	Conserved protein (F420-dependent protein)	
Rv3036c	**esterase**	25224799	Probable conserved secreted protein TB22.2	Probable conserved secreted protein TB22.2	Possible membrane protein
Rv3354	**protein kinase**	25139900	Conserved hypothetical protein	Lipoprotein	Possible lipoprotein LprJ
			Orthogonal annotation		
Rv0256c*	**B cell Antigen/probable INOS promoter binding protein**	23827809; 28071726	PPE family protein PPE2	Uncharacterized PPE family protein PPE2	Predicted cobalt transporter in Mycobacteria
Rv2204c	**probable serine hydrolase**	26298037; 26853625; 26536359	Conserved protein	Protein Rv2204c	probable iron binding protein from the HesB_IscA_SufA family
Rv3779	**polyprenylphosphomannosyl synthase/galactosaminyltransferase**	21030587; 19717608	Probable conserved transmembrane protein alanine and leucine rich	Membrane protein (Probable conserved transmembrane protein alanine and leucine rich)	

*PE/PPE family protein

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Annotations are separated into those completely novel, those with similar annotations but with greater specificity, and those with an additional, orthogonal annotation compared to what is in the databases of Table 2. Pubmed IDs (PMID) from which annotations for each product were derived are included as well. Members of the PE/PPE family are indicated by asterisk. For the full set of such annotations, see S8 Table.

EC number assignments for the 1725 GUF were identified in BioCyc, TubercuList, UniProt, and KEGG and compared with the EC numbers assigned during manual curation (Materials and Methods). The presence of the EC number annotation was compared between the databases and manual annotation and depicted in Fig 2A. EC numbers were assigned manually to 111 GUF, 59 of which are genes newly characterized with EC numbers in this study (Fig 2A). EC numbers for the 1725 GUF were retrieved for the above-mentioned databases and compared across the databases. The 111 GUF were annotated with one or more of 98 unique EC number assignments with varying degrees of specificity (Fig 2B). Of the 98 EC numbers assigned to the GUF, 59 were unique to this study resulting in potential expansion of the enzymatic capabilities of *M. tuberculosis*. BioCyc was excluded from Fig 2A due to sparse annotation of the 'hypotheticome' GUF.

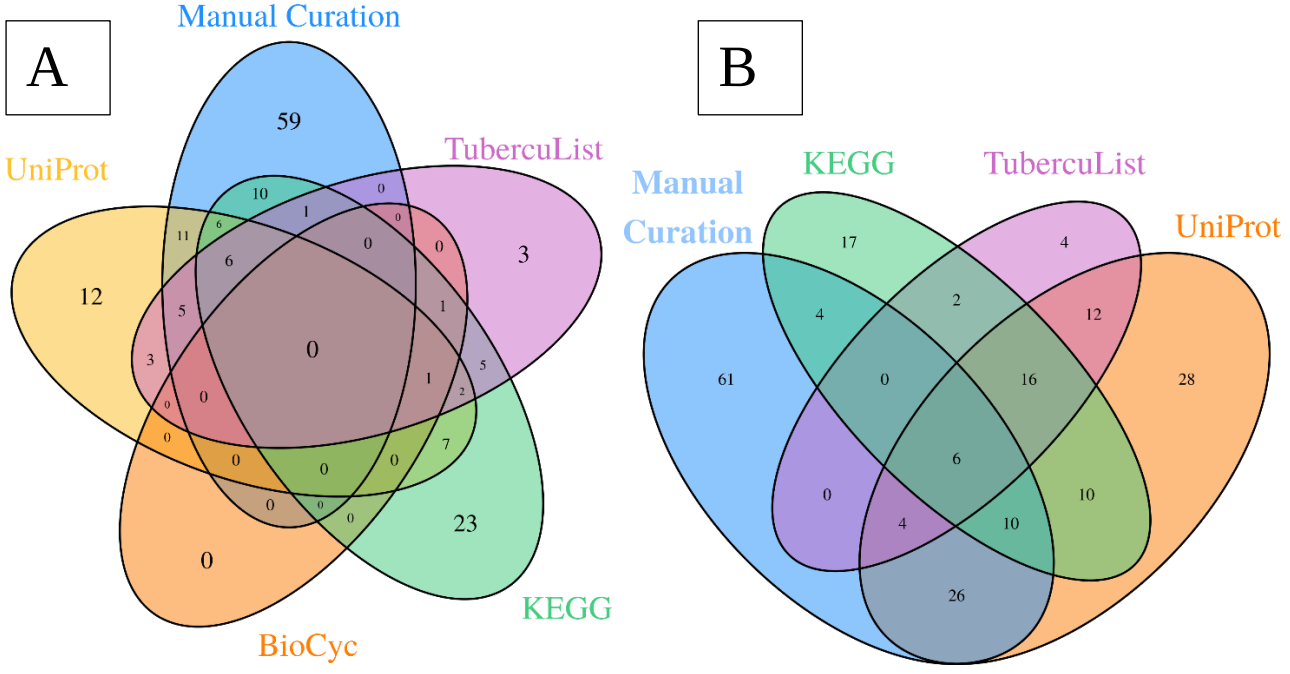


Fig 2. EC number annotation in the manual curation effort compared to widely used databases. (A) Mannotations were compared with those in the databases in Table 2 to identify the presence or absence of EC number in GUF annotations. The ovals in plot A represent the set of GUF annotated with an EC number in each of the five databases compared. The non-overlapping segments indicate the number of GUFs annotated uniquely in the that database. For example, 59 GUFs were annotated with an EC number in the manual curation effort in this study and these genes are not annotated with an EC number in TubercuList, UniProt, KEGG and BioCyc. (B) EC numbers were enumerated for the GUF for each of the database and unique EC numbers were identified and compared to the unique EC numbers manually curated. The ovals in plot B represent the set of unique EC numbers annotated for the GUF in each of the 4 databases compared. For example, the manual curation effort yielded 61 unique EC number assignments to the GUF indicating curation of previously unannotated function to these genes in this study.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene products inferred from structure despite low AA% similarity

Annotations in Table 5 demonstrate the utility of structure-based annotation for inferring function where other methods cannot. The table contains annotations inferred from high structural similarity to a PDB template despite lacking appreciable sequence similarity. Some of these affirm tentative or unreviewed annotations in UniProt or Mtb Systems Portal, others expand on them, and others project functions where there were previously none. Affirmatory annotations support the existing annotations and indicate their predictive strategies and structural inference converge while novel annotations mark exciting prospects for experimental validation.

Table 5. Novel annotations transferred through structural similarity despite low sequence similarity.

Locus	Top I-TASSER hit	AA%	TM_{ADJ}	PDB	Final annotation	TubercuList	UniProt	Mtb Net Portal	Type
Rv3507	Fatty acid synthase Subunit beta	41.5	0.87	2pff	Probable Fatty acid Synthase subunit	PE-PGRS family protein PE_PGRS53	PE-PGRS family Protein PE_PGRS53	None	Novel
Rv2998A	Osmolarity sensor Protein EnvZ	22.4	0.85	3zrw	putative phosphorelay sensor kinase	Conserved hypothetical Protein	Two-component system, OmpR family, Sensor kinase (unreviewed)	None	Affirmatory
Rv1139c	Integral membrane Methyltransferase	18.2	0.86	**4a2n**	putative integral Membrane Methyltransferase	Conserved hypothetical Membrane protein	Conserved hypothetical membrane protein (Membrane protein)	None	Novel
Rv1766	Copper-sensing transcriptional Repressor CsoR	29.2	0.84	4m1p	Putative Transcription factor	conserved protein	Conserved protein	None	Novel
Rv3192	5,10-methylenetetrahydro Methanopterin Reductase	15.6	0.83	1z69	Putative Monooxygenase	Conserved hypothetical alanine And Proline-rich protein	Conserved hypothetical Alanine and Proline-rich protein	Oxidoreductase	More specific
Rv2141	M20 family Metallo peptidase	19.5	0.82	2pok	putative linear amide hydrolase	conserved protein	conserved protein	FIG016551: Putative peptidase	Affirmatory
Rv0052	Isonitrile hydratase InhA	32.6	0.81	3noo	Putative Cyclohexyl-isocyanide hydratase/ putative peroxiredoxin	Conserved protein	Conserved protein	ThiJ/Pfpl Family protein	Novel

A subset of the novel annotations from I-TASSER models with high structural similarity to solved PDB crystal structures of known function. All annotations either extend those found in frequently updated databases or are new entirely. Sequence similarities among proteins displayed range from well-below (<20%) to modestly above 40% sequence similarity, which unofficially demarcates the "twilight zone" of sequence homology, beyond which only structure-based similarity methods can detect remote homology[46]. High TM_{ADJ} reflects that the true structure is similar to that of the PDB template. TM_{ADJ} above 0.52 indicates that the template and the GUF share the same structural fold, and higher scores indicate greater degrees of similarity in structure and, by extension, function. For each locus tag, annotations from UniProt, Mtb systems Portal, TubercuList, the highest structural similarity (after adjusting for expected error, "TM_{ADJ}"; Equation 1, Supplemental Note) PDB template and its identifier ("PDB"), and its final annotation (File S1) are displayed, along with the amino acid identity, and type of characterization. Those listed as "affirmatory" corroborate the annotations in UniProt or Mtb Network Portal. "Novel" are entirely novel annotations to those in UniProt and Mtb Network Portal, while "more specific" are in accord with annotations in other databases but describe product function in greater detail. S9 Table contains all PDB matches with TM_{ADJ} greater than 0.52.

Structural analogs of our GUF span diverse functional classes. We highlight analogs relevant to open questions in *M. tuberculosis* pathogenesis, along with some that are otherwise novel or surprising.

The structural analogy of Rv1139c to an integral membrane methyltransferase implies a potential role in virulence, as one-carbon transfers modify mycolic acids embedded in the *M. tuberculosis* cell wall.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mycolic acids feature prominently in the repertoire of pathogenic weaponry wielded by *M. tuberculosis* and are essential to virulence[47].

Rv1766 is a putative transcription factor that structurally resembles a copper-responsive metalloregulatory protein. Transcription factors of *M. tuberculosis* are considered mostly identified, with substantial efforts undertaken toward their systematic characterization, so if this gene truly encodes a transcription factor it would be significant[48,49]. Meanwhile, Rv0052 and Rv3192 play putative roles in the redox response, critical to enduring host attacks in macrophage[50]. These highlight a few of the many annotations afforded by inference through structural similarity.

Similarity to host-like structures suggest host-mimicry in many *M. tuberculosis* hypothetical proteins

Though our inclusion criteria balance stringency and quality, proteins with structural analogs (as opposed to homologs) may be overlooked, due to their low AA% (Materials and Methods). As an obligate pathogen, *M. tuberculosis* likely harbors proteins convergently evolved to mimic host protein structure[51,52]. To scan for such cases, we re-examined hits with TM_{ADJ} values that indicated matching topology (Supplemental Note, Equation 1) and low sequence similarity. Remarkably, several were analogous to host-proteins involved in TB infection dynamics, but not yet attributed to a protein *M. tuberculosis* (Table 6).

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 6. Protein models with high similarity to known players in immunity and the host-pathogen interface.

locus	PDB analog	TM_{ADJ}	mechanism	species	TubercuList
			alter cell fate		
Rv3304	Gamma-glutamylcyclotransferase	0.88	inhibit apoptosis Cascade	*Homo sapiens*	Conserved protein
			host transcriptome modification		
Rv1179c	ATP-dep Type ISP Restriction-modification enzyme	0.62	methylate host DNA	*Lactococcus lactis*	Unknown protein
			host organelle entry		
Rv0538	human secretory Immunoglobulin A	0.65	evade immune Detection	*Homo sapiens*	Possible conserved membrane protein
Rv3737	Importin subunit beta-3	0.64	translocate DNA & Proteins to nucleus	*Saccharomyces cerevisiae*	Probable conserved membrane protein
Rv0585c	Human Transportin 3	0.68	translocate DNA & Proteins to nucleus	*Homo sapiens*	Probable conserved integral Membrane protein
Rv1278	Human Transportin 3	0.73	translocate DNA & Proteins to nucleus	*Homo sapiens*	Hypothetical protein
			vesicular trafficking		
Rv2998A	Arf1 GTPase activator Sec7	0.68	control vesicle traffic	*Thielavia terrestris*	Conserved hypothetical protein
Rv0500B	Vesicle-associated Membrane protein 8	0.59	control vesicle contents	*Rattus norvegicus; Mus musculus*	Conserved hypothetical protein
Rv0500B	ARF GTPase-activating Protein GIT1	0.59	control vesicle traffic	*Rattus norvegicus*	Conserved hypothetical protein
Rv2082	Copi Coat vesicular Coatamer transport protein	0.59	Clandestine Vesicular transport	*Saccharomyces cerevisiae; Mus musculus*	Conserved hypothetical protein
Rv3425	RHO GTPase-activating Protein RGD1	0.53	control vesicle traffic	*Saccharomyces cerevisiae; Mus musculus*	PPE family protein PPE57
Rv3362c	Probable gliding protein mglA	0.52	control vesicle traffic	*Thermos thermophila*	Probable ATP/GTP-binding protein
			other		
Rv0193c	Niemann-Pick C1	0.63	acquire cholesterol; Expunge Ca2+	*Homo sapiens*	Hypothetical protein
Rv1191	Lysosomal protective protein	0.60	reduce proteolytic Stress	*Homo sapiens*	Conserved protein
Rv1276c	Ubiquitin-associated and SH3 Domain-containing protein B	0.66	Dsrupt host Ubiquitination pathways	*Mus musculus*	Conserved hypothetical protein
Rv0416	Ubiquitin-related modifier 1	0.69	Dsrupt host Ubiquitination pathways	*Mus musculus*	Possible protein ThiS

All matches in the table exceed TM_{ADJ} of 0.52 and mediate functions associated with host subversion in other pathogens, or are structurally similar to host proteins implicated in immunity, suggesting analogy. These functions were not necessarily included in the annotation .gff file (File S1) because their low AA% prevented them from meeting inclusion criteria, but represent interesting candidates for experimental testing.

These may clarify incompletely characterized elements of mycobacterial lifestyle in the host, particularly in immune-cell fate and transport into, and regulation of the cellular environment in which *M. tuberculosis* resides during infection. More broadly, the recurrence of these analogs suggests that pathogen host-mimicry for manipulating host cell phenotype may be a more widespread and tightly orchestrated phenomenon than currently appreciated.

Perhaps most fascinating of these putative analogs is Rv0193c, which encodes a protein structurally resembling Niemann Pick 1 protein (NPC-1) (Fig 3). NPC-1 mediates an inherited lipid storage disorder in human lysosomes[53]. The overlapping portion of the two structures span multiple regions of NPC-1 annotated on UniProt as dipping into the lumen, suggesting this potential mycobacterial analog may attach to the phagosomal lumen. A potential hypothesis is that within the lysosome, secreted NPC-1 analogs competitively antagonize native host proteins, initiating cholesterol accumulation. This protein's effect on the NPC-1 pathway may contribute to phagosomal maturation arrest, calcium dyshomeostasis,

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

altered mycolic acid production or influence host metabolism and transport of cholesterol and other lipids[54]. Potential roles of these proteins are expanded upon in the Discussion.

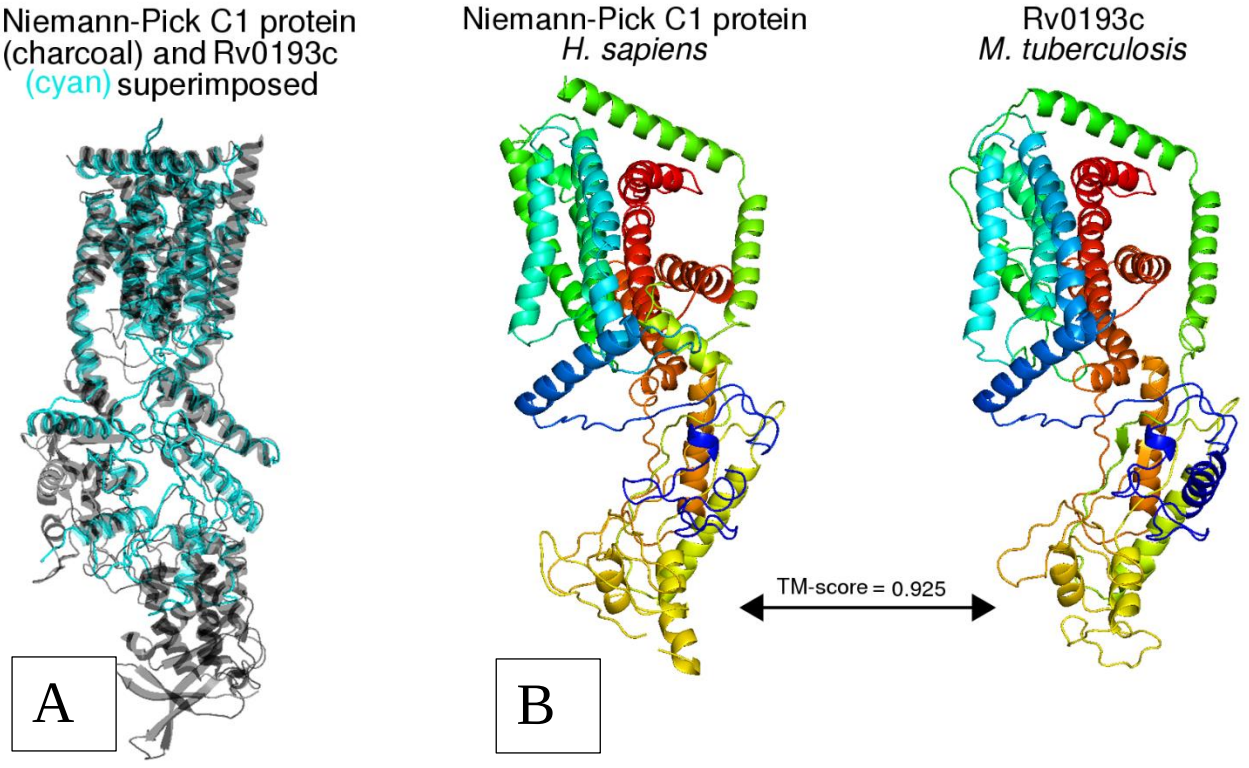


Fig 3. Structural similarity between Rv0193c modeled structure and Niemann-pick C1 protein (NPC1) from *homo sapiens*.

(A) Superimposition of the larger NPC-1 crystal structure (PDB Template 3jdB, charcoal) and the predicted model structure of Rv0193c (.pdb structure predicted from I-TASSER, cyan). Structures are rendered translucent to depict their relative positions. (B) The region of NPC-1 which Rv0193c aligns to (RMSD = 2.72). All structures were derived from PDB files and visualized in pymol.

PE/PPE protein structure models similar to diverse proteins

Candidate structural matches for 7 of 36 (19.4%) PE family proteins and 11 of 68 (16.2%) PPE family proteins were identified through structural alignment to solved structures from PDB. However, these proportions are well below the genome-wide proportion of proteins with matches by the same criteria (521/1725, 30.2%), likely due to the uniqueness of the PE/PPE families. Any clues to the function of these hardly characterized protein functions are valuable nonetheless, as they are unique to the MTBC complex and repeatedly implicated in pathogenesis. Effector proteins from other pathogens and human proteins were prevalent among those with similar PDB templates (Table 7), supporting the notion that PE/PPE proteins are integral to host-pathogen interaction in *M. tuberculosis*.

Table 7. Protein models of PE and PPE family genes with similar proteins

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene	Protein	homolog	organism	Homolog 2	organism	Homolog 3	organism
			PPE family				
Rv0388c*	PPE9	PPE family protein PPE41	*M. tuberculosis*				
Rv0442c	PPE10	**Putative flagellar Hook-associated protein**	*B. Pseudomallei*				
Rv0755c	PPE12	Rubber oxygenase A	*Xanthomonas Sp*				
Rv1548c	PPE21	Pentamodular Arabinoxylanase	*C. Thermocellum*				
Rv2430c*	PPE41	PPE family protein PPE41	*M. tuberculosis*				
Rv3135*	PPE50	PPE family protein PPE41	*M. tuberculosis*	**small G-protein—cytoplasmic effector adapter protein**	*H. sapiens*	Methyl-accepting Chemotaxis protein	
Rv3159c	PPE53	**Serine protease Pet autotransporter**	*E. Coli*	**Serine protease EspP**	*E. Coli*	**Hemoglobin-binding protease Hbp autotransporter**	*E. Coli*
Rv3425*	PPE57	PPE family protein PPE41	*M. tuberculosis*	**cytokinesis protein**	*S. serevisae*	Haptoglobin-hemoglobin Receptor	*T. congolense*
Rv3533c	PPE62	**Immunoglobulin A1 Secretory component**	*Homo sapiens*				
Rv3558	PPE64	**Serine protease Pet autotransporter**	*E. Coli*	**Serine protease EspP**	*E. Coli*		
Rv3739c*	PPE67	PPE family protein PPE41	*M. tuberculosis*				
			PE family				
Rv3893c*	PE36	PE family protein PE25	*M. tuberculosis*				
Rv3477	PE31	**toxin translocase subunit**	*P. luminescens*				
Rv1791	PE19	**Colicin**	*E. Coli*	Methyl-accepting Chemotaxis protein I	*E. Coli*	**F-BAR domain Only protein 2**	*H. sapiens*
Rv1788	PE18	**Colicin**	*E. Coli*	Methyl-accepting Chemotaxis protein I	*E. Coli*	**F-BAR domain Only protein 2**	*H. sapiens*
Rv1430	PE16	**Putative flagellar Hook-associated protein**	*B. Pseudomallei*				
Rv1386	PE15	**Colicin**	*E. Coli*				
Rv0916c	PE7	Methyl-accepting Chemotaxis protein I	*E. Coli*	**Cytoskeleton Remodelling protein**	*D. discoideum*		
Rv0160c	PE4	Dinuclear copper Monoxygenase (Tyrosinase)	*A. oryzae*				

*Homologous to one of the members of the PE25/PPE41 heterodimer with solved structure shown to mediate host-pathogen interaction

Potential analogs to host proteins, which may mimic or act to subvert the host are bolded (both the template and the locus tag). Only models with at least one match exceeding TM_{ADJ} of 0.52 (the TM-score corresponding to matching topologies > 50% of the time)[25] are shown. Proteins and the organism expressing them are listed. Where multiple matches exceeded the threshold, only the top three hits are displayed.

Though lacking obvious relevance to the host-environment, the analog of Rv0755 is interesting: Rubber oxygenase (RoxA) from plant pathogen *Xanthomonas*, which aligns across 97% of the polypeptide. Latex degradation would be an unexpected capability of *M. tuberculosis* without known utility in the lung, to our knowledge. Alternatively, this protein may catalyze degradation of some form(s) of isoprenes, a group of latex-related compounds which mediate important processes in virulence and cell wall metabolism[55].

The strong similarity between PE_PGRS family Rv3507, and a fatty acid synthase (FAS) subunit is intriguing, as functions of the PE_PGRS protein family are largely unknown. On the one hand, this annotation should be taken with skepticism; PE_PGRS protein models closely resembled FAS subunit protein structures with conspicuous frequency, particularly *Saccharomyces cerevisiae* PDB template 2pff. Perhaps this similarity is merely an artifact of their high GC-content and the resulting glycine abundance aligning to the hydrophobic region of large eukaryotic synthases, inflating their predicted similarity score. On the other hand, Rv3507 and a handful of others exhibit similarity in structure, but not sequence, to additional eukaryotic FAS enzymes, suggesting a true relationship, though this could indirectly result from threading guided by 2pff during I-TASSER modelling[19].

CATH structural topologies

In parallel with functional annotations, structure can be annotated systematically through transfer of CATH "topologies", often referred to as protein "folds" from structurally similar PDB entries[56].Protein

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

structure models for 373 GUF had matches meeting the CATH inclusion threshold with PDB entries annotated with CATH topologies, 39 of which had no gene product annotation (S10 Table).

Some CATH topologies encompass multiple protein superfamilies that carry out diverse reactions, while others are nearly invariant in function across all members of the topology[57]. Both cases are represented in commonly transferred CATH annotations (Table 8). For example, members of the "TIM Barrel" topology (Table 8) vary wildly in function[58,59], and "alpha-beta plaits" comprise over 90 superfamilies and thousands of domains[39,60]. In contrast, "Tetracycline repressor; domain 2" (TetR) topology members vary in sequence, are structurally homogeneous[61], and function nearly exclusively as concentration-dependent transcriptional activators. Transcriptional repressors dissociate from their target DNA in the presence of substrate, activating transcription, which, in some cases, are antibiotics, or other compounds of clinical interest[62]. TetR repressors serve important functions in *M. tuberculosis*. They control expression of the primary proteasomal complex responsible for degrading damaged proteins[62]; of isocitrate lyase, a critical metabolic switch that initiates the glyoxylate shunt to reprioritize nutrient utilization[63]; and of drug efflux pumps in response to antibiotic concentration[61]. All GUF with TetR CATH topologies are listed S11 Table, along with the PDB templates from which they were transferred.

Table 8. Most commonly transferred CATH topologies and their associated functions

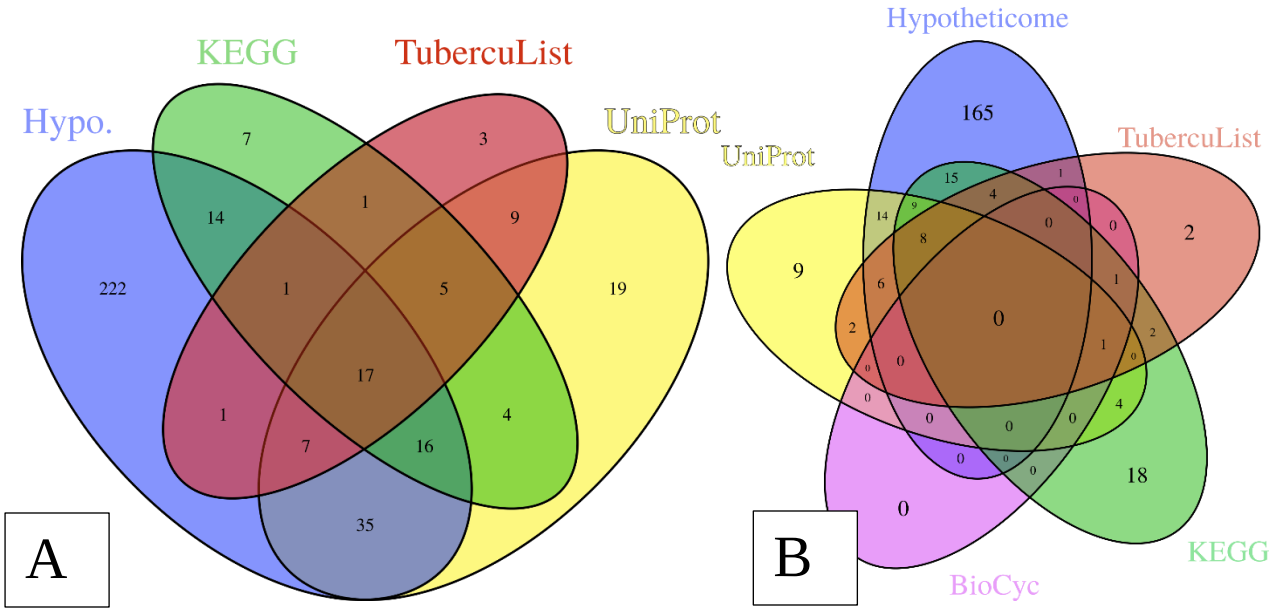
Associated function(s)	CATH topology	Genes
nucleotide/cofactor binding	Rossmann fold	82
Transcriptional regulation	Arc Repressor Mutant, subunit A	37
Many, diverse	TIM Barrel	29
Many, diverse	Alpha-Beta Plaits	20
membrane structure & processes	Helix Hairpins	20
Transcriptional repression	Tetracycline Repressor; domain 2	17

Expanded metabolic capabilities of *M. tuberculosis*

EC numbers describe catalytic function hierarchically, through a four-tiered numerical identifier, and allow systematic description and classification of the enzymatic capabilities. Levels of this hierarchy funnel from general enzyme class (e.g. ligase, oxidoreductase) down to substrate specificity with atomic precision[29].

Cumulatively, annotations yielded 222 unique EC number assignments spread across 313 GUF. Of this unique set of EC numbers, 165 are specific to this *M. tuberculosis* annotation, and 222 (not to be confused with the 222 unique EC numbers) of the 313 genes are annotated with an EC number unique to this annotation (Fig 4).

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

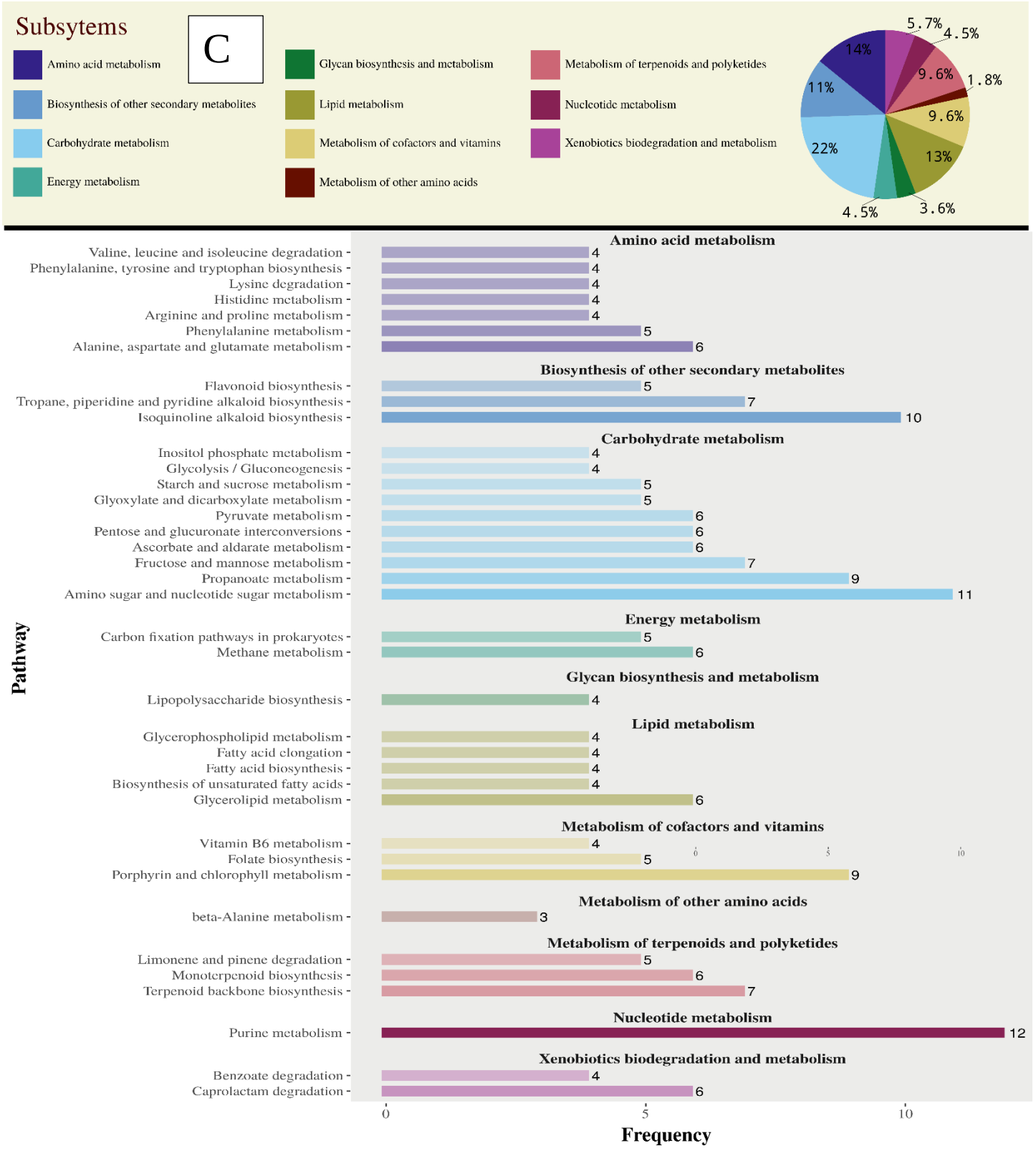


Fig 4. Updated *M. tuberculosis* EC number annotation compared to widely used databases.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

EC number annotations assigned for manually curated gene products as well as the assignment using structural homology for the 1725 GUF were compared with EC number annotation in existing databases. The annotation effort yielded one or more EC number curation for 313 genes assigned manually or using structural homology prediction in the 1725 GUF. (A) Existing databases were compared to identify the presence of EC number annotations for the GUF. (B) EC numbers were enumerated for the GUF for each of the database and unique EC numbers were identified and compared to the unique EC numbers annotated in this study. (C) Distribution of EC numbers annotated across KEGG Subsystems and Pathways. Subsystems depicted are the generic KEGG subsystems, since they have yet to be assimilated into the metabolism of *M. tuberculosis*. Pathways within each subsystem are also depicted and the number of EC numbers mapping to them are displayed or all pathways with at least four genes, or the highest total within the subsystem if no pathways had four or more genes. The pie chart shows the proportion each subsystem comprised of the total added.

These expand the annotated metabolic capabilities of *M. tuberculosis* across numerous pathways and subsystems (Fig 4C), several of which are underrepresented, scarcely characterized, or integral to pathogenesis. For instance, the novel candidate proteins mediating reactions involved in lipid metabolism may address open questions in how the elaborate lipidomic profile of *M. tuberculosis* is orchestrated, and how it contributes to pathogenesis.

GO terms implicate proteins in physiological contexts and processes of the host environment

GO terms describe gene products through three structured ontologies: biological processes (the processes in which the product plays a part), cellular components (its location within the cell), and molecular functions (its specific function)[28,64].

The distribution of GO terms transferred to GUF because of structural similarity indicates common themes in the new annotations, with several particularly relevant to pathogenic processes of *M. tuberculosis* (Fig 5).

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

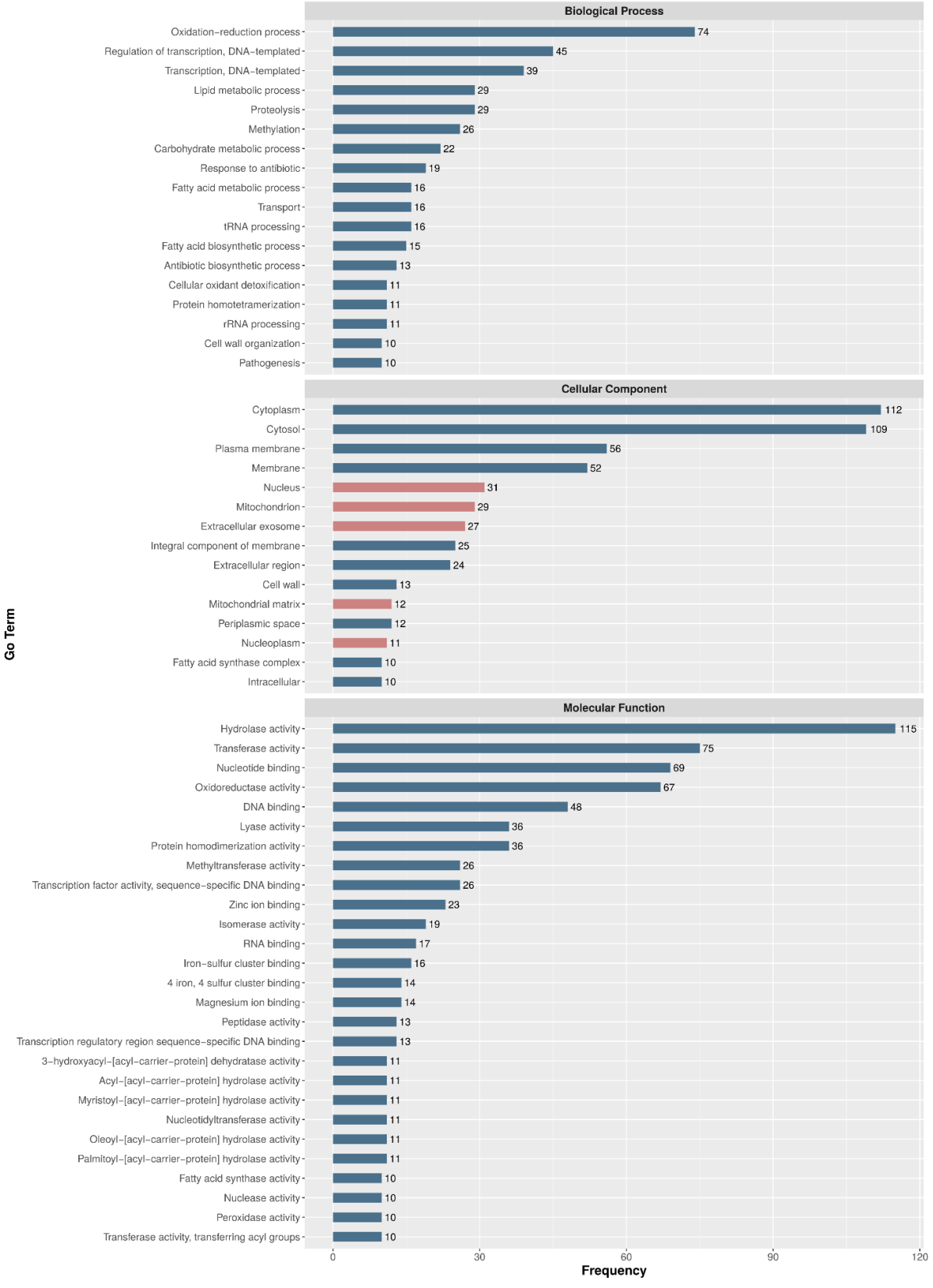


Fig 5. Most frequently annotated GO terms.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene Ontology (GO) terms describe proteins by the cellular location where they perform their function (Cellular Component), their precise function (Molecular Function), and the processes/ pathways in which they are involved (Biological Processes). GO terms were incorporated as described in the flow diagram of Fig 8B. Terms from all three ontologies were included, and very general terms were culled (e.g. "catalytic activity", "growth", "metal ion-binding", etc.) from the plot, but remain in the .gff (File S1). Counts in the histogram span all 1,725 GUF. Only those terms with ten or more occurrences are plotted, and no attempt was made to collapse child ontologies into parents. GO terms implicating function specific to eukaryotic environment are shown in red.

Top hits in the biological process ontology align with classic *M. tuberculosis* features: Myriad proteins to overcome oxidoreductive stress in the host environment, an elaborate array of lipid metabolic capabilities, and sophisticated transcriptional regulatory programs to cater phenotype to what best facilitates survival in any one of the many environments it must persevere in the host[65].

Several cellular component GO terms reference eukaryote-specific organelles, such as "nucleus" and "mitochondrion", indicating candidate host-interaction proteins. Further implicating function at the host-pathogen interface are the 27 GUF annotated with "extracellular exosome" (Fig 5). Exosomes provide a means of intercellular communication between species, and have recently gained recognition for their role is host-pathogen interaction[66], particularly as an access point to the immune system for pathogen-derived effectors[67].

Potential novel efflux pumps uncovered through structural similarity

Apparent structural analogs of known transporters emerged from our 1,725 GUF; 60 met our annotation transfer criteria (S12 Table). These comprised diverse transporter classes, but sodium/hydrogen antiporters (Na+/H+ antiporters) were particularly abundant. (Na+/H+ antiporters) are implicated in proton import necessary to increase proton motive force (PMF) which, in turn, fuels the electron transport chain, generating ATP for active drug efflux posited to mediate intrinsic antibiotic resistance in *M. tuberculosis*[68]. Sixteen GUF shared structural similarity with drug transporters, the majority exhibiting similarity to PDB template 4zow, a Major Facilitator Superfamily family PMF antiporter isolated from *E. coli*[69], totaling 24 likely analogs of PMF-driven antiporters among the GUF, a sizable addition to gene products implicated in drug efflux.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

We isolated 1,725 genes products (GUF) in need of updated annotation on TubercuList, the most frequently cited database for *M. tuberculosis*[6]. Through manual literature curation, inference by structural similarity, and consolidation of existing databases, we annotated 987 of these GUF with functional notes, gene products, or CATH topologies, 713 of which are gene products, a vast improvement over other *M. tuberculosis* annotation resources Major implications for *M. tuberculosis* physiology and infection dynamics stem from these annotations. These implications span core metabolism, transport, pathogenesis, host-interaction, regulation, stress responses, DNA damage repair, drug resistance, and more. We highlight several of these to demonstrate how the new information from this article illuminates potential phenotypic inferences, and increases such studies' value. We present these annotations in both human (S1 Table) and machine readable (S1 File) format to facilitate genome-wide bioinformatics analyses, and targeted reference of particular genes for applications of narrower focus. This work advances the *M. tuberculosis* reference annotation and provides multiple resources to integrate into future analyses and exploration of those still unannotated.

In the following sections we discuss the importance of manually curated annotations generally, and for *M. tuberculosis* particularly. We then discuss the novel annotations feasible through structure-based methods, focusing on PE/PPE genes and on annotations with mearing on the host-pathogen interface. We next discuss the implications of our annotation update to *M. tuberculosis* metabolism, newly identified transport and efflux proteins, and how our annotation enriches an example transcriptomics study. Next, we discuss how our inclusion criteria were informed, and how to interpret each of the types of annotation included in this effort, and the limitations of our annotation. We then discuss the genes that remain uncharacterized, before exploring the potential application of these methods to other genomes.

Regular Manual Curation is essential for well-studied organisms

Manual curation is the most accurate form of extracting annotations from literature, and a critical component of annotation. While automation can help curators prioritize which papers to read in detail, current software tools developed to "read" scientific articles and extract information cannot reliably gauge strength of determination methods, leaving manual curation as the gold standard[45]. The curators of UniProt estimate they evaluated over 4,500 papers, but only used 1,368 of these in their annotation[45]. Our manual annotation cited 656 papers, but required reading of thousands, underscoring the magnitude of the task of manual curation. UniProt curators address this problem by 1) regularly reading through published articles (every 6 months) so literature review remains feasible 2) automated filtering of articles to determine which are worth curating manually for functional annotation[45]. We recommend these approaches to anyone desiring specific functional annotation for an organism of interest. The subjectivity of these calls should be minimized to keep a high standard of annotation accuracy, though attaining absolute objectivity is challenging. A key step in maximizing objectivity is following controlled vocabularies known as ontologies, several of which we implemented in our annotation procedures. Ontologies for several facets of annotation exist, including evidence assignment, function, and relationships between gene products[28,70]. The volume of new publications is simply too large for global curation efforts like UniProt to manually annotate all experimental characterizations. Research communities of widely studied organisms must initiate their own, complementary manual curation efforts to remain updated. To mitigate the manual component of this process, PubMed Central (PMC) can be accessed programmatically, and queried by a date range of publication dates, organism, and gene, and is a good place to start for those desiring the most up-to-date functional annotation for a particular organism.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Significance of Manually Curated Gene Products

Mannotation provided important functional assignments, absent from regularly updated databases, to 85 GUFs (29 excluding antigens) with high certainty. Particularly striking were annotations of PE and PE_PGRS genes with experimentally characterized functions (Table 4). These gene families are unique to mycobacteria, sparsely characterized, and difficult to accurately sequence due to highly repetitive regions and high GC-content[71]. Notably, Rv1430 (PE 16) encodes an esterase[72]. This catalytic capacity by a PE/PPE family gene is a crucial discovery that should not be obscured by the current disconnect between the *M. tuberculosis* bibliome and annotation databases.

Highlighting more examples of important annotations absent from frequented databases demonstrates the depth of this issue: We mannotated Rv0024 as a probable peptidoglycan hydrolase, and noted its involvement in isoniazid and pyrazinamide (two first line antituberculosis drugs) resistance and biofilm formation[73]. This gene had been annotated as a putative secreted protein, which does not indicate its involvement in drug resistance and biofilm formation. Also among our unique mannotations is Rv1337 as a rhomboid protease[74], a far more informative annotation than that from TubercuList (possible integral membrane protein), PATRIC, RefSeq, or Mtb Network Portal (S7 Table). Also absent from these databases is the involvement of Rv1337 in ciprofloxacin and novobiocin resistance and in biofilm formation[74], which we added as functional notes. Another mannotated gene, Rv3005c, encodes an oxidoreductase responsive to oxidative stress[75]. This stress response is important for in-host survival of *M. tuberculosis*[76,77], but lacks annotation implicating this gene in common databases. Many other functional features and discoveries pertinent to pathogenesis, host-pathogen reaction, and antibiotic resistance are present in functional notes (File S1). We annotated 599 GUF with such notes, providing information on many lacking product annotation (S5 Table).

Knowing these genes' precise functions provides researchers greater context of the physiological capabilities when piecing together the mechanisms of resistance to drugs and the genomics underpinning their emergence. A more informed research community can speed the drug discovery process by clarifying fertile routes to pursue and by pruning dead-ends in these efforts.

Structural analogs reveal candidate host-interaction proteins

Identifying proteins at the host-pathogen interface is challenging. Technical limitations in faithfully representing the host environment hinder laboratory approaches, while sequence homology is ineffective, as these proteins often evolve convergently with low AA similarity rather than from common ancestry[52].

Annotation derived from structural similarity is data-driven, removing both the bias toward *a priori* assumptions and the constraint to evolutionarily close relatives inherent in approaches based on AA sequence-similarity. Therefore, structural similarity-based approaches can furnish annotations that run contrary to conventional predictions, while imparting greater significance to findings aligning with dogma, since they are found independent of prior assumption. These advantages manifest repeatedly in this annotation: I-TASSER runs revealed dozens of GUF models structurally analogous to crystallized proteins with functions specific to eukaryotic cellular environments, unmasking potential host-interacting proteins. Many of these putative analogs fell in protein classes known to subvert host immunity in other intracellular pathogens, while others matched novel targets that are implicated in processes that contribute to *M. tuberculosis* pathogenesis, but lack known pathogenic effectors (Table 6).

The mechanisms of these putative pathogenic effectors can be divided into four broad categories: (1) Influencing vesicle trafficking to various organelles, (2) Manipulating host transcription, (3) Manipulating immune cell fate, and (4) Localizing pathogen-derived cargoes across membranes into host

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

organelles. Analysis of putative structural analogs also added pathogenic functions to those already known for the Rv3365c-Rv3361c operon, and revealed an unexpected pattern of *M. tuberculosis* proteins analogous to human proteins implicated in lysosomal storage diseases.

Vesicle trafficking.

One of two primary routes to lysosomal degradation of *M. tuberculosis* is mediated by small GTPases. Small GTPases influence vesicular transport to orchestrate *M. tuberculosis* delivery to the lysosome[78]. Interestingly, Rv3362 matched six Ras GTPase proteins as structural analogs (S9 Table). Ras GTPases interact with host effectors to regulate membrane trafficking and vesicle transport to coordinate immune responses and influence how immune cells mature into different states (maturation into phagosome, for instance)[79]. Some known *M. tuberculosis* effectors act on Ras proteins to increase *M. tuberculosis* replication[78,80].

Rv3362 may interfere with native eukaryotic Ras signaling, uncoupling or modifying host signals to serve interests of the pathogen. In addition, Rv3362 may contribute to the observed subversion of autophagy by *M. tuberculosis*, or perhaps has distinct, complementary roles in host subversion.

Another surprising apparent host analog is Rv2082, the structural model of which resembles a "coatomer" protein. Coatomer proteins drive vesicular formation and subcellular trafficking from the Golgi, and are repurposed similarly in phagosomes[81]. This leaves one to wonder whether *M. tuberculosis* is mimicking the coatomer to disrupt proper function of the phagosomal transport machinery, or if this is a 'Trojan horse' of sorts; a means for entry and delivery of nefarious signals to host nuclei for manipulation of programmed immune responses.

Subverting host immune defense programs by manipulating transcription.

Direct manipulation of host transcription has recently been characterized as a route for pathogenic control of host processes. At least two known proteins of *M. tuberculosis* exert epigenetic control on host transcription: a histone methyltransferase to modify higher order chromatin organization, and a DNA methyltransferase to modify host DNA directly, but others may have similar roles[82].

One such candidate arose in the hypotheticome GUF: Rv1179c structurally resembles a type ISP restriction-modification enzyme of *Lactococcus lactis*, which comprises DNA endonuclease and DNA methyltransferase activity. It is the only structural match for Rv1179c and overlaps with the C and N-terminal helicase as well as the N6-adenine DNA methylase domains annotated on InterPro. DNA methylation in *M. tuberculosis* has been reported but accounted for by three other enzymes[83,84], suggesting that if truly a DNA methyltransferase, Rv1179 may be host-directed. Alternatively, it could be a self-directed RM-system that is yet undiscovered.

Driving immune cell fate for replication-permissive niches *in vivo*.

Another potential effector modulating host cell fate is Rv3304, which structurally resembles human gamma-glutamyl transferase (Table 6), which upregulates cytochrome C from the mitochondria, a central player in apoptosis induction[85,86]. This potential mimicry may allow *M. tuberculosis* to manipulate host immune cells toward or away from apoptotic cell death. Modeled Rv2998A bears a structure analogous to a sec7 guanine exchange factor (GEF), which can affect host GTPases to influence directionality of vesicular transport. This phenomenon is one of few well-characterized in host mimicry, which has recently gained appreciation as a widespread means of pathogen adaptation[52,81]. Pathogens can mimic their host through horizontal acquisition or convergent evolution. Host mimicry of

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sec7 has been characterized in fellow obligate pathogens *Legionella pneumophila* and *Rickettsia prowazekii*, as an effector molecule translocated into host cytosol via their type IV secretion system[87,88].

Localizing pathogen products to host organelles.

The discovery of eukaryotic nuclear import analogs provides candidates for mediating delivery of pathogen-created elements to the host nucleus for transcriptional manipulation. The Rv3737 model structurally resembles a cell-cycle dependent nuclear import subunit from *S. cerevisiae* (Table 6), with slightly lesser similarity to human importins (S9 Table). Similarly, models of Rv1278 and Rv0585c resemble human Transportin 3, a nuclear import factor that localizes extranuclear cargoes to the nucleus, particularly splicing factors[89]. Intriguingly, Transportin-3 is also implicated in HIV-1 replication, presenting a theoretical mechanism for potentiating HIV-1 and TB coinfection beyond their known synergies[90].

The Rv3361c-Rv3365c operon is heavily implicated in pathogenic processes.

Though known as an important virulence operon, several structural analogies suggest multiple pathogenic mechanisms are encoded by the Rv3361c-Rv3365c operon, which is starkly upregulated upon macrophage entry[91]. Rv3361c confers resistance to quinolones, ciprofloxacin, and sparfloxacin[92], while its structure also closely resembles a patatin autotransporter, a virulence factor contained in the whooping cough vaccine, which provides immunity to *Bordetella pertussis*, its etiological agent[93]. Rv3362c is a structural analog of the small GTPases that orchestrate vesicular transport. Together with previously characterized effectors Rv3364c (a serine protease inhibitor that binds Cathepsin G to halt a pro-apoptotic signaling cascade[91]), this operon comprises a formidable cluster of virulence factors.

Potential analogs of proteins implicated in lysosome storage diseases.

Perhaps most unique among these potential effector proteins is Rv0193c (Table 6), which appears analogous to human protein "Niemann-Pick protein 1" (NPC-1). Dysfunctional NPC-1 causes a eukaryotic disease characterized by lipid accumulation and defective lipid transport in lysosome[53]. Incredibly, *M. tuberculosis*-infected macrophage phenotypically resemble cells afflicted with Niemann-Pick disease in their accumulation of cholesterol and other lipids, and decreased calcium concentration. This phenotypic parallel Niemann-Pick diseased cells and *M. tuberculosis* infected macrophages between prompted investigation earlier this year, which heralded inhibition of the Niemann-Pick pathway as a mechanistic explanation for intracellular persistence, but did not identify a corresponding genetic determinant in mycobacteria[54]. Rv0193c may be a principal element of *M. tuberculosis* responsible for their findings.

Like NPC-1 and Rv0193c, Rv1191c structurally resembles a human protein that, when dysfunctional, characterizes a lysosomal storage disease: cathepsin A. Cathepsins are proteolytic, and degrade pathogenic proteins in the lysosome. *M. tuberculosis* appears to avoid this fate by knocking down cathepsin transcription and activity in macrophage, which improves survival. The authors noted that cystatins—natural cathepsin inhibitors—were upregulated early in the course of infection in activated macrophage, implying *M. tuberculosis* acted early to curtail cathepsin production[94]. An intriguing hypothesis is that through its similarity to cathepsins, Rv1191c signals a cystatin-cathepsin negative feedback loop without expression of host cathepsin, but rather with the *M. tuberculosis* analog (presumably inactive). An alternative hypothesis is that the *M. tuberculosis* cathepsin analog competes for limited cofactors or binding partners, perhaps the host cathepsin precursor for the (yet unknown) activator it requires in the lysosome[95], antagonizing formation of active cathepsins, thereby reducing its proteolytic activity and enhancing mycobacterial survival.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The convergence of structural analogs of *M. tuberculosis* proteins associated with lysosomal storage disorders paints a picture where, through structural mimicry of host proteins, *M. tuberculosis* induces a lysosomal phenotype resembling multiple diseased states. These present exciting opportunities for future work and foreshadow a larger set of effector proteins tailored to subvert the host immune system than currently characterized. A larger set would be unsurprising; many obligate pathogens have dozens, or even hundreds (*Legionella*, for example[87]) of secreted effectors that reconfigure the wiring of host immunity programs[85].

These candidate pathogenic effectors compel further investigation. If legitimate, they are appealing drug targets; their inhibition could potentiate the mechanisms of immunity they normally disrupt. Host-interaction is widely acknowledged to mediate fundamental *M. tuberculosis* survival strategies *in vivo*, but most such mechanisms are characterized scarcely, at best[96]. These candidate effector proteins are high priority candidates for experimental elucidation.

Structural analogs among PE/PPE proteins

Several trends among the elusive PE/PPE genes emerged based on the proteins they shared structural similarity to (Table 7). While PE_PGRS proteins mainly matched structures of eukaryotic fatty acid synthases, there emerged a more diverse array of structural matches among the other two PE/PPE families.

Five from the PPE family had significant homology to PPE41, one of the better understood PPE proteins. Among these, the structure model of PPE57, (encoded by Rv3425) resembles both a Ras GTPase and PPE41, which, in its heterodimeric form (with PE25) preferentially induces necrosis over apoptosis in murine macrophage, though the precise mechanism is unclear[97]. This concomitant similarity with PPE41 and Ras GTPases suggest PPE57 may encode a protein that carries out a similar function, and perhaps provides the unknown mechanism of necrosis inhibition carried out by the PE25/PPE41 heterodimer, since it shows homology to both PPE41 and to a reasonable cause of immune cell fate manipulation (through interfering with Ras signaling). The homology with PPE41 of four other PPE proteins suggests they may also be secreted effectors, perhaps after dimerizing with a PE gene, or, like PPE57, after being expressed on the mycobacterial membrane surface.

Several matches to PPE structure models suggest function in the host-environment. PPE64 and PPE53 structurally resemble serine protease autotransporters, which are secreted effectors widely implicated in bacterial pathogenesis. They play a variety of immunomodulatory and cytotoxic roles, such as degrading host proteins[98].

Interestingly, three PE proteins' top structural matches were colicins. Colicins are often secreted and translocated into neighboring bacteria, where they degrade nucleic acids to reduce competition[99]. *M. tuberculosis* may utilize these compounds to outcompete its non-pathogenic counterparts of the healthy lung microbiome. Alternatively, these matches could be host-directed, and damage host DNA and RNA through a similar mechanism.

Novel metabolic annotations fill knowledge gaps and may improve metabolic modeling

Newly annotated metabolic products fill pathway gaps and further characterize genome-wide metabolic reconstructions *in silico* for systems inquiry. Functional annotation of genes and proteins involved in carbohydrate metabolism point to novel substrate utilization capabilities in *M. tuberculosis*, and perhaps identify alternative in known metabolic pathways and help explain the remarkable resilience of *M. tuberculosis*[100].

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Polyketide and terpene metabolism mediate pathogenic processes unique to *M. tuberculosis* and are overrepresented in (Fig 9C). Polyketides are implicated heavily in virulence[101–103], make up part of the outer lipid layer of the *M. tuberculosis* bacillus, the host-pathogen interface[104], where they mediate processes to subvert mechanisms of host immunity[105]. Meanwhile, terpenes play an immunomodulatory role during early stages of *M. tuberculosis* infection and phagosomal maturation[106–108]. Recently, terpene products unique to *M. tuberculosis* expressed on the cell membrane surface were discovered[109] and others show potential as agonists of antibiotics for TB treatment[110], underscoring the importance of understanding their metabolism. Identifying products that metabolize virulence and antibiotic resistance components should accelerate advancements in understanding and treating *M. tuberculosis*.

Expanded set of putative transport and drug efflux proteins identified

Particularly relevant to *M. tuberculosis* research and clinical communities, are 60 GUFs implicated in membrane transport and efflux (see S12 Table for the complete list). These proteins are prime targets for inference through structural homology: due to structural preservation in their membrane-spanning regions, transport proteins exhibit a far lower degree of sequence conservation than globular proteins[42,111] relative to structure. This structural conservation likely owes to restricted structural freedom while embedded in the membrane. These putative transporters are candidate contributors to intrinsic antibiotic resistance and tolerance, which is often enacted through upregulation of membrane efflux in *M. tuberculosis*[112]. Others may serve homeostatic roles, such as regulating pH, the proton-motive force, or the relative concentration of metal ions. These are each tightly regulated to maintain a viable internal state, *in vivo*[113,114] . These GUFs are important to characterize experimentally, since efflux proteins are integral to understanding short-term drug response and baseline homeostasis. Their characterization would enhance the utility of *in silico* systems modeling approaches, and contribute to our basic understanding of *M. tuberculosis* homeostasis.

Greater annotation coverage in omics studies

To illustrate this immediate relevance to the *M. tuberculosis* research community, we chose two recent, genome-wide transcriptomics screens under novel conditions—iron deprivation[115] and exposure to the soluble fraction of activated lysosome[116]—to see if genes they reported as hypothetical are characterized in our annotation. We found that our annotations provided products for 16/31 and 18/37 genes listed as unknown or hypothetical among the differentially transcribed genes unique to the iron and lysosomal studies, respectively (S13 Table). This amounts to 50% of genes unannotated by their sources, similar to the fraction annotated in this work (45%). If available to the authors, these annotations could have affected the conclusions drawn these studies, or enabled more informed discussion.

Interpreting computationally derived annotations

To complement our manual literature curation, we further annotated the GUF through inferential methods. Each of these annotation sources have their own set of advantages, assumptions, limitations, discussed below.

Metabolic annotation: enzyme commission number.

The binomial regression in Fig 8 does not definitively predict precision at a given value of TMID, nor was it designed to. Rather, it provides a clear interpretation of how terms were derived, and an approximate, conservative projection of precision. This projected precision likely underestimates true precision: False negatives occur when the query protein EC is predicted and either 1) The EC of the PDB template is incorrectly annotated 2) The PDB template is incompletely annotated (the reaction described by the EC has multiple valid EC assignments) 3) The protein of the PDB template has multiple true ECs, of which only a subset are annotated 4) Our EC assignment was incomplete or incorrect for reasons 1-3.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In contrast, false positives are vanishingly rare; either the query or PDB EC assignment would have to be incorrect on a particular EC number of over 7,000[29].

Annotations were incorporated hierarchically using boundaries of 50% and 75% projected precision for the 3rd and 4th EC tiers (Fig 8C). The 3rd level of EC numbers describe detailed catalytic function (e.g., 3.1.6.- describe "Sulfuric ester hydrolases"), only lacking their specific preferred substrate(s).

CATH topologies.

Topology annotations derived through CATH had limited AA% in most instances (S2 Fig) which makes their incorporation valuable where genes would have no annotation otherwise[117]. Closer inspection of topologies with little functional diversity could allow valid inference that the protein of interest shares the function common to the topology, but we did not attempt to make such distinctions in this work.

Gene Ontology Terms.

The GO framework is unique in that it describes gene products in a species-independent manner and at varying degrees of specificity[70,118]. These features make GO terms useful for relating gene products across databases and drawing parallels between products from different species share function not apparent in their primary names. For example, "multidrug resistance protein" and "neurexin 1" sound unrelated, but can be unified by the GO term "transmembrane transport". This cross-species unification is particularly useful for reconciling annotation transfers of analogs and distant homologs into gene product names relevant to the organism of interest.

GO terms convey meaning at multiple functional levels and should be interpreted differently depending on their ontology and specificity. Molecular function ontology terms are likely the most reliable; they typically convey information that depends less on the genetic background of the organism than the other ontologies and are thus more likely conserved across similar structures in different organisms.

We implemented inclusion criteria for GO terms mirroring EC number cutoffs because they ultimately convey similar information: the propensity of structural similarity to PDB templates to imply shared function. GO terms encode diverse biological information and are frequently used to identify enriched processes and functions in a set of genes, such as those with upregulated expression under a particular condition, or those found mutated across clinical isolates during treatment of a particular drug. These GO terms increase the annotation coverage of *M. tuberculosis* H37Rv reference genome, and may help to uncover latent commonalities among gene sets, making them a useful component of this annotation update.

Ligand-binding sites.

Ligand-binding sites (LBS) were included from COFACTOR at C-score$_{LBS}$ values corresponding to precision > 0.6 in a recent benchmarking study[27]. LBS have narrower use than EC or GO terms, but can identify putative ligand-protein interactions, and elemental requirements. These binding site predictions and the residues predicted to coordinate binding are contained in S14 Table. These can be interpreted as being at least 60% likely to be true[27], though most have greater confidence.

Rationale for inclusion criteria and related procedures

Manual annotations, EC numbers, and GO terms were annotated hierarchically among one another because they describe general function (except for GO terms describing local properties, e.g. "metal-ion-binding protein"), and thus a GUF correctly annotated with a function is less likely to truly have an additional, distinct function (though possible, hence the inclusion of multiple EC numbers in some cases,

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig 9). In contrast, CATH topologies annotate structure alone, which, while ultimately giving rise to function, is orthogonal to GO and EC annotations in meaning and in method of derivation. Ligand-binding site (LBS) annotations are orthogonal to both annotation of large structural features and overall gene product function, as they are dictated primarily by local structure. Therefore, LBS and CATH topologies were annotated irrespective of EC, GO, or manual annotation, as well as one another.

"Overannotation"[119] with low-confidence predictions is the primary cause of annotation errors. To mitigate these errors, we sought to maximize annotation coverage only insofar as prior data (e.g. benchmarks, annotation status) suggested it more likely true than false. We chose 50% as the lowest precision incorporated so that any annotation incorporated was more likely correct than not (Figs 3-5). We used few broad classifications rather than many granular ones to 1) minimize false prioritization of PDB templates that appear more similar to native protein structure than true matches, due to inherent variability in structure prediction, and 2) to identify instances where templates with a similar degree structural similarity differed in function, for careful assessment. We employed one threshold above 50% to demarcate "putative" and "probable" qualifying adjectives. For parsimony, we chose 75%, the midpoint between 50% ("putative") and 100% (absolute certainty). TM-score error levels prevent more granular classification, as they would introduce false annotations at the expense of correct ones.

This approach of iterative inclusion with coarse hierarchical steps allowed differentiation of degrees of similarity when their difference was large (and thus more often encoding true differences in degree of similarity) without errant detection owing to noise from the limitations of structural prediction.

Error and bias in resolving protein structure and function are numerous, incompletely characterized, and distributed unevenly across protein classes and families[119]. This uncertainty and heterogeneity make accounting explicitly for sources of error challenging and time-consuming, if not intractable. To circumvent complications in accounting for these errors explicitly, we used the precision:TMID relation to capture PDB annotation reliability in a single metric and inform our inclusion criteria (Fig 8) With this approach, precision is considered independently of how PDB entries were annotated, and the observed precision (Fig 8B) reflects what was attained despite potential flaws or inconsistencies in PDB annotation procedures.

Remaining uncharacterized genes

Despite our multifaceted approach to annotation, many gene products remain unannotated. Among these are members of the enigmatic PE/PPE gene family. These genes are unique in their anomalously high GC% content and are specific to mycobacteria[120]. Their uniqueness to the MTBC complex makes finding homologs in PDB unlikely, particularly if their catalytic domains are absent or represented sparsely in PDB. Further clouding their characterization is the intracellular lifestyle of *M. tuberculosis*, which may render some of these proteins dependent on metabolic contexts or immunological cues specific to host microenvironments, and thus inactive *in vitro*. Ultimately, full characterization of the *M. tuberculosis* hypotheticome likely requires high-throughput biochemical assays, perhaps following methods development that allow direct assay or precise reconstruction of particular host microenvironments, *in vitro* or *ex vivo*. These present formidable logistical and technological challenges that may well take decades to resolve. In the meantime, our most capable inferential methods serve as valuable surrogate, albeit with caveats, limitations, and assumptions of their own.

Unfortunately, of the 1,711 GUF for which we were able to complete an I-TASSER run, over half (871, S15 Table) produced models of insufficient quality (C-score > -1.5)[25] to confidently imply structural similarity. Several phenomena may challenge effective modeling of this gene set: 1) No proteins of similar folds have been solved, leaving little to thread to 2) The protein is highly disordered[121] 3) These

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

are multi-domain proteins that need to be split into individual domains[19] 4) sequencing errors or gene coordinate misannotation[122].

We suspected reason 3 might be a primary cause, as I-TASSER documentation recommends breaking up multi-domain polypeptides into their constituent domains[19], which we did not. However, upon evaluating whether protein length differed between proteins of high (greater than -1.5) and low (below -1.5) C-scores, we found their respective distributions nearly indistinguishable (S4 Fig). This suggests multi-domain proteins did not cause poor modeling, at least in most cases.

The unexpected trove of apparent homologies/analogies where *M. tuberculosis* proteins appear to masquerade as host proteins for manipulation of host immune responses suggest that many of the genes not yet characterized play their roles in the host environment, perhaps in unique ways that challenge discovery through homology or analogy.

Limitations

Manual annotation.

While manual literature annotations are the most reliable, they are not immune to inaccuracy. For example, in one paper[123] Rv1818c, annotated as PE_PGRS33 by TubercuList, is mentioned as being down-regulated during nutrient starvation and oxygen depletion; however, their cited source[124] says that PE_PGRS33 showed no significant change in expression in any condition they tested, including nutrient starvation. Illustrating a more serious error, a 2013 paper claimed that Rv1749c is part of a VapBC toxin-antitoxin system in *M. bovis*[125], but their cited source[126] mentioned Rv1749c neither in the paper nor in the supplementary material. The same paper claims Rv0988 is a hydrolase, but provides no citation for this claim[125], and no major database supports this claim[2,5,17,34–36,38]. These examples underscore the need for skepticism and caution when evaluating evidence for functional annotation in literature.

Functional annotations in several genes may be contradicted by information published since the original annotation. We did not attempt to resolve such conflicts, as they should be resolved by specialists, particularly when the reason for discordant results between two groups were unclear.

Functional inference through structural similarity.

Limitations of our approach to functional inference from PDB similarity include our focus on global, rather than local, structural similarity. This approach makes functional inference challenging for proteins with functionally diverse folds[127] Other proteins that challenge functional annotation inference through this approach include those with active sites exhibiting a high degree of dynamism[128] or context-specific conformation and activity[129].

Not all GUF ran through I-TASSER produced complete models; fourteen of the 1,725 GUF failed their I-TASSER runs. Of these, six are annotated as pseudogenes in TubercuList, and their annotated AA may be unthreadable due to early termination codons. The remaining eight sequences belong to either PPE or PE_PGRS gene families, which are unique to mycobacteria, especially prone to sequencing errors, and intrinsically hypervariable[71] This may explain their inability to be resolved by I-TASSER's threading algorithm.

Another limitation lies in the overrepresentation of model organisms and human proteins that have been crystallized and uploaded to PDB, which could have exaggerated the prevalence on apparent human analogs to an extent.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Resources for further characterization

For accessibility, we provide final annotations in common machine and human readable formats. We provide a machine-readable GFF3 file containing our updated annotation (File S1). This file includes EC numbers, GO terms, CATH topologies, and product name annotations, and is ideal for reference transfer, variant effect prediction, and other bioinformatic analyses (File S1). If one wishes to create a GFF3 file with thresholds of greater or lesser stringency, they can do so using the Supplementary code and specifying altered thresholds.

Annotations in the GFF3 are defined by our inclusion criteria. PDB templates with structures similar yet below our criteria can consult S16 Table, which contains the top 3 PDB templates for each of the 1,711 GUF based on TMID, and S9 Table, which holds all matches where a $TM_{ADJ} > 0.52$ (Equation 2, Supplementary note) and/or TMID > EC3 (putative) threshold. Further exploration of GUF structure can be carried out by analyzing the I-TASSER output for each of the GUF at www.tuberculosis.sdsu.edu/resources/annotation/I-TASSER once files are made available, including functional predictions by COFACTOR, predicted ligand binding sites, local secondary structure confidence, and other potentially useful metrics.

Application of approach to other genomes

Our hierarchical, precision-guided approach to incorporating annotations protects against overannotation while increasing annotation coverage. These are desirable features for expanding annotation of any organism and should become more effective over time: Structural prediction methods should improve in accord with algorithm design, become more accessible as computational costs lower, and provide greater coverage as more protein structures and functions are elucidated. While we implemented this approach to maximize annotation of *M. tuberculosis*, it can be applied to other species. Employing these approaches to other species would help reconcile functional characterizations in the literature with what can be inferred from conservation of structure and function, and increase quality of functional data. We next plan to develop these methods into a tool with tunable inclusion criteria parameters, incorporate additional annotations from local features, and make the precision regression more robust and capable of considering additional parameters, such as overlap length.

Conclusion

To our knowledge, this work comprises the most comprehensive functional annotation of *M. tuberculosis* to date. Though inevitably containing some misannotations, this update provides a more complete view of the metabolic and pathogenic capabilities of *M. tuberculosis*, and clearly defines how annotations were derived. We will update this annotation on our GitLab site (see Data availability), where others can also submit merge requests to incorporate recent functional characterizations, which will be added following quality control (Methods).

We hope these annotations help the TB research community interpret and design future studies and prioritize candidate gene products for experimental characterization. Literature-curated annotations will better inform research design, and interpretation of functional implications from omics studies. Structurally inferred annotations supplement the manual curations, and enable target prioritization for confirmatory wet-lab work by the TB community, perhaps accelerating discovery in function of those gene with functions that remain unconfirmed. This update should be seen as an iteration, and re-implemented periodically to keep pace with future *M. tuberculosis* GUF characterizations. Periodic updates will also leverage the expanding set of solved structures to screen against and continuously integrate into reference and clinical strain annotation.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In principle, three main factors limit annotation: (i) Experimental and structural characterization of gene product function. (ii) Faithful annotation of function to sequence and function to structure in centralized databases. (iii) The capability of transferring annotations to similar products. With our strategy of controlled curation of empirically determined gene function followed by structured, informed incorporation of gene product annotation, the community can actualize the growing knowledge base of structure-function relationships in PDB, experimental characterizations of *M. tuberculosis* gene product function, and capability of sequence-structure-function prediction tools such as I-TASSER as they become more capable and resource-efficient. If we fail to implement such a workflow, we risk functional misattribution and propagation of aberrant annotation. Such gaps and falsehoods in our collective knowledge confuse and mislead research initiatives: well-reasoned decisions can misallocate resources, subvert the validity of high-throughput analyses, and, ultimately, stunt progress toward eradicating TB. The annotations collected, collated, and organized in this work step toward a more unified and effective path where gene product annotations are collected and stored in a standardized manner. Continuing and integrating this effort with the larger TB research community is essential to maintaining and accelerating progress toward understanding *M. tuberculosis* pathogenesis and reducing the global burden of Tuberculosis on public health.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and Methods

Systematic manual literature curation

We followed a standardized procedure to systematically curate gene function annotations from the literature, as outlined in Fig 6. The annotation of each GUF was extracted from TubercuList, comprising 1,057 unannotated ("hypothetical" or "unknown" functional categories) and 668 with ambiguous annotations (totaling 1,725 GUF). These annotations were converted to feature table format for ease of analysis and editing by curators, and version control through git was used to allow curators to work in parallel and track annotation progress of each gene. Each GUF was assigned to two curators to ensure no existing literature annotations were missed, and to remove any human bias in annotation. Each GUF was searched in Google scholar in format "mycobac* <GUF locus tag>" (e.g. "mycobac* Rv0004") and confined to publications between January 1st, 2010 and June 30, 2017. Earlier work was included when referenced in publications from the primary search. Patents and non-English articles were excluded. For each publication returned by the query, every mention of the locus tag or gene name was inspected manually unless it became apparent that the article did not contain information relevant to gene product function (e.g., purely an association study). Orthology and domain-based computational annotations were excluded when the sole basis of evidence, since TrEMBL annotations are regularly updated and are quality-controlled, and would likely catch such cases[17]. In contrast, orthology and domain-based annotation were included when combined with other evidence, such as using domain annotations to identify candidate genes for subsequent molecular docking simulations that putatively demonstrated product function. Curators evaluated experimental evidence for functional characterization and noted the methodology used to connect gene product and function according to NCBI Feature Table Format[33], using as descriptive language as possible while remaining concise and accurately representing the methods used and conclusions drawn by the primary authors.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

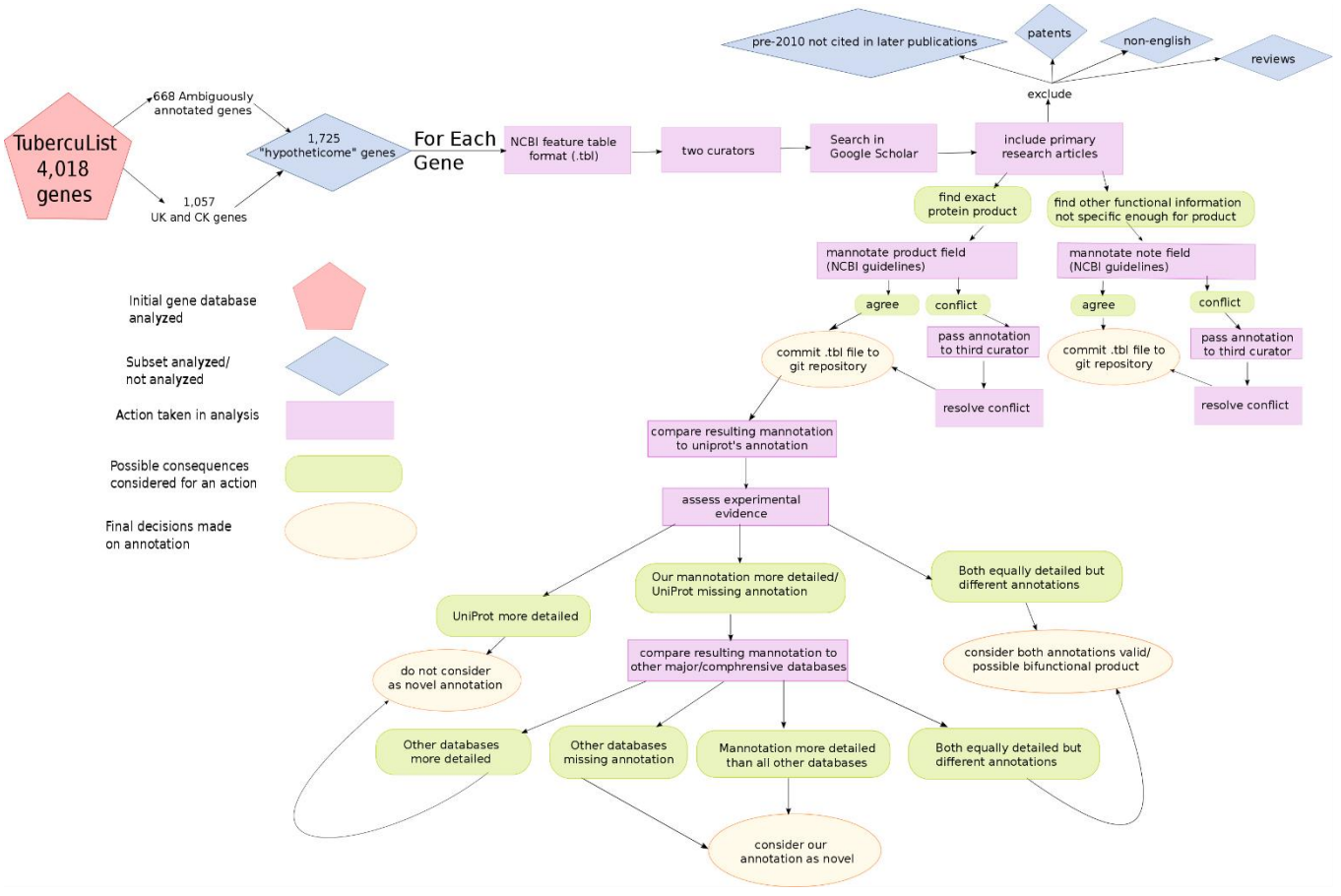


Fig 6. **Information flow for producing annotations from literature curation.**

An initial extraction of the existing annotation every "conserved hypothetical" and "unknown" protein from TubercuList totaled 1,057 unannotated protein-coding genes. Additionally, a 668 "ambiguous set" was manually determined from the annotations on TubercuList, and these annotations were extracted and combined with the 1,057 hypothetical and unknown proteins to give a total of 1,725 GUF. These genes were then searched in Google Scholar, and pertinent articles were analyzed for annotation information, which was recorded in NCBI's Table File Format (.tbl extension) for each gene, one file per GUF. Every GUF annotated with a novel product was compared to annotation in other databases (Results). Decisions were made based on the criteria described above.

Guidelines for manual product annotation.

To facilitate consistent annotation, we described products in one of three ways:

1. ***high-confidence*** - derived from evidence that, barring human error or data fabrication, definitively prove the annotated function is carried out by the GUF. This highest confidence assignment is implied by the absence of a qualifying adjective. Techniques warranting this annotation include protein purification with subsequent functional characterization through enzymatic assays, and gene knockout/complementation studies that isolate the GUF as the causal mechanism.
2. ***probable*** - used for experiments that provide strong evidence that the GUF carries out a certain function, but require minor assumptions or rely on strong but fallible inferences. Such techniques include X-ray crystallography with molecular docking and transposon mutagenesis studies.
3. ***putative***- used for experiments where a non-trivial assumption or inference with well-known exceptions is required. Examples include gene-knockout and complementation studies where an

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

observed phenotypic change (e.g. localization of substrate to inner cell-membrane) correlates with KO/complementation status, indicating mediation by the GUF. The putative qualifier was also used for *in silico* functional predictions based on three-dimensional methods, such as molecular docking simulations (PatchDock).

Experiments with insufficient evidence to assign "putative" annotation or higher were left with their product field unchanged, but annotated with a note, using either "potential" or "possible" as qualifiers, where the former connotes relatively higher confidence. The methods justifying these notes require significant assumptions, or are derived from incomplete information, and should be treated as tentative. Notes were also included for well-justified functional information not useful for defining a product name, such as "overexpression increases susceptibility to isoniazid *in vitro*".

Manual curation quality assurance.

Each gene query was assigned to two investigators, while a third was designated as "polisher" and served to assure quality and consistency. To hedge against human error, two investigators curated annotation for each gene, independently. To resolve discrepant annotations, the two investigators would consult with one another to produce a consensus annotation. If two investigators could not resolve the annotation, a third investigator (not necessarily the investigator serving as "polisher") would act as an arbitrator, and break the deadlock if no consensus could be reached.

After manual curation, every gene for which a new function had been assigned was inspected by the polisher. Polishers went to the source from which the new annotation was derived, and verified that the conclusions drawn by the initial manual curation were valid, correctly cited, and properly formatted. When the polisher felt the original annotation was inconsistent with annotation guidelines, they and the original curator(s) would discuss discrepant interpretations, and form a mutual consensus. If they could not, an additional arbitrator would confer with them to break the deadlock.

Comparison to existing databases.

To assess the novelty of manual product annotations, we compared each to their counterparts on UniProt[17], Mtb Network Portal[5] (which included annotations from TBDB[1]), PATRIC[2], RefSeq[34], BioCyc[35], and KEGG[36]. We obtained the UniProt, Mtb Network Portal, and RefSeq annotations for each of our genes with new product annotation programmatically. No parsable HTML could be obtained from the PATRIC website for the feature view of each gene, so PATRIC gene annotations were obtained manually. We then compared mannotations to annotations from each database. For each GUF, we determined whether the database annotations agreed or disagreed with our mannotations. For annotations that agreed, we recorded which source had the more descriptive annotation. Annotation from our literature curation absent in the other databases were considered candidate novel gene annotations. If the annotations disagreed, we considered our annotation a candidate for additional gene product annotation, since both our annotation and those in other databases may describe true functions (bifunctional/moonlighting proteins). Existence of functional annotations for these genes were tallied for each database to assess their comprehensiveness and identify discrepancies between them. Furthermore, genes unannotated in any of the listed databases, but with annotations assigned in this study, were identified and enumerated. EC number assignments were also compared among the databases (Results).

Incorporation of previous *in silico* annotation efforts for *M. tuberculosis*.

Previous large-scale *in silico* attempts at predicting hypothetical gene function were considered[12–16], but we did not incorporate their results in this work because they were not assigned clear confidence metrics. We instead opted to run the 1,725 GUF through I-TASSER[19] (See "Annotating genes of unknown

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

function via structural similarity") and evaluated whether the suite could produce informative *in silico* predictions.

Enzyme Commission number assignment.

Enzyme Commission (EC) numbers are period-delimited, hierarchical descriptors of enzyme-catalyzed reactions. The four EC number digits correspond to progressively granular description of reactions, and when assigned to a protein, imply that it catalyzes the described reaction[29] (Table 2). We annotated all literature-curated gene products with experimentally verified enzymatic activity with EC numbers. We used the EC number assigned by the authors of the source article of the annotation unless it did not follow International Union of Biochemistry and Molecular Biology (IUBMB) conventions[29]. When the authors assigned no EC number we assigned one ourselves according the reaction/EC number relations on the official IUBMB database[29]. EC numbers were assigned only to the degree of specificity warranted by experimental evidence (e.g. 3.1.-.- when esterase activity was shown, but evidence of no further substrate specificity was provided).

Annotating genes of unknown function via structural similarity

We designed our structural similarity-based annotation procedure to address the question "How likely is it that this annotation reflects true function?", and reasoned the question could be applied to our data as: "How do structural and sequence similarities correlate with the likelihood of matching annotations?". To answer this, we ran training genes products (TGP) of known function through I-TASSER (Iterative Threading ASSEmbly Refinement) suite (standalone, version 5.1) to observe how similarity metrics (TM-score, AA%, C-score, etc.) correlated with precision (Equation 2).

Following GUF selection, amino acid (AA) sequences were extracted from TubercuList and run through a local installation of I-TASSER for structural prediction and identification of candidate homologs and analogs in the protein data bank (PDB) through structural alignment[19,26]. From I-TASSER, metrics are computed that describe similarity between the GUF structure model and solved structures on PDB. For each query:PDB template proteins pair, similarity metrics were extracted along with EC, GO, and CATH annotations of the known protein (from PDB)[18]. We focused on two of these metrics that best correlated with precision (supplementary note): AA identity (AA%) and Template-modelling score ("TM-score", a measure of structural similarity independent of protein length).

TM-score describes structural similarity ranging from 0 and 1. It represents the average root-mean squared deviation across all atoms in the structural prediction with respect to the PDB template, normalized to remove apparent deviation arising falsely due to local differences[19,37] (Equation 1, as calculated by Zheng and Skolnick[37]).

$$TM\text{-}score = (1/L_N)\sum_{i=1}^{L_T}\frac{1}{(1+d_i^2+d_0^2)} \quad (1)$$

where L_N is protein length, L_T is the length of the aligned residues to the template, d_i is the distance of the i^{th} pair of residues between two structures after an optimal superposition, and $d_0 = 1.24\sqrt[3]{L_N - 15} - 1.8$, as described by Zhang *et al.*, normalizes for protein length. This metric allows structural similarity to be compared across proteins of different length[37].

We then assessed how precision of EC number and GO term predictions (Equation 2) correlated with similarity metrics (S17 Table) through logistic regression to identify those metrics most predictive of precision. We calculated precision as follows:

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$precision = \frac{(TP)}{(TP + FP)} \quad (2)$$

where TP = "True Positive", and FP = "False Positive". We used 50% and 75% as the cutoff for usage of "putative" and "probable" adjectives, respectively. We derived our training data from mannotated GUF with "probable" confidence or higher (which totaled 163 GUF, S2 Table) and a randomly chosen set of 200 *M. tuberculosis* genes with products of known function from TubercuList[38]. EC numbers were extracted from these 363 genes. GO terms were restricted to those marked as experimentally verified in UniProt[17] for the 200 random known genes and any in our GUF set. We ran these 363 sequences through I-TASSER to benchmark precision as a function of several similarity metrics, among which TM-score and AA% prevailed as the metrics most predictive of precision (Supplementary note).

CATH topologies (see next section for background) were annotated according to a previously established posterior probability distribution[25]. These EC number, GO term, and CATH thresholds were applied hierarchically (See next section for details) to update GUF annotations beyond what we could curate from the literature, and incorporated programmatically in NCBI's Table File Format (.tbl files) using Genbank Prokaryotic Annotation Guide[33] syntax and guidelines. Finally, we parsed the updated .tbl files with custom scripts (Supplemental Code) to produce a final GFF3 file (S1 File) of the updated H37Rv annotation for programmatic access and integration with common genomics programs. Fig 7 summarizes how the annotations were selected and integrated into a gene-based format distributable and integrable with common bioinformatics and annotation pipelines. Ligand-binding site (LBS) predictions from COFACTOR and product names inferred directly from homologous structure were included in a similar process but are omitted from Fig 7 to avoid crowding.

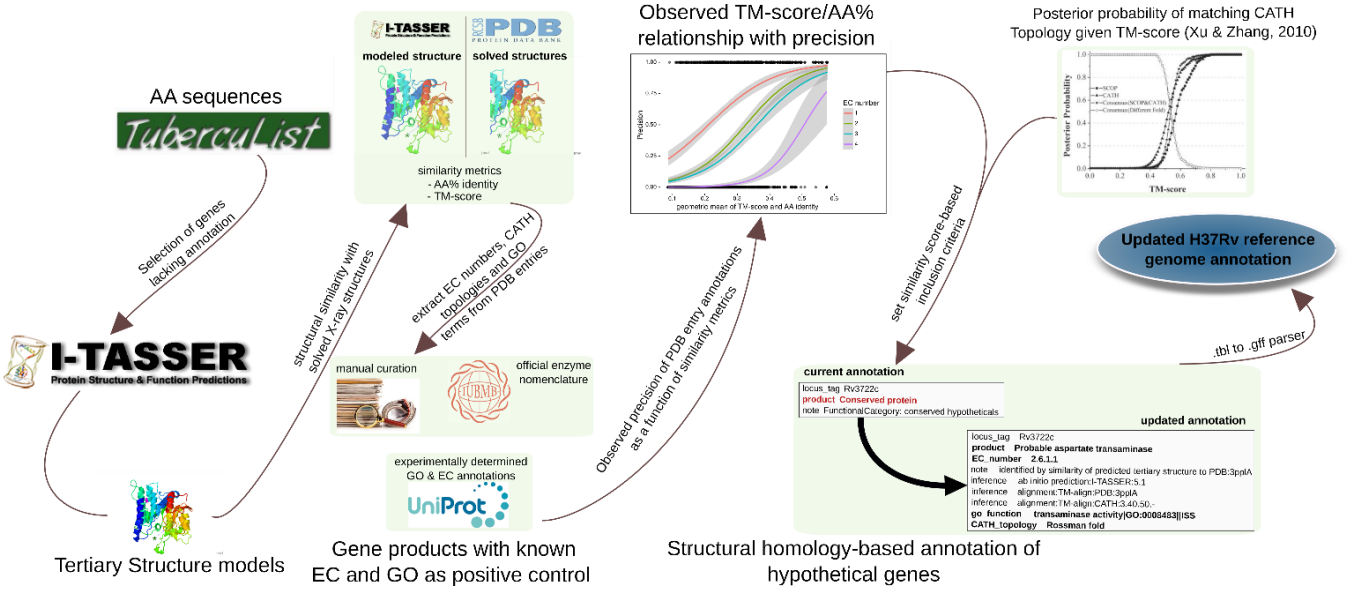


Fig 7. Information flow for producing annotations from structural homology.

The flow of information and procedures for acquiring, processing, filtering, and representing information, running from retrieval of amino acid sequences to the final updated H37Rv annotation. Some details are omitted for clarity.

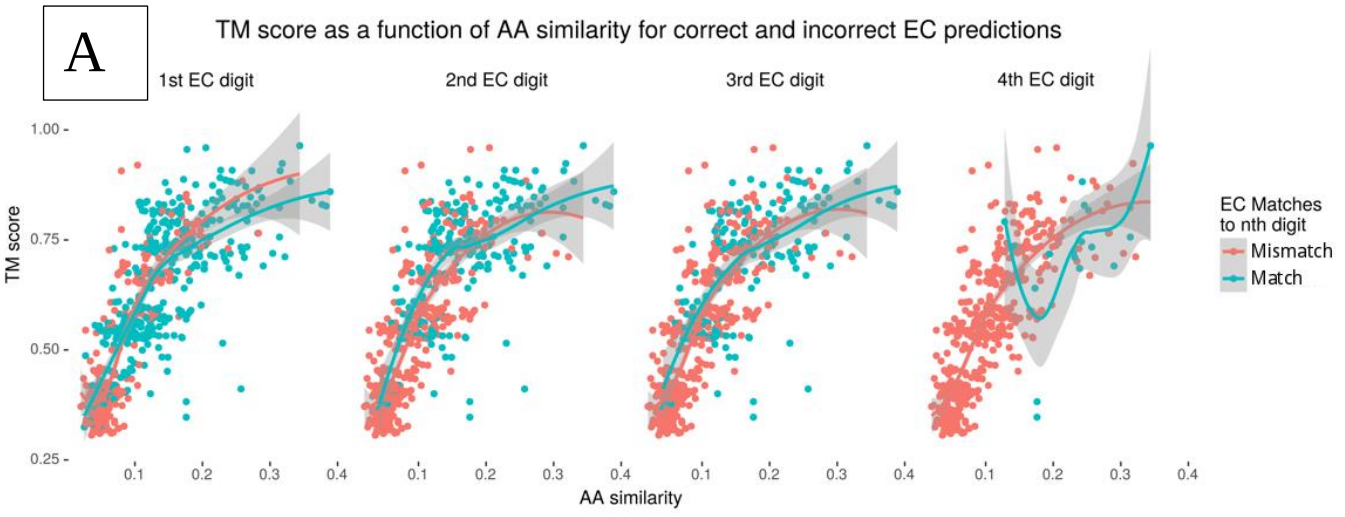
bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

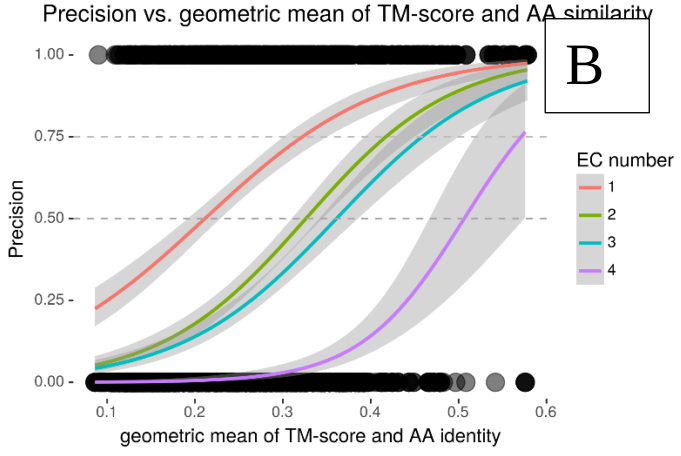
Annotation describing structure (CATH) and function (EC, GO, and product name-based transfers) were transferred from qualifying PDB templates to our GUF. Inclusion criteria were handled differently depending on the assumptions of each annotation framework, and the degree to which each could be inferred from global structural similarity, and other available metrics (e.g., LBS predictions from COFACTOR).

Inclusion criteria for annotations derived from structural similarity

We surveyed how matching and discordant EC assignments between GUF and PDB templates similar to I-TASSER structure predictions distributed with respect to sequence identity (AA%) and structural similarity (TM-score). As expected, the proportion of EC assignments discordant between training gene products and predicted homologs increased with the number of EC tiers evaluated (Fig 8A). AA% and TM-score correlated strongly with one another (R=0.784, Pearson correlation coefficient), among both concordant and discordant EC numbers (Fig 8A), suggesting both are informative for setting inclusion thresholds. To represent both AA% and structural similarity, we took their geometric mean (which we call “TMID”).

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.





C

Precision	EC digit			
	1	2	3	4
	TMID value at specified precision			
0.5	0.201	0.331	0.374	0.508
0.75	0.306	0.424	0.479	0.576

Fig 8. Determination of inclusion criteria for EC and GO annotations.

(A) TM-score and amino acid sequence identity (AA%) colored by correctness in the sample data. Dots represent pairwise relations between query protein and PDB template. Their position indicates structural similarity (TM-score, y-axis) and AA% (x-axis) and their color indicates concordant (red) or discordant (blue) EC number, to the specificity indicated in the pane. (B) Precision of EC number as a function of the geometric mean of TM-score and AA% ("TMID"). Precision was regressed on TMID for each of the four tiers of EC specificity. Horizontal lines indicate the precision cutoffs used to set thresholds for hierarchical incorporation of annotations. Circles at the bottom and top are individual data points (representing 0 for incorrect and 1 for correct, at a particular TMID value). Circles are rendered at 10% opacity to visually depict observation density (C) Table showing TMID cutoffs corresponding to 50 and 75 percent precision for each of the 4 EC number digits. In all analyses, templates with AA% > 40% were excluded to isolate matches due to structural similarity rather than AA identity.

We binomially regressed precision (Equation 2) against TMID using the training gene products to determine expected precision of EC of our GUF and PDB entries similar to its predicted structure (Fig 8B). The resulting regression lines for each EC digit informed cutoffs in our inclusion criteria based on expected precision (Fig 8C).For benchmarking based on structure alone, only templates with C-scores above -1.5 were included, as structural predictions with lower confidence are unlikely to reflect correct protein topology[19].

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Since TM-score and AA% were similarly predictive of precision (Fig 8), we used their geometric mean (Equation 1) to estimate precision of EC and GO predictions based on the logistic regression curves shown in Fig 8B. We used this estimated precision as the 50% and 75% cutoffs for "putative" and "probable" qualifying adjectives (Fig 8C).

Though TMID proved useful for estimating precision, the sample distribution of AA% and TM-score did not cover areas where one was high and the other low (Fig 8A). While high AA% should be captured by annotation methods based on sequence identity, high TM-score with low AA% is only amenable to annotation through structural similarity. However, the relationship between TM-score and precision of functional annotations in such cases is unclear. For these cases with high TM-score, we transferred "CATH structural annotations" from similar PDB entries based on a precision:TM-score relation determined previously[25]. We also implemented special inclusion criteria for protein classes that typically lack sequence similarity despite sharing structure and function.

CATH is a hierarchic classification system of protein domain structures, in which "topology" is the third level of the hierarchy, more specific than "architecture" and more general than "superfamily"[39]. Structural fold annotations can be functionally informative in some cases.

For Ligand-binding site (LBS) and CATH predictions, we applied benchmarks previously established[25,40].

The relationship between precision and TM-score has been rigorously benchmarked in prior work by Zheng and colleagues[25]. This precision:TM-score relation for matching CATH topologies between two proteins follows an extreme value distribution[25], which confers high discriminatory power to a binary inclusion threshold. We set the minimum TM-score for inclusion as that which corresponded to 50% precision, after correcting TM-score for the expected modeling error (encoded by the C-score, Supplementary Note). CATH annotations were retrieved using the REST API of PDB. All analyses were implemented in R[41].

Hierarchical annotation

To determine final GUF product annotations we prioritized more reliable methods of inference before deferring to less reliable methods. We first included mannotations (highest priority) and high-confidence EC-based annotations from structural similarity. EC-based annotations were considered with mannotations so that secondary functions of moonlighting proteins wouldn't be precluded from annotation. We incorporated annotations in a gene-wise manner, hierarchically between annotation frameworks (mannotation and EC > GO) and iteratively (iterating over progressively looser thresholds within GO matches) within each GUF. EC annotations were included when projected precision exceeded 50% (Fig 8A).

Rather than directly mapping PDB template name to product annotations, we derived product names using EC number and GO terms. This enabled transfer of shared aspects of similar proteins without implying identical function. For example, a GUF may be similar enough to confidently transfer "methyltransferase" annotation to our GUF, but not "6-methyladenine DNA methyltransferase", the product of the PDB template. Using the name of exact PDB match would imply greater specificity than the degree of structural similarity warrants, which is often misleading.

Enzyme commission numbers.

EC number annotations were transferred to the degree of specificity dictated by the threshold they met (Fig 9). If no EC-bearing PDB:GUF matches passed threshold, the GUF progressed to the GO protocol,

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

during which EC numbers were inferred for some products (see next subsubsection). Conflicting EC numbers were pruned to a common digit if possible, and collapsed into unique ECs at the degree of specificity dictated by their similarity if not (Supplemental Note). Final EC numbers informed product field annotations. In cases where product field had a prior entry, if the EC annotation of greater confidence was less specific, then product annotation was modified to reflect this EC curation instead of a more specific, less confident EC annotation. Threshold boundaries were set at 75% and 50% precision for the 4th and 3rd digits of specificity in EC numbers (Fig 9).

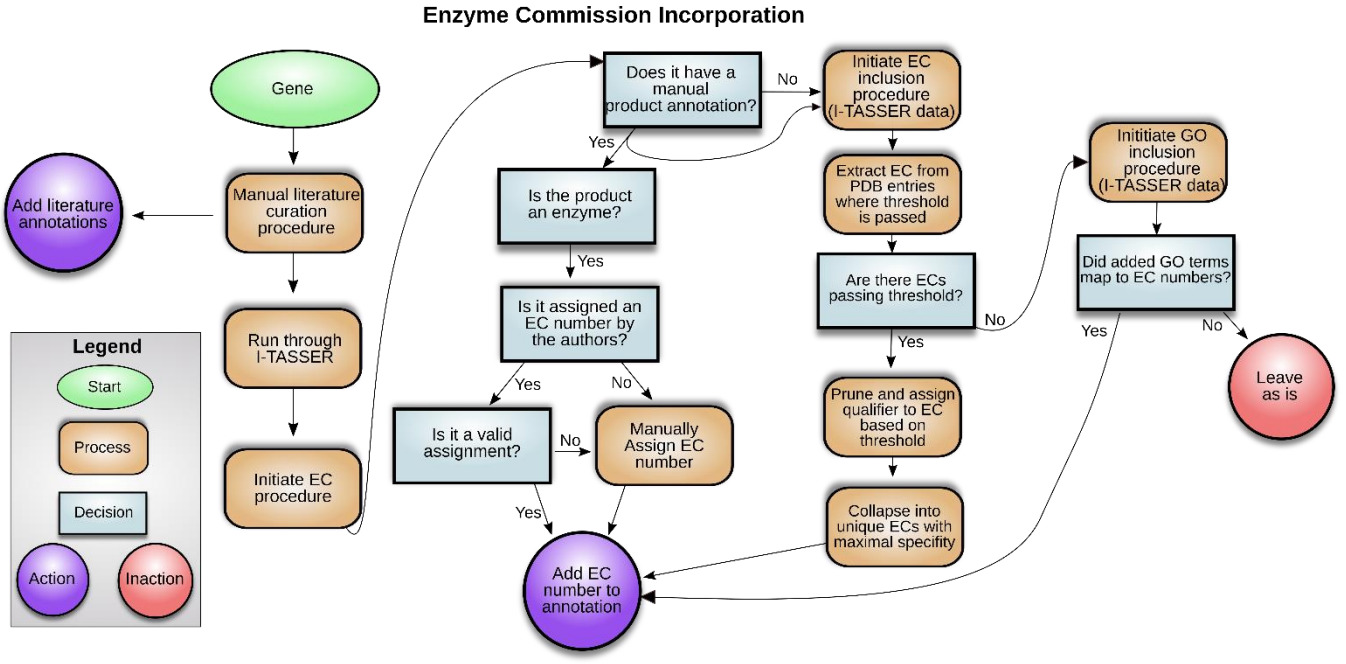


Fig 9. Enzyme Commission (EC) number inclusion protocol.

The flow of processes and decisions each GUF was subjected to for determining EC based annotations from structural homologs. Most processes and decisions were implemented in a fully automated manner, but some corner cases had to be resolved manually. These cases were handled algorithmically, or by previously established procedures where possible. For example, when EC numbers had to be assigned manually, the procedures put forth by IUBMB were consulted and followed directly[29].

Gene Ontology Terms.

GUFs unannotated by the EC inclusion protocol were screened for PDB matches at progressively lower TMID thresholds for GO term transfer (Fig 10). With this approach, GUF receive only the best available annotations, since GUF already annotated (from sources of higher expected precision) exit the iteration before progressing to weaker thresholds. Unlike using a single threshold, this approach mitigates overannotation with assignments of lower confidence, without lowering the number of annotated genes. Each GO term transferred was labeled with all the PDB templates supporting it. EC numbers were inferred programmatically and included where a direct mapping between EC and GO term existed. If the GO term implied enzymatic activity, relevant terms were searched in EXPaSY ENZYME and an EC number was GUF assigned and used to derive the product name from the ENZYME.dat file obtained from the EXPaSY database. When GO terms mapped to multiple EC numbers, the EC numbers were merged at the most specific level at which they converged (e.g., 3.2.1.5 and 3.2.2.4 would resolve to

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2.-.-). Such instances of EC assignments from GO terms occurred often because of inconsistent annotation across frameworks in PDB entries: many PDB entries were not assigned EC numbers, but were assigned GO terms based on enzymatic activity. If GO terms were transferred but none mapped to EC numbers, GO terms were examined for terms sufficiently descriptive to constitute a product name (e.g., "DNA binding transcription factor activity" is sufficiently descriptive whereas "pathogenesis" is not). All GO terms remain in the annotation file (S1 File), but where no GO term warranted product annotation, product field was left unchanged, and not included in the final counts (Results).

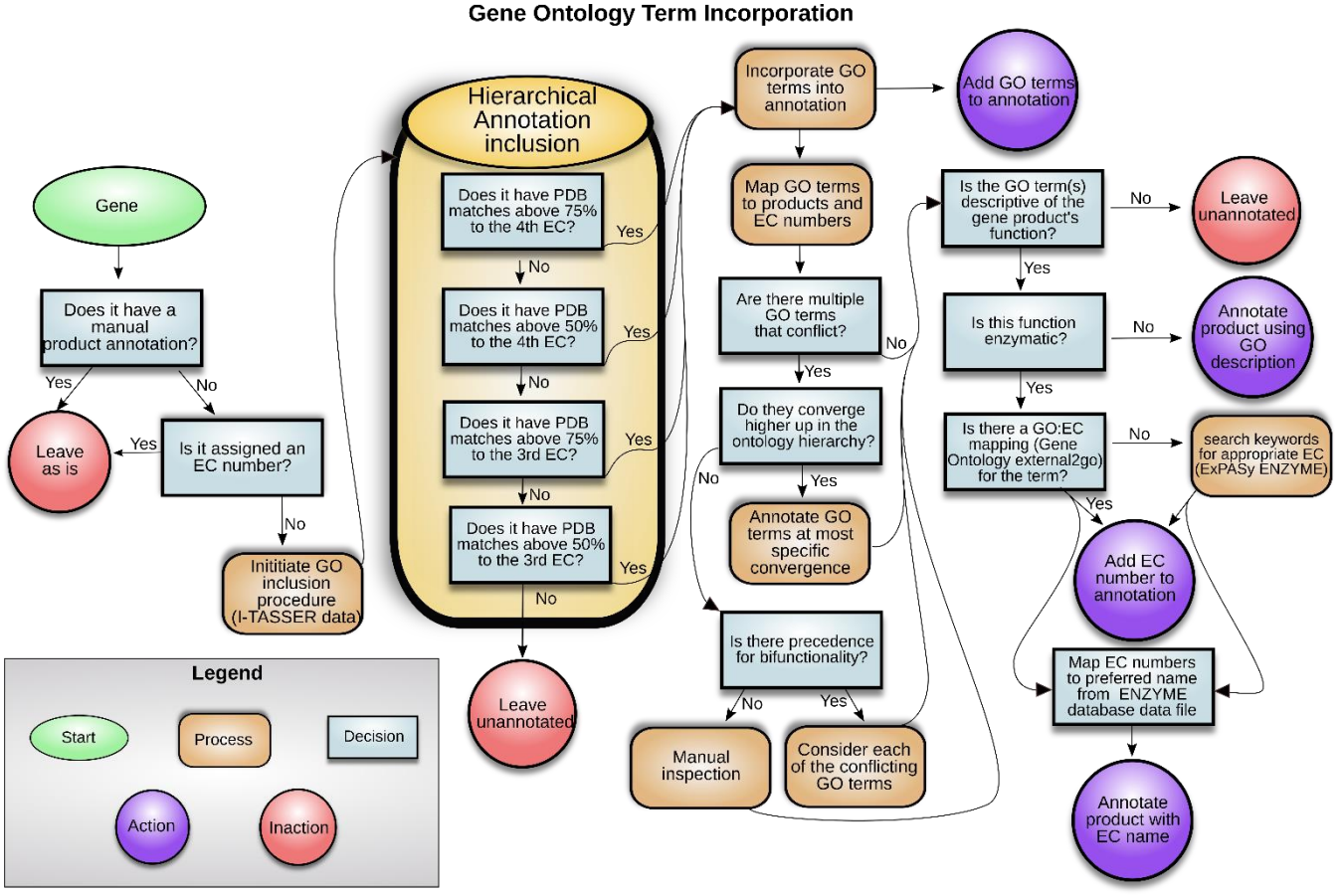


Fig 10. Gene Ontology (GO) term inclusion protocol.

The flow of processes and decisions each GUF was subjected to for determining GO based annotations from structural homologs. All processes and decisions were implemented in a fully automated manner, up until product assignment, those of which did not map to EC number had to be resolved manually.

Name-based product annotation from structurally similar PDB templates.

Many GUF with quality structure models (C-score > -1.5) had similarity only to PDB templates that lacked EC or GO term annotations. These GUF were algorithmically included using the same TMID thresholds described for EC and GO inclusion (Figs 4 and 5) but assigned product names manually. The top 3 PDB templates of GUF lacking functional annotation (EC, GO, or manual) were examined. Each query:template pairwise relation with a TM-score greater than 0.85 and/or TMID meeting the inclusion criteria for putative EC 3rd digit (0.374, corresponding to a precision > 0.5) were considered for annotation by these criteria: If (i) the portion of the GUF similar to the PDB entry structurally aligned with the PDB entry crystal structure along the coordinates annotated with the functional motifs

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

responsible for the function under consideration for transfer and (ii) the PDB template had the function verified experimentally in its UniProt entry or in the primary publication that claimed to elucidate its function (typically in the same article that determined its structure). When both criteria were met, annotation was transferred to the GUF to the degree of specificity warranted by the TMID, and product names were assigned as detailed for GO terms (Fig 8).

Structure-based transport protein annotation.

Transport proteins require different inclusion criteria than globular proteins. These proteins are especially difficult to characterize experimentally, and more conserved in structure than in sequence, relative to globular proteins[42]. To accommodate these unique features, inclusion criteria for transferring annotations from PDB templates of transport proteins weighted structural similarity more heavily than AA%: These annotations were transferred if greater than 90% of the PDB implicated in transport aligned to by the GUF, and structural similarity exceeded the threshold for CATH topology transfer. Transport protein annotations were transferred to GUF in a less specific form than that given the PDB templates (e.g. "transport protein" instead of "Na+/H+ antiporter"), unless all three highest templates matched a more specific description and the TM-score exceeded 0.85, in which case a more specific product name was transferred.

Data acquisition

We extracted functional annotations, EC numbers, and GO terms for each gene in our set of 1,725 GUF, from Entrez[34], Mtb Network Portal[5], TubercuList[38], PATRIC[2], KEGG[36], UniProt[17], and BioCyc[35]. Counts were obtained using custom python html scraper scripts.

Data Availability

Our final updated annotation of the 1,725 ambiguously annotated GUF has been provided in Supplementary file 1 and will be available on the Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence (LPCDRP) website at https://tuberculosis.sdsu.edu/. Continued updates will be made available on our soon-to-be public Lab GitLab site: https://gitlab.com/LPCDRP/Mtb-H37Rv-annotation/

Author contributions

SM, AE, and FV conceptualized the overarching research goals and aims; SM, AE, DG, AZ, NK, and CC developed the manual curation protocol; SM, DG, AZ, NK, and CC manually curated annotations from literature, and independently validated annotations; SM, AE, DG, AZ, and NK developed and tested code for data processing and analysis; AE implemented the I-TASSER suite on local resources; SM, AE, DG, and AZ prepared the manuscript; SM, DG, and AZ prepared figures and tables; SM and FV managed the project.

Funding

This work was funded by a grant from National Institute of Allergy and Infectious Diseases (NIAID Grant No. R01AI105185). SM, AZ, DG, AE, NK, CC and FV were supported by this grant. SM was also supported by scholarships from a National Science Foundation Grant (no. 0966391). The funding bodies had no role in the design of the study or collection, analysis, and interpretation of data or in writing the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

We thank Sarah Ramirez-Busby and Derek Conkle-Gutierrez for their extensive review of the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Reddy, T. B. K. *et al.* TB database: an integrated platform for tuberculosis research. *Nucleic Acids Res.* **37,** D499–D508 (2009).
2. Gillespie, J. J. *et al.* PATRIC: the Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species ▿. *Infect. Immun.* **79,** 4286–4298 (2011).
3. Garcia, B. J., Datta, G., Davidson, R. M. & Strong, M. MycoBASE: expanding the functional annotation coverage of mycobacterial genomes. *BMC Genomics* **16,** 1102 (2015).
4. Schito, M. & Dolinger, D. L. A Collaborative Approach for 'ReSeq-ing' Mycobacterium tuberculosis Drug Resistance: Convergence for Drug and Diagnostic Developers. *EBioMedicine* **2,** 1262–1265 (2015).
5. Ma, S. *et al.* Integrated Modeling of Gene Regulatory and Metabolic Networks in Mycobacterium tuberculosis. *PLOS Comput. Biol.* **11,** e1004543 (2015).
6. Faksri, K., Tan, J. H., Chaiprasert, A., Teo, Y.-Y. & Ong, R. T.-H. Bioinformatics tools and databases for whole genome sequence analysis of Mycobacterium tuberculosis. *Infect. Genet. Evol.* (2016). doi:10.1016/j.meegid.2016.09.013
7. Steenken, W., Oatway, W. H. & Petroff, S. A. BIOLOGICAL STUDIES OF THE TUBERCLE BACILLUS. *J. Exp. Med.* **60,** (1934).
8. Cole, S. T. *et al.* Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393,** 537–544 (1998).
9. Camus, J.-C., Pryor, M. J., Médigue, C. & Cole, S. T. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology* **148,** 2967–2973 (2002).
10. Ijaq, J., Chandrasekharan, M., Poddar, R., Bethi, N. & Sundararajan, V. S. Annotation and curation of uncharacterized proteins- challenges. *Front. Genet.* **6,** (2015).
11. Sivashankari, S. & Shanmughavel, P. Functional annotation of hypothetical proteins – A review. *Bioinformation* **1,** 335–338 (2006).
12. Ramakrishnan, G. *et al.* Enriching the annotation of Mycobacterium tuberculosis H37Rv proteome using remote homology detection approaches: insights into structure and function. *Tuberculosis (Edinb).* **95,** 14–25 (2015).
13. Mao, C. *et al.* Functional assignment of Mycobacterium tuberculosis proteome revealed by genome-scale fold-recognition. *Tuberculosis (Edinb).* **93,** 40–6 (2013).
14. Al-Khafaji, Z. M. In Silico Investigation of Rv Hypothetical Proteins of Virulent Strain Mycobacterium tuberculosis H37Rv. *http://ijpbr.in/wp-content/uploads/2013/12/15-In-Silico-Investigation-of-Rv-Hypothetical-Proteins-of-Virulent-Strain-Mycobacterium-tuberculosis-H37Rv.pdf* (2013).
15. Doerks, T., van Noort, V., Minguez, P. & Bork, P. Annotation of the M. tuberculosis hypothetical orfeome: adding functional information to more than half of the uncharacterized proteins. *PLoS One* **7,** e34302–e34302 (2012).
16. Mazandu, G. K. & Mulder, N. J. Using the underlying biological organization of the Mycobacterium tuberculosis functional network for protein function prediction. *Infect. Genet. Evol.* **12,** 922–932 (2012).
17. UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).
18. Rose, P. W. *et al.* The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45,** D271–D281 (2017).
19. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12,** 7–8 (2015).
20. Moult, J. *et al.* Critical assessment of methods of protein structure prediction—Round VII. *Proteins Struct. Funct. Bioinforma.* **69,** 3–9 (2007).
21. Kryshtafovych, A., Krysko, O., Daniluk, P., Dmytriv, Z. & Fidelis, K. Protein structure prediction center in CASP8. *Proteins Struct. Funct. Bioinforma.* **77,** 5–9 (2009).
22. Moult, J., Fidelis, K., Kryshtafovych, A. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)—round IX. *Proteins Struct. Funct. Bioinforma.* **79,** 1–5 (2011).
23. Huang, Y. J., Mao, B., Aramini, J. M. & Montelione, G. T. Assessment of template-based protein structure predictions in CASP10. *Proteins Struct. Funct. Bioinforma.* **82,** 43–56 (2014).
24. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins Struct. Funct. Bioinforma.* **84,** 4–14 (2016).
25. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26,** 889–895 (2010).
26. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33,** 2302–9 (2005).
27. Zhang, C., Freddolino, P. L. & Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.* doi:10.1093/nar/gkx366
28. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29. Webb, E. C. & others. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* (Academic Press, 1992).
30. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **41,** D1096-103 (2013).
31. Sangrador-Vegas, A., Mitchell, A. L., Chang, H.-Y., Yong, S.-Y. & Finn, R. D. GO annotation in InterPro: why stability does not indicate accuracy in a sea of changing annotations. *Database* **2016,** baw027 (2016).
32. Liberal, R. & Pinney, J. W. Simple topological properties predict functional misannotations in a metabolic network. *Bioinformatics* **29,** i154–i161 (2013).
33. Detailed Annotation Guide.
34. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* **42,** D553–D559 (2014).
35. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* **28,** 1–6 (2017).
36. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45,** D353–D361 (2017).
37. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Bioinforma.* **57,** 702–710 (2004).
38. Lew, J. M., Kapopoulou, A., Jones, L. M. & Cole, S. T. TubercuList--10 years after. *Tuberculosis* **91,** 1–7 (2011).
39. Orengo, C. A. *et al.* CATH – a hierarchic classification of protein domain structures. *Structure* **5,** 1093–1109 (1997).
40. Roy, A., Yang, J. & Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **40,** W471–W477 (2012).
41. Team, R. C. R: A language and environment for statistical computing. (2014).
42. Theobald, D. L. & Miller, C. Membrane transport proteins: surprises in structural sameness. *Nat. Struct. Mol. Biol.* **17,** 2–3 (2010).
43. Olive, A. J. & Sassetti, C. M. Metabolic crosstalk between host and pathogen: sensing, adapting and competing. *Nat. Rev. Microbiol.* **14,** 221–234 (2016).
44. Ortega, C. *et al.* Systematic Survey of Serine Hydrolase Activity in Mycobacterium tuberculosis Defines Changes Associated with Persistence. *Cell Chem. Biol.* **23,** 290–298 (2016).
45. Poux, S. *et al.* On expert curation and scalability: UniProtKB/ Swiss-Prot as a case study.
46. Chen, J., Guo, M., Wang, X. & Liu, B. A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief. Bioinform.* doi:10.1093/bib/bbw108
47. Slama, N. *et al.* The changes in mycolic acid structures caused by hadC mutation have a dramatic effect on the virulence of Mycobacterium tuberculosis. *Mol. Microbiol.* **99,** 794–807 (2016).
48. Minch, K. J. *et al.* The DNA-binding network of Mycobacterium tuberculosis. *Nat. Commun.* **6,** 5829 (2015).
49. Galagan, J. E. *et al.* The Mycobacterium tuberculosis regulatory network and hypoxia. *Nature* **499,** 178–83 (2013).
50. Singh, A. *et al.* Mycobacterium tuberculosis WhiB3 Maintains Redox Homeostasis by Regulating Virulence Lipid Anabolism to Modulate Macrophage Response. *PLoS Pathog.* **5,** e1000545 (2009).
51. Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W. & Wilson, D. J. Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* **14,** 150–162 (2016).
52. Drayman, N. *et al.* Pathogens Use Structural Mimicry of Native Host Ligands as a Mechanism for Host Receptor Engagement. *Cell Host Microbe* **14,** 63–73 (2013).
53. Vanier, M. T. Complex lipid trafficking in {Niemann}-{Pick} disease type {C}. *J. Inherit. Metab. Dis.* **38,** 187–199 (2015).
54. Fineran, P. *et al.* Pathogenic mycobacteria achieve cellular persistence by inhibiting the Niemann-Pick Type C disease cellular pathway. *Wellcome Open Res.* **1,** 18 (2016).
55. Arora, N. & K. Banerjee, A. Targeting {Tuberculosis}: {A} {Glimpse} of {Promising} {Drug} {Targets}. *Mini Rev. Med. Chem.* **12,** 187–201 (2012).
56. Moult, J. & Melamud, E. From fold to function. *Curr. Opin. Struct. Biol.* **10,** 384–389 (2000).
57. Vijayabaskar, M. S. & Vishveshwara, S. Insights into the Fold Organization of TIM Barrel from Interaction Energy Based Structure Networks. *PLOS Comput. Biol.* **8,** e1002505 (2012).
58. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492,** 193–198 (2001).
59. Goldman, A. D., Beatty, J. T. & Landweber, L. F. The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism. *J. Mol. Evol.* **82,** 17–26 (2016).
60. Dawson, N. L. *et al.* CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45,** D289–D295 (2017).
61. Evans, J. C. & Mizrahi, V. The application of tetracyclineregulated gene expression systems in the validation of novel drug targets in Mycobacterium tuberculosis. *Front. Microbiol.* **6,** (2015).
62. Gandotra, S., Schnappinger, D., Monteleone, M., Hillen, W. & Ehrt, S. In vivo gene silencing identifies the

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mycobacterium tuberculosis proteasome as essential for persistence in mice. *Nat. Med.* **13,** 1515–1520 (2007).
63. Blumenthal, A., Trujillo, C., Ehrt, S. & Schnappinger, D. Simultaneous Analysis of Multiple Mycobacterium tuberculosis Knockdown Mutants In Vitro and In Vivo. *PLoS One* **5,** (2010).
64. Holliday, G. L., Davidson, R., Akiva, E. & Babbitt, P. C. in *The Gene Ontology Handbook* 111–132 (Humana Press, New York, NY, 2017).
65. Cambier, C. J., Falkow, S. & Ramakrishnan, L. Host Evasion and Exploitation Schemes of Mycobacterium tuberculosis. *Cell* **159,** 1497–1509 (2014).
66. Schorey, J. S., Cheng, Y., Singh, P. P. & Smith, V. L. Exosomes and other extracellular vesicles in host–pathogen interactions. *EMBO Rep.* **16,** 24–43 (2015).
67. Smith, V. L., Cheng, Y., Bryant, B. R. & Schorey, J. S. Exosomes function in antigen presentation during an in vivo Mycobacterium tuberculosis infection. *Sci. Rep.* **7,** (2017).
68. Black, P. A. *et al.* Energy Metabolism and Drug Efflux in Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **58,** 2491–2503 (2014).
69. Heng, J. *et al.* Substrate-bound structure of the E. coli multidrug resistance transporter MdfA. *Cell Res.* **25,** 1060–1073 (2015).
70. Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45,** D331–D338 (2017).
71. Fishbein, S., van Wyk, N., Warren, R. M. & Sampson, S. L. Unbound MEDLINE : Phylogeny to function: PE/PPE protein evolution and impact on Mycobacterium tuberculosis pathogenicit. *Molecular Microbiology* **96,** n/a-n/a (2015).
72. Sultana, R., Vemula, M. H., Banerjee, S. & Guruprasad, L. The PE16 (Rv1430) of Mycobacterium tuberculosis Is an Esterase Belonging to Serine Hydrolase Superfamily of Proteins. *PLoS One* (2013). doi:10.1371/journal.pone.0055320
73. Padhi, A., Naik, S. K., Sengupta, S., Ganguli, G. & Sonawane, A. Expression of Mycobacterium tuberculosis NLPC/p60 family protein Rv0024 induce biofilm formation and resistance against cell wall acting anti-tuberculosis drugs in Mycobacterium smegmatis. *Microbes Infect.* **18,** 224–236 (2016).
74. Kateete, D. P. *et al.* Rhomboids of Mycobacteria: Characterization Using an aarA Mutant of Providencia stuartii and Gene Deletion in Mycobacterium smegmatis. *PLoS One* (2012). doi:10.1371/journal.pone.0045741
75. Nambi, S. *et al.* The Oxidative Stress Network of Mycobacterium tuberculosis Reveals Coordination between Radical Detoxification Systems. *Cell Host Microbe* (2015). doi:10.1016/j.chom.2015.05.008
76. Flentie, K., Garner, A. L. & Stallings, C. L. Mycobacterium tuberculosis transcription machinery: Ready to respond to host attacks. *Journal of Bacteriology* (2016). doi:10.1128/JB.00935-15
77. Primm, T. P. *et al.* The Stringent Response of Mycobacterium tuberculosis Is Required for Long-Term Survival. *J. Bacteriol.* **182,** 4889–4898 (2000).
78. Hu, D. *et al.* Autophagy regulation revealed by SapM-induced block of autophagosome-lysosome fusion via binding RAB7. *Biochem. Biophys. Res. Commun.* **461,** 401–407 (2015).
79. Prashar, A., Schnettger, L., Bernard, E. M. & Gutierrez, M. G. Rab GTPases in Immunity and Inflammation. *Front. Cell. Infect. Microbiol.* **7,** (2017).
80. Chandra, P. *et al.* {Mycobacterium tuberculosis Inhibits RAB7 Recruitment to Selectively Modulate Autophagy Flux in Macrophages. *Sci. Rep.* **5,** 16320 (2015).
81. Personnic, N., Bärlocher, K., Finsel, I. & Hilbi, H. Subversion of Retrograde Trafficking by Translocated Pathogen Effectors. *Trends Microbiol.* **24,** 450–462 (2016).
82. Khosla, S., Sharma, G. & Yaseen, I. Learning epigenetic regulation from mycobacteria. *Microb. Cell* **3,** 92–94
83. Zhu, L. *et al.* Precision methylome characterization of Mycobacterium tuberculosis complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Res.* 1–14 (2015). doi:10.1093/nar/gkv1498
84. Shell, S. S. *et al.* DNA Methylation Impacts Gene Expression and Ensures Hypoxic Survival of Mycobacterium tuberculosis. *PLOS Pathog* **9,** e1003419 (2013).
85. Escoll, P., Mondino, S., Rolando, M. & Buchrieser, C. Targeting of host organelles by pathogenic bacteria: a sophisticated subversion strategy. *Nat. Rev. Microbiol.* **14,** 5 (2016).
86. Yuan, S. & Akey, C. W. Apoptosome structure, assembly, and procaspase activation. *Struct. (London, Engl. 1993)* **21,** 501–515 (2013).
87. So, E. C., Mattheis, C., Tate, E. W., Frankel, G. & Schroeder, G. N. Creating a customized intracellular niche: subversion of host cell signaling by {Legionella} type {IV} secretion system effectors. *Can. J. Microbiol.* **61,** 617–635 (2015).
88. Gillespie, J. J. *et al.* The {Rickettsia} type {IV} secretion system: unrealized complexity mired by gene family expansion. *Pathog. Dis.* **74,** (2016).
89. Maertens, G. N. *et al.* Structural basis for nuclear import of splicing factors by human Transportin 3. *Proc. Natl. Acad. Sci.* **111,** 2728–2733 (2014).
90. Bell, L. C. K. & Noursadeghi, M. Pathogenesis of HIV-1 and Mycobacterium tuberculosis co-infection. *Nat. Rev.*

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Microbiol. (2017). doi:10.1038/nrmicro.2017.128

91. Danelishvili, L., Everman, J., McNamara, M. & Bermudez, L. Inhibition of the plasma-membrane-associated serine protease cathepsin G by Mycobacterium tuberculosis Rv3364c suppresses caspase-1 and pyroptosis in macrophages. *Cell. Infect. Microbiol. - closed Sect.* **2,** 281 (2012).
92. Hegde, S. S. *et al.* A Fluoroquinolone Resistance Protein from Mycobacterium tuberculosis That Mimics DNA. *Science (80-.).* **308,** 1480–1483 (2005).
93. Leo, J. C., Grin, I. & Linke, D. Type {V} secretion: mechanism(s) of autotransport through the bacterial outer membrane. *Philos. Trans. R. Soc. B Biol. Sci.* **367,** 1088–1101 (2012).
94. Pires, D. *et al.* Role of {Cathepsins} in \textit{{Mycobacterium} tuberculosis} {Survival} in {Human} {Macrophages}. *Sci. Rep.* **6,** 32247 (2016).
95. Kolli, N. & Garman, S. C. Proteolytic Activation of Human Cathepsin A. *J. Biol. Chem.* **289,** 11592–11600 (2014).
96. Gröschel, M. I., Sayes, F., Simeone, R., Majlessi, L. & Brosch, R. ESX secretion systems: mycobacterial evolution to counter host immunity. *Nat. Rev. Microbiol.* (2016). doi:10.1038/nrmicro.2016.131
97. Tundup, S., Mohareer, K. & Hasnain, S. E. Mycobacterium tuberculosis {PE}25/{PPE}41 protein complex induces necrosis in macrophages: {Role} in virulence and disease reactivation? *FEBS Open Bio* **4,** 822–828 (2014).
98. Ruiz-Perez, F. & Nataro, J. P. Bacterial serine proteases secreted by the autotransporter pathway: classification, specificity and role in virulence. *Cell. Mol. Life Sci.* **71,** 745–770 (2014).
99. Stubbendieck, R. M., Vargas-Bautista, C. & Straight, P. D. Bacterial {Communities}: {Interactions} to {Scale}. *Front. Microbiol.* **7,** (2016).
100. Titgemeyer, F. *et al.* A genomic view of sugar transport in Mycobacterium smegmatis and Mycobacterium tuberculosis. *J. Bacteriol.* **189,** 5903–5915 (2007).
101. Onwueme, K. C., Vos, C. J., Zurita, J., Ferreras, J. A. & Quadri, L. E. N. The dimycocerosate ester polyketide virulence factors of mycobacteria. *Prog. Lipid Res.* **44,** 259–302 (2005).
102. Lee, J. S. *et al.* Mutation in the Transcriptional Regulator PhoP Contributes to Avirulence of Mycobacterium tuberculosis H37Ra Strain. *Cell Host Microbe* **3,** 97–103 (2008).
103. Elghraoui, A., Modlin, S. J. & Valafar, F. SMRT genome assembly corrects reference errors, resolving the genetic basis of virulence in Mycobacterium tuberculosis. *BMC Genomics* **18,** 302 (2017).
104. Minnikin, D. E., Dobson, G. & Hutchinson, I. G. Characterization of phthiocerol dimycocerosates from Mycobacterium tuberculosis. *Biochim. Biophys. Acta - Lipids Lipid Metab.* **753,** 445–449 (1983).
105. Passemar, C. *et al.* Multiple deletions in the polyketide synthase gene repertoire of Mycobacterium tuberculosis reveal functional overlap of cell envelope lipids in host–pathogen interactions. *Cell. Microbiol.* **16,** 195–213 (2014).
106. Mann, F. M., Xu, M., Davenport, E. K. & Peters, R. J. Functional characterization and evolution of the isotuberculosinol operon in Mycobacterium tuberculosis and related Mycobacteria. *Front. Microbiol.* **3,** (2012).
107. Mann, F. M. *et al.* Characterization and Inhibition of a Class II Diterpene Cyclase from Mycobacterium tuberculosis. *J. Biol. Chem.* **284,** 23574–23579 (2009).
108. Mann, F. M. & Peters, R. J. Isotuberculosinol: the unusual case of an immunomodulatory diterpenoid from Mycobacterium tuberculosis. *Medchemcomm* **3,** 899–904 (2012).
109. Layre, E. *et al.* Molecular profiling of Mycobacterium tuberculosis identifies tuberculosinyl nucleoside products of the virulence-associated enzyme Rv3378c. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 2978–2983 (2014).
110. Sieniawska, E., Swatko-Ossor, M., Sawicki, R., Skalicka-Woźniak, K. & Ginalska, G. Natural Terpenes Influence the Activity of Antibiotics against Isolated Mycobacterium tuberculosis. *Med. Princ. Pract. Int. J. Kuwait Univ. Heal. Sci. Cent.* **26,** 108–112 (2017).
111. Olivella, M., Gonzalez, A., Pardo, L. & Deupi, X. Relation between sequence and structure in membrane proteins. *Bioinformatics* **29,** 1589–1592 (2013).
112. da Silva, P. E. A., Von Groll, A., Martin, A. & Palomino, J. C. Efflux as a mechanism for drug resistance in Mycobacterium tuberculosis. *FEMS Immunol. Med. Microbiol.* **63,** 1–9 (2011).
113. Abramovitch, R. B., Rohde, K. H., Hsu, F.-F. & Russell, D. G. {aprABC}: {A} {Mycobacterium} tuberculosis complex-specific locus that modulates {pH}-driven adaptation to the macrophage phagosome. *Mol. Microbiol.* **80,** 678–694 (2011).
114. Agranoff, D. & Krishna, S. Metal ion transport and regulation in {Mycobacterium} tuberculosis. *Front. Biosci. A J. Virtual Libr.* **9,** 2996–3006 (2004).
115. Kurthkoti, K. *et al.* The Capacity of Mycobacterium tuberculosis To Survive Iron Starvation Might Enable It To Persist in Iron-Deprived Microenvironments of Human Granulomas. *MBio* **8,** e01092–17 (2017).
116. Lin, W. *et al.* Transcriptional Profiling of Mycobacterium tuberculosis Exposed to In Vitro Lysosomal Stress. *Infect. Immun.* **84,** 2505–2523 (2016).
117. Zhang, Y. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins* **82 Suppl 2,** 175–187 (2014).
118. Hill, D. P. *et al.* PROGRAM DESCRIPTION: Strategies for Biological Annotation of Mammalian Systems: Implementing Gene Ontologies in Mouse Genome Informatics. *Genomics* **74,** 121–128 (2001).

bioRxiv preprint doi: https://doi.org/10.1101/358986; this version posted July 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

119. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Comput. Biol.* **5,** (2009).
120. McEvoy, C. R. E. *et al.* Comparative analysis of Mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. *PLoS One* **7,** e30593–e30593 (2012).
121. Jensen, M. R., Zweckstetter, M., Huang, J. & Blackledge, M. Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chem. Rev.* **114,** 6632–6660 (2014).
122. Ioerger, T. R. *et al.* Variation among genome sequences of H37Rv strains of Mycobacterium tuberculosis from multiple laboratories. *J. Bacteriol.* **192,** 3645–3653 (2010).
123. Vallecillo, A. J. & Espitia, C. Expression of Mycobacterium tuberculosis pe_pgrs33 is repressed during stationary phase and stress conditions, and its transcription is mediated by sigma factor A. *Microb. Pathog.* (2009). doi:10.1016/j.micpath.2008.11.003
124. J. Bacteriol.-2006-Dheenadhayalan-3721-5.
125. Golby, P. *et al.* Genome-level analyses of Mycobacterium bovis lineages reveal the role of SNPs and antisense transcription in differential gene expression. *BMC Genomics* **14,** (2013).
126. Pandey, D. P. & Gerdes, K. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res.* (2005). doi:10.1093/nar/gki201
127. Nagano, N., Orengo, C. A. & Thornton, J. M. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321,** 741–765 (2002).
128. Murphy, G. S., Greisman, J. B. & Hecht, M. H. De Novo Proteins with Life-Sustaining Functions Are Structurally Dynamic. *J. Mol. Biol.* **428,** 399–411 (2016).
129. Miskei, M. *et al.* Fuzziness enables context dependence of protein interactions. *FEBS Lett.* **591,** 2682–2695 (2017).