# BMC Systems Biology

Research article

# targetTB: A target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis

## Karthik Raman, Kalidas Yeturu and Nagasuma Chandra*

Address: Supercomputer Education and Research Centre and Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, INDIA

Email: Karthik Raman - karthik@rishi.serc.iisc.ernet.in; Kalidas Yeturu - kalidas@rishi.serc.iisc.ernet.in;
Nagasuma Chandra* - nchandra@serc.iisc.ernet.in

* Corresponding author

This article is available from: http://www.biomedcentral.com/1752-0509/2/109

## Abstract

**Background:** Tuberculosis still remains one of the largest killer infectious diseases, warranting the identification of newer targets and drugs. Identification and validation of appropriate targets for designing drugs are critical steps in drug discovery, which are at present major bottle-necks. A majority of drugs in current clinical use for many diseases have been designed without the knowledge of the targets, perhaps because standard methodologies to identify such targets in a high-throughput fashion do not really exist. With different kinds of 'omics' data that are now available, computational approaches can be powerful means of obtaining short-lists of possible targets for further experimental validation.

**Results:** We report a comprehensive *in silico* target identification pipeline, targetTB, for *Mycobacterium tuberculosis*. The pipeline incorporates a network analysis of the protein-protein interactome, a flux balance analysis of the reactome, experimentally derived phenotype essentiality data, sequence analyses and a structural assessment of targetability, using novel algorithms recently developed by us. Using flux balance analysis and network analysis, proteins critical for survival of *M. tuberculosis* are first identified, followed by comparative genomics with the host, finally incorporating a novel structural analysis of the binding sites to assess the feasibility of a protein as a target. Further analyses include correlation with expression data and non-similarity to gut flora proteins as well as 'anti-targets' in the host, leading to the identification of 451 high-confidence targets. Through phylogenetic profiling against 228 pathogen genomes, shortlisted targets have been further explored to identify broad-spectrum antibiotic targets, while also identifying those specific to tuberculosis. Targets that address mycobacterial persistence and drug resistance mechanisms are also analysed.

**Conclusion:** The pipeline developed provides rational schema for drug target identification that are likely to have high rates of success, which is expected to save enormous amounts of money, resources and time in the drug discovery process. A thorough comparison with previously suggested targets in the literature demonstrates the usefulness of the integrated approach used in our study, highlighting the importance of systems-level analyses in particular. The method has the potential to be used as a general strategy for target identification and validation and hence significantly impact most drug discovery programmes.

## Background

It is estimated that about two billion people, equalling one-third of the world's total population are infected with *M. tuberculosis* (*Mtb*) [1]. In 2006 alone, 1.7 million people died of tuberculosis (TB). TB is also the leading killer among HIV-infected people with weakened immune systems. The disease is also of particular interest to India and Asia, with more than half of all deaths occurring in Asia. Further, about 500,000 new multi-drug resistant TB cases are estimated to occur every year [1].

Currently, over 20 drugs are available for TB, of which, four of them, viz. isoniazid, rifampin, pyrazinamide and ethambutol are used as front-line drugs. Injectable drugs such as kanamycin, amikacin, capreomycin and viomycin are preferred next for treatment. Fluoroquinolones such as ciprofloxacin, ofloxacin have been found to be indispensable in the treatment of multi-drug resistant TB. Second-line bacteriostatics, such as *p*-aminosalicylic acid, ethionamide and cycloserine have established clinical efficacy but have more prominent side effects [2]. Isoniazid and ethionamide are inhibitors of mycolic acid synthesis [3,4], while cycloserine and ethambutol inhibit synthesis of peptidoglycan [5] and cell wall arabinogalactan [6,7] respectively, weakening the cell wall of the bacterium. Rifampin and Amikacin exert their pharmacological action by inhibiting bacterial RNA or protein synthesis [8-10]. As in the case of most other prescription drugs used currently, these were also discovered without the advantage of detailed molecular level information about the targets. A common strategy used in the past few decades for drug discovery involves finer structural optimisations, by starting with a lead compound that has already shown some success. Very often, this amounts to finding a newer improved drug, which modifies the function of the same target as the lead compound. This does not automatically lead to consideration of newer targets or even newer mechanisms of action. It is no surprise, therefore, that only a small fraction of the proteins in the bacterial genome have been explored as drug targets.

The existing drugs, although of immense value in controlling the disease to the extent that is being done today, have several shortcomings, the most important of them being the emergence of drug resistance rendering even the front-line drugs inactive. In addition, drugs such as rifampin have high levels of adverse effects making them prone for patient incompliance. Another important problem with most of the existing anti-mycobacterials, is their inability to act upon latent forms of the bacillus. In addition to these problems, the vicious interactions between the human immunodeficiency virus and TB have led to further challenges for anti-tubercular drug discovery [11]. For example, protease inhibitors have been shown to be incompatible with rifampin-containing anti-TB regimens

[12]. As drug discovery efforts are increasingly becoming rational and much less dependent on trial and error, identification of appropriate targets becomes a fundamental pre-requisite.

Traditionally, targets have been identified through knowledge of the function of individual protein molecules, where their function has been well-characterised. Potential targets thus identified are generally taken through a validation process involving whole-cell or animal experiments, gene knock-outs or site-directed mutagenesis that lead to loss-of-function phenotypes. Target validation is one of the critical steps in drug discovery, where a lot of time and money is spent in the pharmaceutical industry. The need for systematic and large-scale validation in the post-genomic era has led to the usage of computational methods for validation [13]. Here, we seek to apply various *in silico* techniques for the identification and validation of drug targets, specifically for *Mtb*. *In silico* methods have the advantage of speed, low cost and even more importantly, provide a systems view of the whole microbe at a time, which enables asking questions that are often difficult to address experimentally. Drug discovery has witnessed a paradigm shift from the traditional medicinal chemistry-based ligand-oriented drug discovery approaches to rational drug target identification and target-driven lead discovery, by targeting the molecular mechanisms of disease. A number of studies have been carried out by various experimental methods to identify drug targets in *Mtb* [14]. Attempts have also been made for the same purpose, based on sequence comparisons of metabolic enzymes [15], and by using various features such as Lipinski druggability at the sequence level and metabolic choke-points at the systems-level [16].

Establishing systems biology concepts and understanding the microbe as a whole opens up new opportunities for computational target identification. Here, we report a comprehensive *in silico* target identification pipeline for *Mtb*, which can also be used as a general framework for *in silico* target identification. We focus our analysis at the systems level, based on network analyses and flux balance analyses (FBA), and further validating it based on sequence analyses and structural comparisons. We have used novel algorithms for the comparison of protein structures and identifying similarity of target pockets with pockets in the human proteome, which could initiate adverse drug effects. Gene expression data have also been considered to render the analysis more comprehensive.

## Methods
### The targetTB Pipeline
A new multi-level target identification pipeline, including a novel method for structural comparison of proteins has been developed. Different *levels of abstraction* are used for

analysis, as discussed below. A summary of the several datasets used in these analyses is given in Table 1.

### Systems Analysis
*Interactome Analysis*
*System Construction*
We have constructed a protein-protein interaction network, based on the STRING database [17] version 7, which includes protein linkages between 3,925 *Mtb* proteins, inferred from published literature describing experimentally studied interactions, as well as those from genome analysis using several well-established methods such as domain fusion, phylogenetic profiling and gene neighbourhood concepts [18]. Thus, the network captures different types of interactions such as (a) physical complex formation between two proteins required to form a functional unit, (b) genes belonging to a single operon or to a common neighbourhood, (c) proteins in a given metabolic pathway and hence influenced by each other, (d) proteins whose associations are suggested based on predominant co-existence, co-expression, or domain fusion. Only the high-confidence interactions that had a STRING score of 0.7 or more were included in the network. We further augmented these with links between proteins that are influenced by the same metabolite, based on the reactions in the genome-scale metabolic reconstruction of *Mtb*, *iNJ*661 [19]. The resulting network contained 3,405 of the 3,925 proteins.

*Node Deletions*
Networks may be perturbed, through the removal of nodes and edges. A typical analysis would be to probe the effect of disrupting a node and its corresponding edges. Networks of different topologies vary in their resilience to various types of perturbations. The effect of node deletions on this network was analysed. Each of the 3,405 nodes was knocked out and the critical network parameters such as clustering coefficient and characteristic path length were monitored. In addition, the number of shortest paths that were disrupted in each deletion were monitored. The shortest paths between all pairs of proteins in the network were computed. Following removal of a node, some of these shortest paths may be disrupted, leading two pairs of nodes becoming unreachable from one another. Based on the change (loss) in the connectivity of nodes in this network and the change in network structure, on the deletion of nodes, we have delineated potential targets.

*Reactome Analysis*
Two independent genome-scale metabolic models for *Mtb* have become available. A genome-scale metabolic network, comprising 849 reactions, mediated by 739 metabolites and involving 726 genes, reported by McFadden and co-workers (GSMN-TB) [20] has been considered. Jamshidi and Palsson have reported another genome-scale metabolic model of *Mtb*, *iNJ*661, comprising 939

**Table 1: Datasets used in this study**

| Reference | Short Description |
|---|---|
| Beste *et al* (2007) [20] (GSMN-TB) | Reports GSMN-TB, a genome-scale metabolic model of *Mtb*, consisting of 849 unique reactions and 739 metabolites, and involving 726 genes. *In silico* gene deletions have been performed, using FBA, identifying genes essential for growth. |
| Jamshidi *et al* (2007) [19] (*Mtb iNJ*661) | Reports the genome-scale metabolic reconstruction of the *in silico* strain *Mtb iNJ*661, comprising 661 genes and 939 reactions. *In silico* gene deletions have been performed, using FBA, identifying genes essential for growth, as well as slow-growing mutants. |
| Raman *et al* (2005) [21] (MAP) | Reports an FBA of the mycolic acid pathway (MAP) in *Mtb*, comprising 217 reactions involving 197 metabolites and mediated by 28 proteins. *In silico* gene deletions have been performed, detailing genes essential for mycolic acid biosynthesis. |
| Sassetti *et al* (2003) [23] (TraSH) | Reports the use of transposon site hybridisation (TraSH) mutagenesis to comprehensively identify the genes required by *Mtb* for optimal growth. |
| ModBase [25] | A database of structural models of proteins from various organisms, including *Mtb* and Human, based on homology modelling |
| Gao *et al* (2005) [30] (Gao-expression) | Reports the variability in gene expression patterns among ten clinical isolates of *Mtb*, as well as the laboratory strains H37Rv and H37Ra, growing in liquid culture. |
| Rachman *et al* (2006) [31] (Rxsachman-expression) | Reports genome-wide expression analysis of *Mtb* from clinical lung samples, as well as *in vitro*. |
| Boshoff *et al* (2004) [32] (Boshoff-expression) | Reports gene expression of *Mtb* in response to several drugs/inhibitors of metabolism, as well as under persistence, starvation and different pH/media. |
| Muttucumaru *et al* (2004) [40] (Muttucumaru-expression) | Reports global gene expression in aerobic, microaerophilic and anaerobic cultures. |
| Voskuil *et al* (2004) [39] (Voskuil-expression) | Reports a genome expression profiling, analysing the adaptive mechanisms initiated by *Mtb* in two common models of *Mtb* non-proliferation. |
| Betts *et al* (2002) [38] (Betts-expression) | Reports the use of gene and protein expression profiling to identify the response of *Mtb* to nutrient starvation, also modeling persistence. |
| Hampshire *et al* (2004) [41] (Hampshire-expression) | Reports the stationary phase gene expression of *Mtb* following a progressive nutrient depletion, proposing a model for persistence. |

reactions mediated by 828 metabolites and 661 genes [19]. We have also earlier published a pathway-level model (MAP) of mycolic acid biosynthesis in *Mtb* [21]. We collated a list of lethal gene deletions for these studies. The essentiality predictions for the *iNJ*661 model were based on growth in Middlebrook 7H9 medium, as detailed in [19], while those for the GSMN-TB model were based on growth in Middlebrook 7H10 medium, as detailed in [20]. Genes whose deletion severely impaired growth (biomass formation) in the medium were designated as essential. The essentiality in [21] was studied using an objective function for optimal production of mycolates; a gene was considered essential if on deletion, most fluxes in the mycolic acid pathway including those of the mycolates dropped to zero. Using the COBRA Toolbox [22] for MATLAB, we also performed double gene deletions for the *iNJ*661 model.

### Essentiality Analysis

Information on gene essentiality from a transposon site hybridisation (TraSH) mutagenesis study for *Mtb* [23] has also been incorporated in the decision criteria.

### Sequence Analysis

Close homologues for the *Mtb* proteins in the human proteome were identified by performing a BLAST search [24]. The BLAST results were parsed using python scripts based on BioPython http://www.biopython.org/. The criteria for regarding a protein as a close homologue were a sequence similarity of greater than 50% using a BLOSUM62 matrix, for a length of more than 50% of the bacterial query protein with an E-value less than $10^{-4}$.

### Structural Assessment of Targetability

#### Obtaining Structures

Crystal structures of 229 proteins from *Mtb* and 3,515 from human are available (excluding those with greater than 70% sequence identity) from the Protein Data Bank (PDB). This translates to a mere 6% of the *Mtb* proteome and under 10% of the human proteome. However, thousands of protein structures from both host and pathogen could be obtained using theoretically calculated structural models, from the ModBase database. Models in ModBase are built on the principles of homology modelling using Modeller [25]. Models of 2,808 proteins from *Mtb* and 16,000 proteins from the human proteome were obtained from this database. The database hosts multiple models for each protein, depending on the number of confident templates available for that protein in the PDB. For this analysis, only the first model for each protein was considered. Also, only those proteins which passed the previous stages of filtering in the target identification pipeline were considered. Of the 942 *Mtb* proteins considered, only 773 had available structures in ModBase.

### Pocket Identification

In order to predict binding sites of a modelled protein, we have used PocketDepth (PD) [26], a geometry-based algorithm that has been developed and validated earlier by our group, to predict potential binding grooves on the surface of the protein. All possible binding sites in the 773 proteins of *Mtb* and the 16,000 human proteins were identified using PD. PD uses the concept of depth, which reflects how central a given pocket is and not merely how deep a subspace is in the pocket. PD outputs predicted binding sites in the form of sets or clusters. From such clusters, protein neighbourhoods within 4.0Å are extracted to obtain the binding sites.

An additional method to identify binding pockets in protein structures was used to obtain a consensus prediction. LigsiteCSC [27], a geometric method based on vectors in eight directions on a grid, also incorporating amino acid conservation information within each protein family, was used for this purpose. Top ten PD clusters were first obtained for each protein, which were compared with the top three pockets obtained from LigsiteCSC. Only the common clusters were retained for further analysis. 767 of the *Mtb* proteins and 15,830 of the human proteins were feasible for analysis, by which 3,500 pockets were identified in *Mtb* and 70,149 pockets in human.

### Pocket Comparison

The next step towards structural assessment of targetability is to compare the binding sites of shortlisted targets of *Mtb* with those of the human proteome. An algorithm developed by us very recently, PocketMatch (PM) [28], has been used for this purpose. PM is based on shape signatures encoded by 90 lists of all-pair distances of residues in the binding site, pre-classified into one of the five standard amino acid types. A similarity score is assigned to each pair of binding sites. Extensive validation for PM, using the PDBbind database [29] of experimentally determined protein-ligand complexes is reported elsewhere [28]. We have now tested the algorithm to compare predicted pockets of all proteins in PDBbind as well. The SCOP-PM comparison for predicted pockets at various thresholds is provided as supplementary material [See Additional file 1].

All the 3,500 identified sites from the 767 short-listed proteins from *Mtb* were compared with the 70,149 identified sites from 15,830 human proteins. The topmost score for every protein pair is then chosen to capture the highest similarity an *Mtb* protein has in any of its pockets with any human protein. The scores are then compared to a predefined threshold as discussed in the results section to infer similarity. The exhaustive pairwise comparison of pockets is highly computationally intensive and was carried out on a massively parallel BlueGene (configuration:

4096 2-way shared memory processor nodes: 8192 IBM PowerPC 440×5 processors operating at 700 MHz, running Linux).

### Further Analysis of Short-listed Targets

The short-listed targets were subjected to further analysis, to retain only those proteins that are highly targetable.

### Transcriptome Analysis/Gene Expression

One of the critical factors influencing the choice of a target would be its expression. Expression profiles related to persistence have been incorporated in [16]. Based on the expression of the genes, we have further filtered our list of targets. For this, we have used data from Small and co-workers [30], who have analysed the expression of genes in ten different strains of *Mtb*, *M. tuberculosis* H37Rv and *M. tuberculosis* H37Ra using cDNA microarrays. We also use data from Kaufmann and co-workers [31], who have performed a genome-wide expression analysis of *Mtb* from clinical lung samples using DNA arrays, and Barry and co-workers [32], who report an expression analysis of *Mtb* under a wide range of conditions. Lists of expressed genes have been reported in [30,31], while in [32], the *z*-scores have been reported for gene expression, in each of the experiments. A gene passed this filter if it was reported to be expressed, by either of [30,31], or in at least one of the studies (where an inhibitor of metabolism was not introduced) reported in [32].

### Comparison with 'Anti-targets'

About seven proteins have been reported to form a set of 'anti-targets' [33], viz. the human ether-à-go-go-related gene (hERG), the pregnane X receptor (PXR), constitutive androstane receptor (CAR), P-glycoprotein (P-gp), as well as membrane receptors like the adrenergic $_{1a}$, the dopaminergic D2, the serotonergic 5 – $HT_{2c}$ and the muscarinic $M_1$. Unintentional binding of drugs to these proteins causes adverse effects, leading to their labelling as anti-targets. The sequences of 306 proteins in the human proteome corresponding to these anti-targets were fetched from the NCBI sequence database. The accession numbers of these protein sequences are provided as supplementary material [see Additional file 2]. The short-listed targets were compared to these anti-targets by standard sequence analysis.

### Similarity to Gut Flora Proteins

A number of organisms are known to inhabit the gut of a normal healthy individual [34]. Inadvertent inhibition of proteins of these organisms is likely to result in side effects. In order to study this possibility, the short-listed *Mtb* proteins were compared to the proteins of the gut flora (296,017 proteins from 95 organisms), again by sequence analysis. Some of these organisms are *Bacteroides intestinalis*, *Bifidobacterium bifidum*, *Bifidobacterium longum*

and *Lactobacillus salivarius*. A full list of the 95 organisms is provided as supplementary material [see Additional file 3].

### Involvement in Persistence

*Mtb* has an unusual capacity to persist in the host at many levels. In the cellular level, it resides in macrophages that typically function to eliminate pathogens and at the systemic level, it resists clearance by the adaptive immunity of the host. Its clearance by anti-bacterials is also very slow [35]. It may be possible to address the problem of persistence by targeting those genes that are implicated in persistence. For example, isocitrate lyase is a well-known persistence factor in mice, whose disruption attenuated bacterial persistence [36]. *pcaA*, a cyclopropane synthase involved in mycolic acid biosynthesis has also been shown to be a requirement for long-term mycobacterial persistence and virulence in mice models of tubercular infection [37]. Targets that passed all the previous filters were examined for expression during persistence based on several microarray expression data [32,38-41].

### Phylogenetic Profiling

Phylogenetic profiling was carried out against 707 fully sequenced bacterial genomes. First, a BLAST was run against each of the 707 genomes, for *Mtb*. The BLAST output was then parsed using python scripts, based on BioPython, to obtain the E-value of the best hit, with a match of more than 50% of the query length, for each sequence in *Mtb*. The E-values thus obtained were converted to scores between 0 and 1, with 0 representing a strong match and 1 representing a weak match. The score was calculated as $-1/\log(E)$. Hits with $E > e^{-4}$ were all neglected and given a score of 1.0. This is identical to the scoring scheme of Protein Link EXplorer (PLEX) [42], which however currently considers only 89 genomes. For each protein in *Mtb*, profile strings comprising scores for the hits of the proteins were generated. Each profile string thus encodes the presence or absence of each of the *Mtb* proteins and where present, the extent of similarity as well. A subset of these results, for 228 pathogenic genomes, was analysed to examine the broad-spectrum nature of an identified target.

### Involvement in Drug Resistance

Proteins involved in emergence of resistance to anti-tubercular drugs have been analysed and reported by us recently [43]. The list of about 25 proteins closely connected to different pathways of resistance were obtained and used for analysis here.

## Results

A range of analyses spanning multiple levels of abstraction have been carried out, to identify plausible drug targets. The methodology can also be used more generally as

a target identification pipeline that would be applicable to many drug discovery programmes. Starting from the entire proteome of *Mtb* H37Rv comprising 3,989 proteins, we have shortlisted 451 proteins as potential drug targets using a variety of filters, as depicted in Figs. 1 and 2. Fig. 1 illustrates a pictorial view of the targetTB pipeline while Fig. 2 shows a simplified view of the pipeline as a flowchart, illustrating the flow of this study. We first carry out a network analysis, where a full genome-scale interactome encoding several types of protein-protein interactions and protein-protein influences from metabolic pathways is reconstructed. Gene deletions that would significantly disrupt the network are then identified (List-A1). Next, we have studied the reactome through FBA (List-A2), to identify lethal gene deletions. This is further augmented with high-throughput gene essentiality data (List-A3). These system-level analyses together comprise Filter A. This is then integrated with sequence-level (Filter B) and structural analyses (Filter C) as described below (see Fig. 1). The expression of the gene encoding for the target is highly desirable (Filter E) and the list is further pruned by eliminating targets with high similarities to known 'anti-targets' in the human proteome (Filter F) and proteins in gut flora (Filter G). Those targets known to contribute to drug resistance in the pathogen are then prioritised. By analysis of similarity against several pathogenic proteomes, broad-spectrum targets as well as those unique to *Mtb* have also been identified. Various filters, lists and the numbers of proteins passed and eliminated at the various stages of the pipeline are given in Table 2.

### Systems Analysis
#### Interactome Analysis
A protein-protein interaction network comprising 3,405 nodes and 29,302 edges was constructed, which covered over 85% of the *Mtb* proteome. To evaluate the importance of a given protein in the context of the large interactome network, each node was individually deleted and its impact measured in terms of the number of shortest paths that are disrupted. Shortest paths in a network are quite critical to the structure of the network. Shortest paths in metabolic networks of *Mtb* and *M. leprae* have been identified and analysed by us earlier [44]. Samson and co-workers have earlier analysed a protein network in *Saccharomyces cerevisiae*, indicating that the analysis of shortest paths may provide an idea of network navigability as well as the efficiency with which a perturbation can spread throughout a network [45]. More recently, Wingender and co-workers have illustrated the importance of a similar metric, a 'pairwise disconnectivity index', for topological analysis of regulatory networks [46]. The disruption of shortest paths is expected to have a substantial effect on the network connectivity in protein networks as well. In the interactome network studied here, most of the node

deletions do not significantly disrupt network connectivity. However, substantial effects (more than 5,000 disrupted shortest paths) were observed upon deletion of 431 of the 3,405 nodes (List-A1). These 431 proteins, for which a critical role in maintaining interactome network structure is suggested, were taken through further steps of filtering, in order to identify most useful drug targets. For example, for BirA (Rv3279c), close to 95,000 shortest paths in the network, were disrupted by its removal. A complete list of these proteins is provided as supplementary material [See Additional file 4].

#### Reactome Analysis
An FBA study, using the *iNJ*661 model [19], identified 188 proteins of the 661 studied, as essential for the growth of the bacterium, whereas an additional 41 also had a significant impact on growth (the *in silico* knock-out mutants were slow growers) [19]. A separate FBA study using an independently derived genome-scale metabolic model (GSMN-TB) identified 259 of the 719 proteins studied as essential for growth [20]. While these two models are similar in many respects, there are subtle differences in their biomass functions for FBA, as well as their coverage of the *Mtb* proteomes. 134 proteins were common to both lists of essential proteins. A third FBA study (MAP), carried out by us previously for the mycolic acid pathway alone identified 15 proteins in the pathway as essential for the microbe. Put together, the three studies suggest 318 proteins to be essential for the microbe. A critical role in maintaining the metabolism of the bacterium is suggested for the 318 proteins (List-A2). We have also carried out a double knockout study, on the *Mtb iNJ*661 model, identifying 49 pairs of genes, which when knocked out together, produce a lethal phenotype.

#### Essentiality Analysis
A high-throughput analysis of gene essentiality, using Transposon Site Hybridisation (TraSH) mutagenesis has been reported earlier. Genes, whose deletion produced slow-growing mutants, were also identified. These proteins (List-A3), taken together with Lists A1 and A2, form the list of proteins (List A) that are implicated to be essential, by systems-level analyses. We have combined the essentiality data, rather than take a consensus from the different system-level models discussed above, since each model has its own strengths and weaknesses. Many proteins are eliminated from the pipeline at this stage. For example, MabA (Rv1483), which has been suggested as a potential drug target [47,48], was not found to be essential in any of the systems-level analyses. MshA (Rv0486), suggested as an essential component of mycothiol biosynthesis and essential for growth in *Mtb* Erdman strain [49], is also not found to be essential in any of the systems-level studies.

**Table 2: Models and methods used in the targetTB pipeline**

| Analysis | Coverage | N/A | 4 | X | | Computation |
|---|---|---|---|---|---|---|
| **(A)** Systems Analysis | | | | | | |
| (i) Network | 3405 | | 431 | 2974 | | Single node deletions |
| (ii) FBA | | | | | | |
| a) *Mtb iNJ*661 | 661 | | 229 | 432 | | FBA |
| b) GSMN-TB | 717 | | 259 | 458 | | FBA |
| c) MAP | 26 | | 15 | 11 | | FBA |
| (iii) TraSH | 3186 | | 656 | 2530 | | |
| *Summary* ((i) or (ii) or (iii)) | 3823 | 166 (A') | 1138 | 2685 | (A$_X$) | |
| **(B)** Sequence Analysis | 3989 | | 3611 | 378 | (B$_X$) | 3,989 *Mtb* vs. 33,453 human sequence comparisons |
| **(A&B)** Systems & Sequence Passed | 3989 | - | 942 | 3047 | (A$_X$ ∪ B$_X$) | |
| **(C)** Structural Assessment (PMScore < 0:8) [942 from A&B only considered] | 767 | 175 (C') | 622 | 145 | (C$_X$) | 3,500 sites of *Mtb* (767 proteins) vs 70,149 sites of Human pocketome (15,830 proteins) = 245,521,500 pairwise comparisons |
| **(D)** A & B & C | 3989 | - | 622 | 3552 | (A$_X$ ∪ B$_X$ ∪ C$_X$ ∪ C') | |
| **(E)** Expression | | | | | | |
| (i) Gao-expression | 3590 | 399 | 2210 | 1380 | | |
| (ii) Rachman-expression | 634 | 3355 | 634 | - | | |
| (iii) Boshoff-expression | 3915 | 74 | 3915 | - | | |
| *Summary* ((i) or (ii) or (iii)) | 3917 | 72 | 3264 | 653 | (E$_X$) | |
| *Summary* (for (D)) | 622 | 1 | 529 | 92 | | |
| **(F)** Non-similarity to Anti-targets | 3989 | - | 3928 | 61 | (F$_X$) | 306 vs. 3,989 sequence comparisons |
| Non-similarity to Anti-targets (for (D)) | 622 | - | 611 | 11 | | |
| **(G)** Non-similarity to gut flora | 3989 | - | 3730 | 259 | (G$_X$) | 296,017 vs. 3,989 sequence comparisons |
| Non-similarity to gut flora (for (D)) | 622 | - | 543 | 79 | | |
| **(H)** D & E & F &; G | 622 | 1 | **451** | 170 | | |
| **(I)** Expression during Persistence | | | | | | |
| (i) Muttucumaru-expression | 3924 | 82 | 639 | 3268 | | |
| (ii) Boshoff-expression | 3915 | 74 | 105 | 3810 | | |
| (iii) Betts-expression | 3649 | 340 | 274 | 3375 | | |
| (iv) Voskuil-expression | 3924 | 82 | 310 | 3597 | | |
| (v) Hampshire-expression | 3757 | 395 | 1349 | 2245 | | |
| *Summary* ((i) or (ii) or (iii) or (iv) or (v)) | 3933 | 73 | 1871 | 2045 | | |
| *Summary* (for (H)) | 451 | - | **216** | 235 | | |
| **(J)** Broad-spectrum Targets [present in > 100/228 pathogenic genomes] | 451 | - | **186** | 265 | | 2,295,901 vs. 3,989 sequence comparisons (phylogenetic profiling against 707 genomes) |
| **(K)** Targets unique to Mycobacteria | 451 | - | **66** | 385 | | |

Number of proteins covered in each study are indicated. A '4' indicates that it passes the filter, while a '**X**' indicates failure. The method used at each step is also indicated. The number of targets identified in the final lists H, I, J and K are boxed and indicated in bold typeface. The symbols A', C', A$_X$-G$_X$ are as described in the text and in Fig. 1.
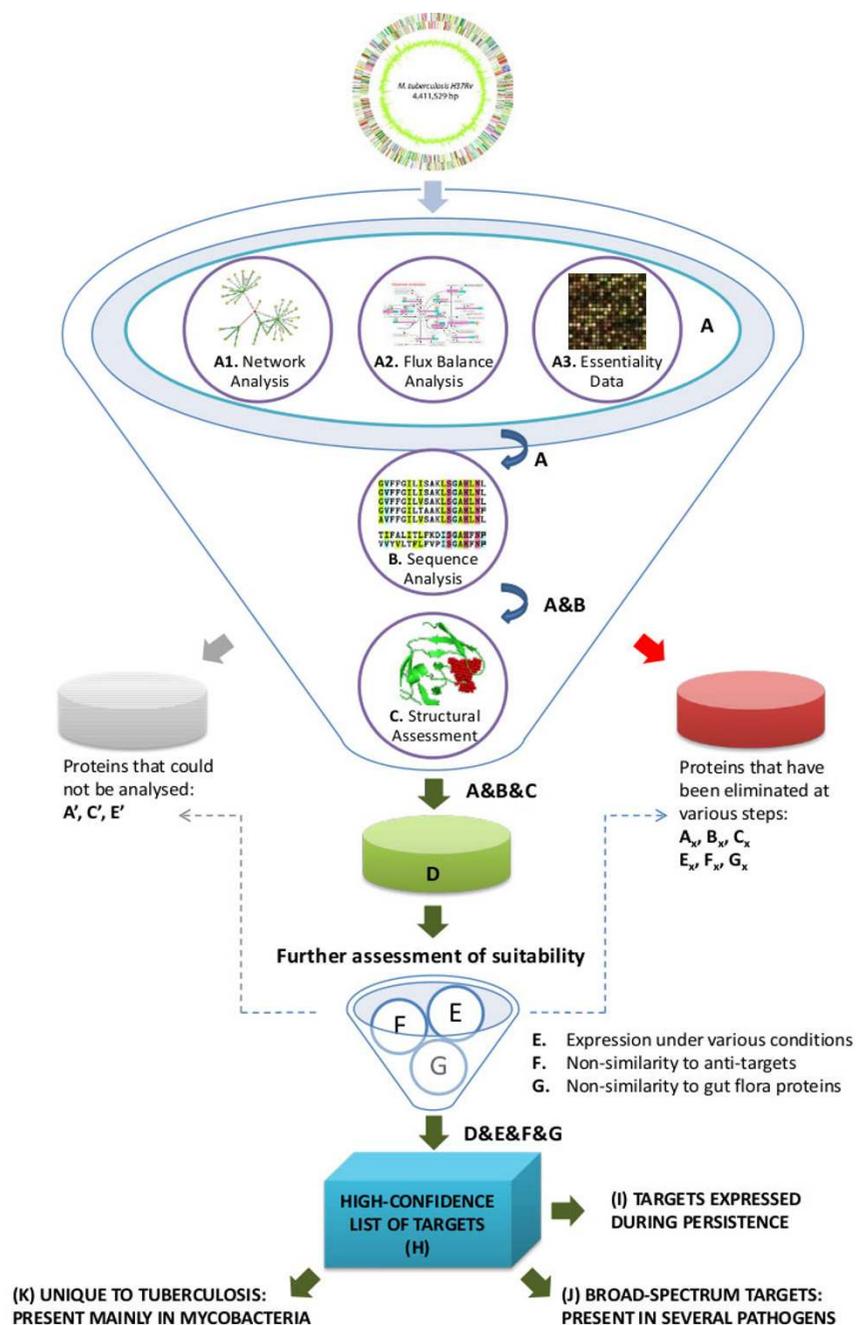
**Figure 1**
**The targetTB Target Identification Pipeline**. The funnel depicts the order in which the entire proteome of *Mtb* is considered and analysed at different layers. 'A' refers to the systems level studies, which includes A1, for network analysis of the interactome; A2, for flux balance analyses of the reactome; and A3, for genome-scale essentiality data determined experimentally as reported by Sassetti *et al* [23]. Those proteins that passed these filters are indicated as 'A', and combined with the results of sequence analysis (B), to derive those that passed both filters (depicted as 'A&B'). These were then taken through Filter C, referring to the structural assessment filter, yielding the list of 622 proteins as the D-List (A&B&C). Further steps of filtering are indicated in the smaller funnel as E (expression under various conditions), F (non-similarity to anti-targets) and G (non-similarity to gut flora proteins). Those proteins that pass all the six levels of filtering (indicated as D&E&F&G) form the H-List comprising 451 targets. Additional filters I, J and K used for analysing the H-List are also indicated. Lists A', C' and E' refer to the set of proteins at A, C and E levels, respectively, that could not be analysed for lack of appropriate data. Lists $A_X$, $B_X$, $C_X$, $E_X$, $F_X$ and $G_X$ refer to sets of proteins that failed in that particular filter, but may have passed at other levels.

### Sequence Analysis

At the sequence level, comparison with the human (host) proteome can be useful in filtering out those targets that have detectable homologues in the human cells, in order to reduce the risk of adverse effects that arise due to unintended interaction of the drug with the host protein. For 3,611 of 3,989 *Mtb* proteins, no close homologues were observed in the human proteome. The remaining 378 proteins, for which close homologues were observed, were eliminated at this step. The 3,611 proteins (List-B) were taken through further steps in the targetTB pipeline. Proteins such as KasA (Rv2245), KasB (Rv2246), MabA (Rv1483), RmlB (Rv3464), which have been suggested as potential targets in earlier studies, have all been eliminated at this stage, due to the presence of close homologues in the human proteome.

Combining the systems and sequence level analyses, 942 proteins were shortlisted for further analysis. A list of these proteins is presented as supplementary material [See Additional file 4].

### Structural Assessment of Targetability

Similarity between proteins is better captured through structural comparisons, where structural data for both proteins are available. In fact, what ultimately matters in determining the pharmacological profiles of drug molecules is the recognition of the drug molecules by various protein molecules at their binding sites. It is therefore important to compare binding sites in the various protein molecules in both the pathogen and the host. At this step, we want to critically weed out targets that share very high similarity with binding sites from the human 'pocketome', since targeting these may lead to adverse drug reactions, due to inadvertent binding with human proteins.

This type of analysis would become more meaningful if carried out at the proteome-scale. Advances in crystallography and various structural genomics projects [50-52] have led to the determination of 229 and 3,515 structures of *Mtb* and human, respectively. In the absence of experimentally determined structures, high-confidence homol-
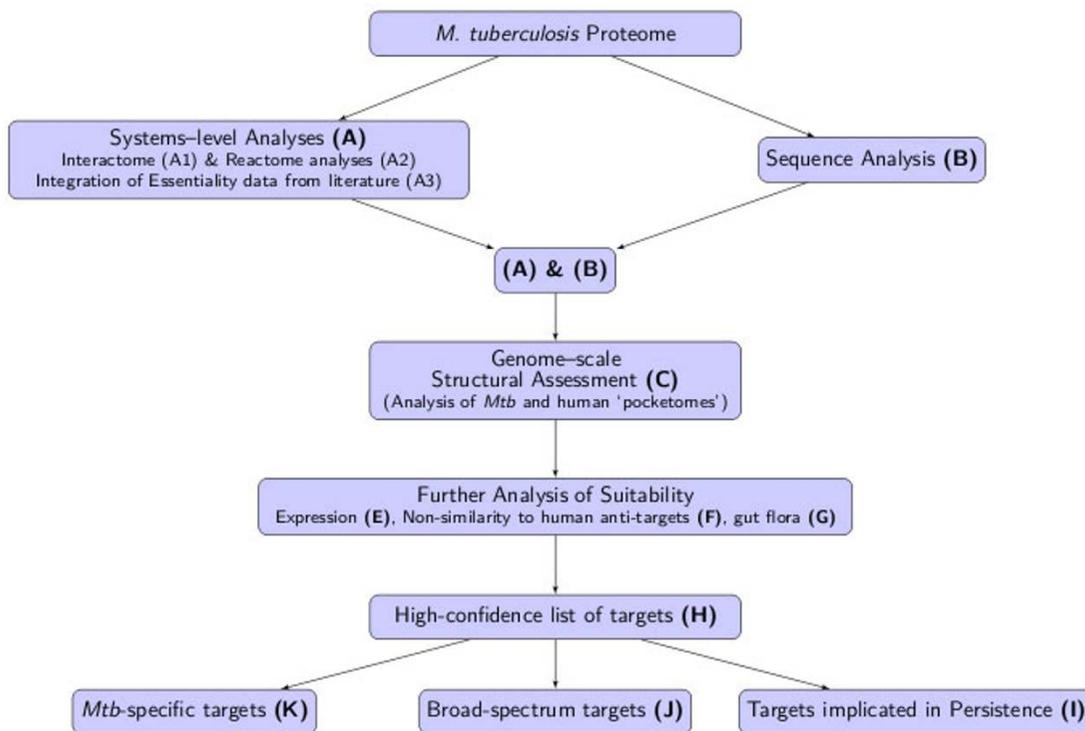


**Figure 2**
**Flowchart illustrating the sequence of analyses in this study**. This flowchart provides a simplified view of the various filters used in this study, in the order in which they are applied, to arrive at the final lists of targets.

ogy models for 2,808 *Mtb* proteins and 16,000 human proteins were obtained from the ModBase database. The availability of such a large number of protein structures in both species makes it feasible to carry out a proteome-scale structural assessment of targetability. Identification of binding sites and further comparison of the identified binding sites are the next two challenging steps towards this goal. Two new algorithms that we have recently developed, PD and PM, enable us to carry out this comparison.

Of the 942 proteins shortlisted earlier in the pipeline, 773 had structures available in the PDB/ModBase databases. For these 773 proteins, the top 10 binding sites for each protein, identified using PD were compared with the top three binding pockets from LigsiteCSC. LigsiteCSC considers amino acid conservation at the putative sites, in the family of proteins. This automatically leads to identifying residues and hence the sites that are likely to be functionally important. Finding a consensus among top predictions between the two methods increases confidence in site prediction significantly. Some proteins such as DesA3 (Rv3229c), EmbB (Rv3795) and AccE5 (Rv3281) passed all other tests, but were not included in the H-List of high-confidence targets, since the structural analysis could not be performed.

A consensus between PD and LigsiteCSC was obtained so as to identify the most probable pockets that also contained conserved amino acid residues at the binding sites. Using this, 3,500 pockets were identified for 767 of the *Mtb* proteins. A similar exercise carried out for the human proteins identified 70,149 pockets. An all-versus-all comparison of the 'pocketomes' of *Mtb* and human was performed, using PM. This translated to 245,521,500 pairwise comparisons, which corresponded to over three years of serial CPU time, that was successfully completed on a BlueGene System, within a week.

A PM score of 0.8 or more indicates high similarity between two binding pockets. This threshold was used as a filter to eliminate all those proteins in *Mtb* whose pockets closely matched with any pocket of any protein in the human proteome. Of the 767 proteins, 145 had closely matching pockets in the human proteomes and were therefore eliminated from the pipeline. It is possible that some of these *Mtb* proteins contain some pockets that are sufficiently different from pockets of human proteins. Such proteins may also be targetable, but would require a close and more detailed analysis of all the pockets in the protein. The remaining 622 form a list of targets for anti-tubercular drugs. These proteins were taken through further steps of filtering to produce lists of highly viable targets.

Thus, of the 767 proteins that passed the A and B filters described above and had available structures, only 622 of

them were found to pass this filter. This is despite the fact that sequence filtering was already carried out, re-emphasising the need for a multi-level target identification and validation scheme. The resulting proteins form the D-List, of targets that can be further explored for TB drug discovery.

### Further Analysis of Short-listed Targets
While the fundamental determinants of the quality of a target have already been considered earlier, the following aspects are also of importance in selecting a quality target for drug design. The following filters were therefore used to further prune the identified list and in some cases to enrich the list with targets having additional benefits.

#### Gene Expression
It is obvious that a target would be desirable only if it is expressed in the organism, at least under disease conditions. Expression data is available for over 3,900 genes in *Mtb* from various studies [30-32]. Of the shortlisted targets in the D-List, 529 are expressed, indicating their high viability as suitable targets. It must be noted here that the expression data are not comprehensive, especially in terms of the conditions that have been tested. The expression filter, while useful in understanding what is expressed and hence what is a useful target, should not be used to rule out otherwise useful targets. Until availability of more comprehensive data, this step is best used at the post-identification analysis stage. For example, proteins such as TrpD (Rv2192c), AroA (Rv3227), RibC (Rv1412) do not appear to be expressed in any of the experiments considered.

#### Comparison with Anti-targets
An ideal target should not only have specific recognition to the drug directed against it, but should also be sufficiently different from the host proteins, which have been termed as anti-targets. Considering this aspect early in the drug discovery pipeline may prove to be very useful in minimising the risk of failure of the drug candidates in the later stages of drug discovery. Anti-targets include proteins such as the transporters and pumps, which modify the bio-availability of a drug by their efflux action, or those proteins that trigger hazardous side effects, such as the hERG protein, which when blocked causes the 'sudden death syndrome' [33]. This list is by no means complete, but has been included here, more from a conceptual perspective, to highlight the need for screening against anti-targets. Sequence comparisons against 306 sequences belonging to the eight categories of anti-targets carried out revealed that sequence homologues at a similarity of 30% for over 30% of the query length were observed for 11 of the targets from the D-List. Such a loose similarity measure is used, since it is desired to rule out even a remote similarity with any anti-target. Moreover, close homo-

logues have already been eliminated by sequence analysis earlier. A structural analysis of the proteins, when more data become available would be of immense utility in this regard. Serine/Threonine protein kinases such as the PknB (Rv0014c), earlier proposed as a target [53], PknL(Rv2176) and PknH (Rv1266c), as well as cytochromes such as Cyp128 (Rv2268c) and Cyp132 (Rv1394c) were eliminated at this stage.

### Similarity to Gut Flora Proteins

The targets from the D-List were further compared to the protein sequences of hundreds of organisms that inhabit the gut of a healthy human. This was carried out to prune the list of identified drug targets, so that the drugs administered do not bind unintentionally to the proteins of the gut flora. Unintentional inhibition of gut flora proteins are known to lead to adverse effects and can promote pathogenic colonisation of the gut [54]. Drug interactions with gut flora are also believed to be the cause of idiosyncratic drug toxicity and reduced bio-availability of the drug [55,56]. Similarity of the identified targets to such proteins therefore affects their suitability. The sequence analyses carried out here indicate that 79 proteins from the D-List had close homologues in the gut flora and were hence removed from the list of most viable targets. For example, FtsZ (Rv2150c), Glf (Rv3809c) have homologues in gut flora and were hence eliminated at this stage. Interestingly, Icl (Rv0467), which has been particularly suggested as an attractive drug target [57] and also implicated in persistence [36], fails at this stage, due the presence of homologues in gut flora.

At this stage of filtering, from the 622 targets in the D-List identified earlier, 163 have been eliminated, leaving behind a high-confidence list of 451 targets (H-List). Several known targets appear in this list. A comprehensive analysis of the passage of several known targets in the targetTB pipeline has been performed. Some of these targets are indicated in Table 3, while the complete list is available as supplementary material [See Additional File 5].

### Involvement in Persistence

The expression of targets in the H-List, under conditions of persistence were analysed, from a set of microarray data. 216 of the H-List targets were up-regulated two-fold or more in at least one of the studies considered. These 216 targets form the I-List of targets, which may be useful in combating persistent *Mtb* infection. Some examples of proteins in the I-List are DesA1 (Rv0824c), DesA2 (Rv1094), DevS (Rv3132c), FadD32 (Rv3801c), KatG (Rv1908c), Pks13 (Rv3800c), CysH (Rv2392) and Wag31 (Rv2145c). CysH has also previously been shown to be important for *Mtb* persistence [58].

### Identification of Broad-spectrum vs. Mtb-specific targets

Phylogenetic profiling of *Mtb* proteins against various genomes gives a measure of the uniqueness of a particular target to the *Mtb* proteome. Phylogenetic profiling can also help in identifying important functional linkages of chosen targets. It is also useful for identifying targets that can be used for designing broad-spectrum anti-bacterials. The 451 shortlisted targets were compared with 228 pathogenic bacterial genomes (provided as supplementary material [See Additional file 6]). If the *Mtb* target has close homologues in more than 100 genomes, we refer to it as a possible broad-spectrum anti-bacterial target (J-List). Several proteins involved in lipid metabolism are present in this list, viz. InhA (Rv1484), FabH (Rv0533c), FabD (Rv2243), PcaA (Rv0470c) and the MmaA's 1–4. IspF (Rv3581), which has been suggested as an attractive target in many pathogens [59], is also in the J-List. A main concern of such a strategy to target a multitude of bacteria in clinical therapy is the emergence of resistance to multiple organisms, which is highly undesirable. However, if the emergence of resistance is countered, as discussed below, having broad-spectrum targets could be of great advantage.

Proteins that were present only in mycobacteria were also identified by this analysis (K-List). This list is rich in mycobacteria PPE proteins and also contains proteins such as DevS, a sensor histidine kinase involved in a two-component signal transduction pathway.

### Involvement in Drug Resistance

In a recent study, we identified possible pathways that would be involved in the emergence of drug resistance in *Mtb* [43]. We also proposed the concept of 'co-targets', referring to those proteins, which when inhibited simultaneously with a corresponding primary target, will help in reducing the emergence of resistance to the drug binding to that primary target. The importance of any protein in the H-List identified here will significantly increase if it also happens to be a constituent of the resistance pathways. These pathways comprise proteins that are predicted to be either directly responsible for generating resistance to the given drug, or serve as an important hub in the flow of information from the target of the given drug to the machinery of resistance. Proteins in the resistance pathways broadly belong to one of the four mechanisms, which are mediated by cytochromes, SOS related genes, antibiotic efflux pumps and genes involved in horizontal gene transfer (HGT). The putative targets in the H-List were analysed for their proximity to resistance-related proteins in the protein-protein interaction network described in [43]. Of 451 proteins in the H-List, 25 were closely involved in the resistance pathways and would therefore be significantly more useful as drug targets.

**Table 3: Results for known and proposed targets in the targetTB pipeline**

| Target | Remarks | targetTB pipeline | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A | B | C | E | F | G | H | I | J | K |
| **I. Cell Wall Biosynthesis** | | | | | | | | | | | |
| DdlA (Rv2981c) | Known target of cycloserine [5] | 4 | 4 | X | 4 | 4 | 4 | X | | | |
| EmbA (Rv3794) | Known target for ethambutol [7,75] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | | |
| AftA (Rv3792) | Suggested as an attractive target [76] | 4 | 4 | X | 4 | 4 | 4 | X | | | |
| AftB (Rv3805c) | Suggested as a potential target [77] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | | |
| MurG (Rv2153c) | Suggested as a potential target [78] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | 1 | |
| **II. Lipid Metabolism** | | | | | | | | | | | |
| FabH (Rv0533c) | Possible target of thiolactomycin; also suggested as potential target [79,80] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | 1 | |
| FabD (Rv2243) | Suggested as a potential target [81-83] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | 1 | |
| AcpM (Rv2244) | Induced on isoniazid treatment [81,84] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | | |
| Pks13 (Rv3800c) | Suggested as a promising target against *Corynebacterineae* [85] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | | |
| InhA (Rv1484) | Known target for isoniazid, ethionamide [4] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | 1 | |
| PcaA (Rv0470c) | Suggested as a possible target of thiacetazone [86] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | 1 | |
| MmaA1 (Rv0645c) | -do- | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | 1 | |
| FadD32 (Rv3801c) | Suggested as a promising target [87] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | 1 | |
| DesA3 (Rv3229c) | Suggested as a possible target [88] | 4 | 4 | ? | 4 | 4 | 4 | X | | | |
| Fas (Rv2524c) | Possible target of pyrazinamide [89] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | | |
| **III. Intermediary Metabolism and Respiration** | | | | | | | | | | | |
| LysA (Rv1293) | Lysine auxotroph has vaccine potential [90]; suggested as potential target [91] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | 1 | |
| TrpD (Rv2192c) | -do- | 4 | 4 | 4 | X | 4 | 4 | X | | | |
| LeuA (Rv3710) | Suggested as potential target [92] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | | |
| DapB (Rv2773c) | Suggested as potential target [93] | 4 | 4 | 4 | X | 4 | X | X | | | |
| AroB (Rv2538c) | Shikimate pathway suggested as an attractive target [94] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | 1 | |
| ArgA (Rv2747) | Essential enzyme catalysing initial step of arginine biosynthesis [95] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | | |
| AlrA (Rv3423c) | Known target of Cycloserine [5] | 4 | 4 | X | 4 | 4 | 4 | X | | | |
| DfrA (Rv2763c) | Important drug target in many pathogens [96]. Suggested as drug target in [96,97] | 4 | X | - | 4 | 4 | 4 | X | | | |
| PanB (Rv2225) | Critical for pantothenic acid synthesis [98] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | 1 | |
| PanC (Rv3602c) | Critical for pantothenic acid synthesis [98]; suggested as potential target [99] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | 1 | |
| PanD (Rv3601c) | Critical for pantothenic acid synthesis [98]; suggested as potential target [100] | 4 | 4 | X | 4 | 4 | 4 | X | | | |
| PanK (Rv1092c)/CoaA | Prokaryotic enzymes involved in the synthesis of CoA are good targets [101]; [102] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | | |
| CysH (Rv2392) | Suggested as an attractive drug target [103-106]; CysH is important for *Mtb* protein during latent infection [58] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | | |
| IspD (Rv3582c) | Potential drug target [107] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | 1 | |
| IspF (Rv3581c) | Potential drug target [107]; attractive target in many pathogens [59] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | 1 | |
| Icl (Rv0467) | Required for persistence of *Mtb* in macrophages and mice [36]; suggested as an attractive target [57]. Icl1 and Icl2 are required for fatty acid catabolism and virulence in *Mtb* [108] | 4 | 4 | 4 | 4 | 4 | X | X | | | |
| AtpE1 (Rv1305) | Inhibited by a diarylquinoline drug R207910 *in vitro* [109] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | | |
| Cyp121 (Rv2276) | Putative essential gene. Possible role in virulence through studies with ΔAraC/XylS gene regulator mutant (ΔRv1931c) [110]. Induced in isoniazid- and thiolactomycin-treated *Mtb* [111] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | | |
| **IV. Information Pathways** | | | | | | | | | | | |
| GyrA (Rv0006) | Known target of uoroquinolones [112,113] | X | 4 | - | 4 | 4 | X | X | | | |
| GyrB (Rv0005) | -do- | 4 | 4 | X | 4 | 4 | X | X | | | |
| RpoB (Rv0667) | Known target of rifampicin [8] | 4 | 4 | 4 | 4 | 4 | X | X | | | |
| RpsL (Rv0683) | Known target of Streptomycin [114] | 4 | X | - | 4 | 4 | X | X | | | |
| **V. Regulatory proteins** | | | | | | | | | | | |

**Table 3: Results for known and proposed targets in the targetTB pipeline** *(Continued)*

| | | A | B | C | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GlnE (Rv2221c) | Essential for growth of *Mtb* [115] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | | | |
| MtrA (Rv3246c) | Essential for growth of *Mtb* [116] | X | 4 | - | 4 | 4 | 4 | X | | | |
| DevR (Rv3133c) | Two-component system is a novel target in dormant mycobacteria [117]; essential for growth of *Mtb* under conditions of low oxygen [118] | 4 | 4 | 4 | X | 4 | 4 | X | | | |
| DevS (Rv3132c) | Two-component system is a novel target in dormant mycobacteria [117]; part of the DevR-DevS two-component signal transduction system [118,119] | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | | 1 |
| PknB (Rv0014c) | Possibly essential for mycobacterial growth and hence possible target [53] | 4 | 4 | 4 | 4 | X | 4 | X | | | |
| PknG (Rv0410c) | Crucial virulence factor [120]; possibly essential for mycobacterial growth and hence possible targets [53] | 4 | 4 | X | X | 4 | 4 | X | | | |
| MbtA (Rv2384) | An important adenylation enzyme required for siderophore biosynthesis [121] | 4 | 4 | X | 4 | 4 | 4 | X | | | |
| IdeR (Rv2711) | Suggested as target [122,123] | X | 4 | - | 4 | 4 | X | X | | | |

An account of the passage of known targets (previously reported in literature) through the targetTB pipeline. The putative targets are classified based on their broad functional categories. A, B, C, E, F, G, H, I, J and K refer to the different filters depicted in Fig. 1 and described in the text. '4' indicates that the given protein passes the filter, while a '**X**' indicates a failure. A '**?**' indicates that the analysis was not performed due to lack of appropriate data, while '-' indicates that the protein was not passed through the filter due to failure at a previous stage. All proteins in the H-List indicated in Fig. 3 would have a '4' at levels A-H. 'l' indicates the additional lists (I/J/K) in which a target from the H-List is present.

Some notable examples are PolA (Rv1629), a protein involved in the SOS response, a cytochrome Cyp121 (Rv2276), which is also connected to 19 other cytochromes, and SecY (Rv0732), a protein connected to DnaE1 (SOS) and two other proteins, SecA1 and SecA2, implicated in HGT. Table 4 gives a list of these proteins and their association with resistance related proteins.

***Targets Identified by the targetTB Pipeline***
The various filters and the corresponding analyses that have been applied in this study, to arrive at the final lists of targets are listed in Table 2. Of the 3,989 proteins that have been annotated in the *Mtb* genome, 622 proteins pass the filters of systems and sequence analyses, as well as the structural assessment (D-List). These proteins are then screened to eliminate those which are not expressed, as well as those which have homologues in gut flora, or with anti-targets in the human proteome. A final list of 451 proteins is arrived at, which comprise the H-List. Of these, 216 proteins satisfy persistence criteria (I-List), while 186 are potential broad-spectrum anti-bacterial targets (J-List), and 66 targets are unique to mycobacteria (K-List). Proteins for which the analysis could not be performed, due to lack of available data at this time are separated as lists A' and C', which may be considered for analysis once more data become available. Proteins that have been eliminated at various stages could still find use as drug targets under different scenarios. For example, those proteins eliminated due to non-essentiality to *Mtb* ($A_X$-List) may contain pairs of proteins that could together be essential and may hence be useful, if targeted concurrently. In fact, the double knock-out studies using FBA carried out here clear demonstrate this aspect. Similarly, proteins that have been eliminated due to some structural similarity with human targets ($C_X$-List) may be useful as drug targets if the structural differences between the host and pathogen proteins could be exploited.

The functional classes of the 451 targets (H-List) identified by this study are indicated in Fig. 3. The list is also available as supplementary material [See Additional file 4]. This list includes several known targets and many that have been proposed as potential targets. Some known targets have been eliminated because they have failed one or more filters in the targetTB pipeline. The passage of known and proposed targets for anti-tubercular drugs in the targetTB pipeline is detailed in Table 3 (also see Additional File 5). Some examples of proteins that are in the H-List include known targets such as InhA, EmbA and FabH, as well as many targets that have been proposed for anti-tubercular drug discovery, such as GlfT2, a bi-functional UDP-galactofuranosyl transferase, the fatty acid synthase Fas, the pantothenate kinase PanK, a glutamine-synthetase adenylyltransferase GlnE and the sensor histidine kinase DevS. The list also indicates several proteins that have been suggested as potential drug targets in literature, but eliminated from the targetTB pipeline on account of failing one or more of the filters.

It is interesting to note that of the 451 targets in the H-List, over a half of them belong to the functional classes of 'lipid metabolism' and 'intermediary metabolism and respiration'. It has been said that metabolism has often not been given sufficient importance in 'intelligent' drug design [60]. Our analysis is in support of that observation, highlighting several targets from lipid metabolism, particularly the critical pathway of mycolic acid biosynthesis, amino acid biosynthesis, menaquinone biosynthesis and mycothiol biosynthesis. Several of the metabolites produced in these pathways are essential for mycobacterial survival and hence, the pathways producing these metabolites are ideal candidates for anti-tubercular drug discovery. Many of these pathways do not have equivalent pathways in the human, making them even more suitable candidates for targeting.

**Table 4: Targets in the H-List that are also involved in drug resistance mechanisms.**

**Resistance related proteins**

| Protein | Description |
|---------|-------------|
| CcdA (Rv0527) | Cytochrome |
| PolA (Rv1629) | SOS |
| LldD2 (Rv1872c) | Cytochrome |
| QcrC (Rv2194) | Cytochrome |
| QcrB (Rv2196) | Cytochrome |
| CtaC (Rv2200c) | Cytochrome |
| Cyp121 (Rv2276) | Cytochrome |
| Rv3660c (Rv3660c) | HGT |

**Proteins closely connected to resistance proteins**

| Protein | Closely connected to |
|---------|----------------------|
| SecY (Rv0732) | DnaE1 (SOS) and SecA1 and SecA2 (HGT) |
| Rv0843 (Rv0843) | Three cytochrome proteins |
| Pdc (Rv0853c) | Rv1988 (Antibiotic Efflux Pump) and two cytochromes |
| Rv1456c (Rv1456c) | Three cytochrome proteins |
| Rv1711 (Rv1711) | RecA, DnaE1 (SOS proteins) and a cytochrome |
| Rv1828 (Rv1828) | SecA2 (HGT) and two cytochromes |
| QcrC (Rv2194) | Four cytochrome proteins |
| QcrA (Rv2195) | Five cytochrome proteins |
| QcrB (Rv2196) | Five cytochrome proteins |
| CtaC (Rv2200c) | Five cytochrome proteins |
| Cyp121 (Rv2276) | 19 cytochrome proteins |
| HemE (Rv2678c) | Seven cytochrome proteins |
| FtsK (Rv2748c) | SOS proteins RecA, PolA and DnaE1 |
| RnhB (Rv2902c) | SOS proteins PolA, DnaE1 and DnaE2 |
| TrmD (Rv2906c) | SOS proteins PolA, DnaE1 and Rv2294 (Antibiotic Efflux Pump) |
| PrfB (Rv3105c) | PolA (SOS) and SecA1 and SecA2 (HGT) |
| IlvX (Rv3509c) | Three cytochrome proteins |
| TrxB2 (Rv3913) | RecA (SOS) and two cytochrome proteins |

The top panel shows targets that are directly implicated in resistance mechanisms. The lower panel indicates targets which are immediately connected to proteins involved in the emergence of resistance in the interactome. These proteins are predicted to be involved in mediating the flow of information from the targets to the resistance machinery.

Desaturases DesA1 and DesA2, which have been shown by us to be hallmarks of the mycolic acid biosynthesis pathway in *Mtb* [61], pass all the filters and are present in the H-List. They are also present in the I-List of targets expressed during persistence. These proteins thus appear to be highly viable targets for anti-tubercular drugs. AcpS (Rv2523c), an acyl-carrier-protein synthase involved in mycolic acid biosynthesis, also passes all the filters and is a potential target. TrxB2 (Rv3913), a probable thioredoxin reductase and LysA (Rv1293), a diaminopimelate decarboxylase which catalyses the conversion of diaminopimelic acid to lysine and ThrB (Rv1296), a probable homoserine kinase, which are also ranked very high (ranked two, four and six, respectively) in the metabolic list of prioritised targets reported by Schreiber and co-workers [16], are also targets of interest.

### Comparison with Earlier Computational Studies

Two computational studies, outlining strategies for target identification, particularly for anti-tubercular drugs, have been reported earlier [15,16]. We present an overview of the passage of the targets suggested in these studies in the targetTB pipeline, also outlining the advantages of the targetTB pipeline over the previously reported methods.

### Anishetty et al (2005) [15]

Based on a sequence analysis study, comparing enzymes in metabolic pathways between human and *Mtb*, Pennathur and co-workers proposed 186 proteins as suitable drug targets. Of these, 51 feature in our H-List and 129 do not, while six could not be considered for lack of sufficient functional data. Some examples of the 51 targets featuring in the H-List are AcpS (Rv2523c), AtpC (Rv1311), FabH (Rv0533c), FbpA (Rv3804c), FolB (Rv3607c), IspE (Rv1011), KatG (Rv1908c), LeuA (Rv3710), MenC (Rv0553), PanB (Rv2225), PanC (Rv3602c), PpdK (Rv1127c), GlfT1 (Rv3782) and TrpA (Rv1613). An account of how each of the 180 proteins proposed as targets in the study reported by Anishetty *et al* (2005) fare in

**CELL WALL AND CELL WALL PROCESSES (57)**

CysT (Rv2399c), CysW (Rv2398c), DrrB (Rv2937), DrrC (Rv2938), EmbA (Rv3794), EmbC (Rv3793), EsxG (Rv0287), FtsK (Rv2748c), GlfT (Rv3808c), GlnH (Rv0411c), LepB (Rv2903c), LppL (Rv2138), Lpql (Rv0237), LpqK (Rv0399c), MmpL8 (Rv3823c), ModA (Rv1857), MurE (Rv2158c), MurF (Rv2157c), MurG (Rv2153c), MurX (Rv2156c), NagA (Rv3332), ParB (Rv3917c), PitA (Rv0545c), PitB (Rv2281), Rip (Rv2869c), RpfB (Rv1009), Rv0102, Rv0284, Rv0290, Rv0479c, Rv0585c, Rv1303, Rv1435c, Rv1456c, Rv1459c, Rv1473, Rv1565c, Rv1795, Rv2113, Rv2136c, Rv2969c, Rv3238c, Rv3448, Rv3604c, Rv3635, Rv3679, Rv3737, GlfT1 (Rv3782), Rv3802c, AftB (Rv3805c), Rv3882c, Rv3887c, SecY (Rv0732), SppA (Rv0724), SugB (Rv1237), Wag31 (Rv2145c), WbbL1 (Rv3265c)

**CONSERVED HYPOTHETICALS (77)**

ArgA (Rv2747), Pra (Rv1078), Rv0060, Rv0163, Rv0269c, Rv0295c, Rv0525, Rv0540, Rv0552, Rv0635, Rv0811c, Rv0813c, Rv0926c, Rv1186c, Rv1204c, Rv1208, Rv1215c, Rv1232c, Rv1277, Rv1287, Rv1339, Rv1421, Rv1461, Rv1462, Rv1466, Rv1505c, Rv1558, Rv1592c, Rv1711, Rv1732c, Rv1738, Rv1828, Rv1836c, Rv1847, Rv1870c, Rv1873, Rv1885c, Rv1897c, Rv2026c, Rv2117, Rv2165c, Rv2166c, Rv2170, Rv2372c, Rv2510c, Rv2515c, Rv2554c, Rv2573, Rv2681, Rv2683, Rv2689c, Rv2787, Rv2837c, Rv2842c, Rv2897c, Rv2910c, Rv2949c, Rv3015c, Rv3031, Rv3037c, Rv3131, Rv3282, Rv3284, Rv3338, Rv3404c, Rv3471c, Rv3627c, Rv3722c, Rv3740c, Rv3755c, Rv3796, Rv3868, Rv3871, Rv3909, Rv2140c, Rv2623

**LIPID METABOLISM (48)**

AccD3 (Rv0904c), AcpM (Rv2244), AcpS (Rv2523c), CmaA2 (Rv0503c), DesA1 (Rv0824c), DesA2 (Rv1094), FabD (Rv2243), FabH (Rv0533c), FadA2 (Rv0243), FadD11 (Rv1550), FadD16 (Rv0852), FadD28 (Rv2941), FadD30 (Rv0404), FadD32 (Rv3801c), FadD36 (Rv1193), FadE24 (Rv3139), FadE33 (Rv3564), FadH (Rv1175c), Fas (Rv2524c), FbpA (Rv3804c), FbpB (Rv1886c), FbpC (Rv0129c), InhA (Rv1484), MbtC (Rv2382c), MmaA1 (Rv0645c), MmaA2 (Rv0644c), MmaA3 (Rv0643c), MmaA4 (Rv0642c), MutA (Rv1492), PapA3 (Rv1182), PapA5 (Rv2939), PcaA (Rv0470c), Php (Rv0230c), PimA (Rv2610c), PimB (Rv0557), Pks1 (Rv2946c), Pks13 (Rv3800c), Pks17 (Rv1663), Pks2 (Rv3825c), Pks4 (Rv1181), Pls81 (Rv1551), PlsB2 (Rv2482c), PpsA (Rv2931), PpsC (Rv2933), PpsE (Rv2935), Rv0914c, Rv1722, Rv2613c

**PE/PPE (14)**

PE_PGRS63 (Rv3097c), PE11 (Rv1169c), PE5 (Rv0285), PE6 (Rv0335c), PPE12 (Rv0755c), PPE2 (Rv0256c), PPE22 (Rv1705c), PPE23 (Rv1706c), PPE24 (Rv1753c), PPE3 (Rv0280), PPE4 (Rv0286), PPE46 (Rv3018c), PPE50 (Rv3135), PPE54 (Rv3343c)

**INTERMEDIARY METABOLISM AND RESPIRATION (187)**

Aao (Rv1905c), Adh (Rv1530), AdhA (Rv1862), AdoK (Rv2202c), AmiB2 (Rv1263), ArcA (Rv1001), ArgC (Rv1652), ArgD (Rv1655), ArgJ (Rv1653), AroB (Rv2538c), AroD (Rv2537c), AroE (Rv2552c), AroF (Rv2540c), AroG (Rv2178c), AroK (Rv2539c), AtpB (Rv1304), AtpC (Rv1311), AtpE (Rv1305), AtpH (Rv1307), BioB (Rv1589), BioF2 (Rv0032), CanA (Rv1284), Cbs (Rv1077), CcdA (Rv0527), CoaA (Rv1092c), CoaE (Rv1631), CobC (Rv2231c), Csd (Rv1464), CtaC (Rv2200c), Cyp121 (Rv2276), Cyp128 (Rv2268c), Cyp132 (Rv1394c), CysD (Rv1285), CysE (Rv2335), CysH (Rv2392), CysK2 (Rv0848), DapA (Rv2753c), DapF (Rv2726c), DeoC (Rv0478), Dfp (Rv1391), Fba (Rv0363c), FbiA (Rv3261), FdhF (Rv2900c), FolB (Rv3607c), GalK (Rv0620), GalTb (Rv0619), GcvT (Rv2211c), GlbN (Rv1542c), GlgP (Rv1328), GlnA3 (Rv1878), GlnA4 (Rv2860c), GlnE (Rv2221c), GlpQ1 (Rv3842c), GltA1 (Rv1131), GltB (Rv3859c), GltD (Rv3858c), GshA (Rv3704c), GuaB1 (Rv1843c), GuaB3 (Rv3410c), HemA (Rv0509), HemD (Rv0511), HemE (Rv2678c), HemK (Rv1300), HemN (Rv2388c), HemZ (Rv1485), HisC1 (Rv1600), HycP (Rv0085), IlvX (Rv3509c), IspD (Rv3582c), IspE (Rv1011), IspF (Rv3581c), Lat (Rv3290c), LeuA (Rv3710), LipC (Rv0220), LipT (Rv2045c), LldD2 (Rv1872c), LpdA (Rv3303c), LysA (Rv1293), ManA (Rv3255c), MenC (Rv0553), MenD (Rv0555), MenE (Rv0542c), MenH (Rv0558), MetH (Rv2124c), MiaA (Rv2727c), MoeA2 (Rv0438c), Mrp (Rv1229c), MrsA (Rv3441c), MscR (Rv2259), MshB (Rv1170), MshD (Rv0819), Mtn (Rv0091), MycP1 (Rv3883c), MycP2 (Rv3886c), MycP3 (Rv0291), NadA (Rv1594), Nat (Rv3566c), NirA (Rv2391), PabB (Rv1005c), PanB (Rv2225), PanC (Rv3602c), Pcp (Rv0319), Pdc (Rv0853c), PepQ (Rv2535c), PfkB (Rv2029c), PheA (Rv3838c), PpdK (Rv1127c), PpgK (Rv2702), PpnK (Rv1695), PrcA (Rv2109c), PrcB (Rv2110c), ProB (Rv2439c), PtrBa (Rv0781), PurE (Rv3275c), PurK (Rv3276c), PyrB (Rv1380), PyrD (Rv2139), PyrE (Rv0382c), PyrF (Rv1385), QcrA (Rv2195), QcrB (Rv2196), QcrC (Rv2194), RibA1 (Rv1940), RibG (Rv1409), RibH (Rv1416), RmlC (Rv3465), RmlD (Rv3266c), Rv0208c, Rv0224c, Rv0228, Rv0296c, Rv0315, Rv0316, Rv0338c, Rv0526, Rv0636, Rv0648, Rv0697, Rv0843, DapC (Rv0858c), Rv1096, Rv1178, Rv1201c, Rv1220c, Rv1257c, Rv1264, Rv1360, Rv1465, Rv1500, Rv1692, Rv1714, Rv1856c, Rv2296, Rv2435c, Rv2715, Rv2952, Rv2957, Rv2959c, Rv2962c, Rv3137, Rv3175, Rv3406, Rv3579c, Rv3712, Rv3752c, Rv3778c, Rv3790, Rv3791, Rv3915, SerA1 (Rv2996c), SerB2 (Rv3042c), SpeE (Rv2601), SseA (Rv3283), SseB (Rv2291), ThiD (Rv0422c), ThiL (Rv2977c), ThiO (Rv0415), ThrA (Rv1294), ThrB (Rv1296), ThrC (Rv1295), ThyX (Rv2754c), Tmk (Rv3247c), TrpA (Rv1613), TrpC (Rv1611), TrpG (Rv0013), TrxB2 (Rv3913), UreC (Rv1850)

**INFORMATION PATHWAYS (37)**

AlaS (Rv2555c), Def (Rv0429c), DnaB (Rv0058), DnaG (Rv2343c), DnaZX (Rv3721c), FusA2 (Rv0120c), GltS (Rv2992c), HelZ (Rv2101), Hup8 (Rv2986c), LigC (Rv3731), NrdZ (Rv0570), NusB (Rv2533c), PheT (Rv1650), PolA (Rv1629), PpiB (Rv2582), PrfB (Rv3105c), RadA (Rv3585), Rho (Rv1297), Rnc (Rv2925c), RnhB (Rv2902c), RnpA (Rv3923c), RplL (Rv0652), RplV (Rv0706), RplW (Rv0703), RpsD (Rv3458c), RpsQ (Rv0710), RpsT (Rv2412), SigA (Rv2703), SigC (Rv2069), SigE (Rv1221), SigF (Rv3286c), SigH (Rv3223c), SigM (Rv3911), TopA (Rv3646c), TrmD (Rv2906c), Tsf (Rv2889c), TyrS (Rv1689)

**VIRULENCE, DETOXIFICATION, ADAPTATION (12)**

BpoC (Rv0554), DnaJ1 (Rv0352), DnaJ2 (Rv2373c), GrpE (Rv0351), HrcA (Rv2374c), KatG (Rv1908c), OtsB2 (Rv3372), Rv0627, Rv1477, Rv3177, Rv3660c, TreS (Rv0126)

**REGULATORY PROTEINS (21)**

DevS (Rv3132c), KdpE (Rv1027c), MprB (Rv0982), MtrB (Rv3245c), NadR (Rv0212c), NarL (Rv0844c), PknB (Rv0014c), PknH (Rv0465c), PknL (Rv2176), PstP (Rv0018c), RraA (Rv3853), Rv0196, Rv0465c, Rv0472c, Rv0474, Rv0586, Rv0894, Rv1846c, Rv3220c, Rv3405c, TcrA (Rv0602c)

**CONSERVED HYPOTHETICALS WITH ORTHOLOGUE IN *M. bovis* (4)**

Rv2319c, Rv3611, Rv3818, Rv3849

**INSERTION SEQUENCES AND PHAGES (2)**

Rv1701, Rv3427c

**Figure 3**
**List of Identified Targets**. Distribution of the functional classes of the 451 targets identified in the H-List. The number of targets present in each of the functional classes is also indicated.

the targetTB pipeline is given as supplementary material [See Additional file 7].

Of the 129 targets that do not pass the filters used in our study, but were predicted by Anishetty *et al*, 77 have been eliminated due to their non-essentiality in *Mtb*, as predicted by systems-level analyses, clearly demonstrating the need for incorporating systems-level studies. Of the remaining 52, one had a close homologue in the human proteome and 16 had a PM score of 0.8 or more, leading to their elimination. Of the remaining 35, 14 are not expressed under any of the conditions considered by the experiments considered (studies [30-32]), while 18 of them had homologues in gut flora (five failing both expression and gut flora filters). For the remaining eight, structural assessment through PD-LigsiteCSC-PM was infeasible due to lack of availability of an appropriate model. These observations reiterate the need for a comprehensive multi-level analysis for target identification, as demonstrated by the targetTB pipeline.

*Hasan et al (2006) [16]*
Schreiber and co-workers have reported a study in which they prioritise all proteins in the *Mtb* genome for use as drug targets. Their ranking is based on a consideration of metabolic choke-points, *in vitro* essentiality for growth and druggability as judged by sequence similarity to proteins capable of binding small molecule ligands, besides sequence analysis to identify unique proteins. Some concepts are similar between our study and that of Hasan *et*

*al*, but our study differs from theirs in a number of ways: (i) to start with, the goal in our study is to identify a very high quality list of drug targets that are also computationally validated, whereas Hasan *et al* have aimed to prioritise all proteins in *Mtb* for their feasibility as drug targets (ii) a pipeline has been developed that filters out proteins at every stage, leading to a final list of very high quality targets at the same time eliminating the need for a blind consideration of all proteins at all stages. The pipeline is also useful for considering proteins eliminated at different steps, if required, with necessary caution. (iii) a rigorous FBA and network analysis have been carried out in our study, making the systems-level analysis much more comprehensive (iv) a comprehensive structural assessment of 767 proteins of *Mtb* that passed other filters in the pipeline, against 15,830 different human proteins, has been carried out. New algorithms developed by us have been used to identify and compare pockets, again rendering the structural analysis efficient and more importantly feasible, since it considers only the relevant features that describe drug recognition. In addition, we have considered (v) elimination of proteins similar to anti-targets and also (vi) those important in countering the emergence of drug resistance.

Hasan *et al* have proposed three lists of prioritised targets, based on different scoring schemes. In the metabolic list proposed by Hasan *et al*, 146 of the targets from the H-List are present in the top 500. Of the rest, 82 were eliminated due to the presence of sequence homologues in the

human proteome. 107 were non-essential by systems analysis, while for eight, no data was available. Of the remaining 154, 43 were not feasible for structural analysis, while 49 had a PM Score of 0.8 or more. Two of the proteins had similarities with human anti-targets. Of the remaining 62, 36 had homologues in gut flora and 32 were not expressed (6 failed both filters). As a result, the final list of proteins that we have identified (H-List) differs significantly from those proposed by Hasan *et al*. A report of how the top 500 targets in each of the three lists proposed by Hasan *et al* (2006) fare in the targetTB pipeline is given as supplementary material [See Additional file 8].

## Discussion

It is now well-established that better insights into biological systems may be obtained by considering large-scale system-level models, since biological systems are complex networks of many processes. The conventional method of focussing on a single protein at a time, however important the protein may be, would mean losing perspective of its larger context and hence may not provide the right answers, especially in drug discovery. Broader insights about the appropriateness of a potential target can be obtained by considering pathways and whole-system models relevant to that disease. For example, an enzyme that may be identified as a good target for a particular disease may not actually be critical or essential, when viewed in the context of the entire metabolism in the cell. Analysing system-level models can help in assessing criticality of the individual proteins by studying any alternate pathways and mechanisms that may naturally exist to compensate for the absence of that protein. This study has demonstrated how systems biology can be used in drug target identification and drug discovery.

As the necessity of systems-level studies is becoming more and more obvious, a wide spectrum of techniques have been developed and applied for the simulation and analysis of biochemical systems [62-65]. These include stoichiometric techniques that rely on reaction stoichiometry and other constraints, kinetic pathway modelling techniques that rely on comprehensive mechanistic models and interaction-based analyses, as well as Petri nets and qualitative modelling formalisms [66]. The FBA carried out in conjunction with gene knock-outs here indicates the criticality of individual reactions and hence the associated proteins. In FBA, knock-outs can in fact be viewed as extreme inhibitions in which the target is totally inhibited by a drug. 188 of the 661 proteins in *Mtb iNJ*661 model resulted in lethal phenotypes when knocked out, indicating their essentiality for producing the required biomass and hence for bacterial growth. The FBA analysis also has the potential to consider multiple knock-outs again amounting to total inhibition at multiple points. Such a phenomenon is known to occur by some drugs individu-

ally and more commonly by a cocktail of drugs. For example, isoniazid is thought to act at two points in the pathway by inhibiting both InhA and KasA [4,67]. The FBA study presents a ready framework to analyse the effects of such drug inhibitions, which would be extremely difficult to judge by inspection of the reaction maps alone. Various combinations of the non-lethal gene deletions leading to about 111,628 different double knock-outs were generated and tested with FBA using the same objective function. 49 of them were found to lead to lethal phenotypes, with growth ratio of zero, as compared to that of the wild-type. Such proteins can be targeted simultaneously to achieve excellent antibacterial effect, although individually either one of them would not be good targets. Some examples of such pairs are Rv0505c (SerB1, non-essential)-Rv3042c (SerB2, in H-List), both phosphoserine phosphatases, Rv2243 (FabD, H-List)-Rv0649 (FabD2, non-essential), both malonyl CoA-ACP transacylases, Rv3273–Rv3588c, both carbonic anhydrases, and non-essential, individually, by systems analyses. It is conceivable that each of these pairs that appear to be isozymes produce a lethal phenotype on deletion, since the functional step of the pathways they catalyse may have proceeded in the absence of one, but would be arrested in the absence of both enzymes. Another example is that of Rv0363c (Fba, a fructose-bisphosphate aldolase)-Rv1237 (SugB, a sugar transport membrane protein ABC transporter). Such studies using FBA, however, can be carried out only for the annotated reactome component of the bacterial cell.

Networks obtained by considering various protein-protein interactions and influences, on the other hand are much more comprehensive and nearly complete in their coverage, especially because of the availability of an integrated database that considers experimentally mapped interactions and those predicted from one or more of the four well-established computational methods [17,18]. A drawback of such a network however, could be a large number of false positives. To minimise the introduction of false positives, we have eliminated all low-confidence interactions from our study. The number of broken paths introduced by a knock-out is taken here as a measure of the essentiality of the protein in maintaining the network. Biological networks typically display a power-law degree distribution. We explore the importance of the disruption of network connectivity that occurs on account of attacking nodes that lie on many shortest paths in the network. The advantage of interaction-based modelling such as this is that it is possible to generate interaction networks from existing databases and it is not constrained by lack of quantitative mechanistic data.

Besides essentiality to the pathogen, an ideal target should have several other properties such as non-similarity with

human proteins whose inhibition could lead to potential adverse drug effects, an aspect that has been analysed at multiple levels in this study (see Fig. 1). The simplest level of course is to check for sequence similarity of the target being queried with all the proteins in the human proteome. Sequence information is readily available for hundreds of bacteria and this type of analysis is reported earlier for pathogenic genomes such as *Burkholderia pseudomallei* [68], *Helicobacteri pylori* [69], *Pseudomonas aeruginosa* [70,71] and even *Mtb* [72]. However, such sequence filtering while important, cannot be the sole criteria for identifying high quality targets, since two proteins that are considerably dissimilar in their sequences could have very similar binding sites [73,74]. Thus, while sequence similarity very often leads to structural and hence functional similarity, it is not a necessary condition for two proteins to have similar ligand binding profiles.

In the process of target identification, what really matters for a good target is to have a binding site in the target protein that is sufficiently different from that of any host protein. This is so that a given drug is both available in intended quantities to the intended target and perhaps more importantly, to avoid adverse effects by the drug binding to another protein from the host and manipulating its function as well, which is unintended and unanticipated. For this purpose, it is not very intuitive to look at structural classes and overall properties such as the structural family or secondary structural types, that might describe a structure. Instead, it is important to study the possible binding profile of a given drug to all those proteins to which it is likely to be exposed. Towards this goal, we first identified possible pockets in the set of *Mtb* and human structures, using PD, a validated algorithm that was recently developed in our lab. All such putative pockets were tested for certain criteria such as size and volume, retaining only those that were likely to bind to small molecules. The filtered pockets from preliminarily shortlisted targets from *Mtb* were then screened for similarity against pockets from the human proteins, which involved over 245 million comparisons, using PM, a site-matching algorithm recently developed in our laboratory. From this, 145 putative targets were eliminated due to high similarity with one or more human proteins. Interestingly, well-known molecules such as AlrA, PanD and GyrB are observed to have high similarities with proteins in the human, perhaps explaining the side effects caused by the drugs targeting them. With a cut-off in PMScore of 60%, molecules such as InhA, EmbA and EmbC, would all have been eliminated from the list for not having the properties of a safe target. However, since it is in principle, possible to design inhibitors that could bind only to the intended target by exploiting subtle structural differences that exist at the sites of the bacterial target in question with those of the human proteins obtained as hits with PM, we chose to

use a high cut-off of 80%, so as to remove only those with very high risk of causing side effects. Some examples of molecules that have failed at this stage are DdlA, GyrB, AftA and AlrA. It must be noted that some of these were ranked as high priority targets by other studies that did not consider the structural aspect explicitly, again emphasising the need for structural level analysis. Eliminating those proteins with high similarity to proteins in the gut flora also helps in ultimately reducing the risk of side effects.

The last stages of filtering and post-identification analysis resulted in identifying two categories of targets: broad-spectrum targets and *Mtb*-specific targets. It is necessary to identify targets in both the categories, since they are required in different situations. *Mtb*-specific targets are believed to be safer since they would not lead to many organisms developing resistance against the drugs of such targets. Broad-spectrum targets, on the other hand, would be extremely useful when multiple infections co-exist or in some cases where a specific diagnosis is not possible. A comprehensive phylogenetic analysis of the shortlisted targets against 228 different pathogenic genomes has been carried out in this study, leading to the identification of broad-spectrum targets. Identification of pathways and proteins involved in generating drug resistance and then targeting them simultaneously as co-targets along with the primary broad-spectrum targets would reduce the risk of drug resistance significantly, making many more molecules accessible for therapeutic intervention.

## Conclusion

In summary, network analysis of the interactome in *Mtb* and flux balance analysis of the reactome, both systems-level studies, have helped in identifying a set of proteins critically required for the survival of the bacterium. By mapping these with experimentally determined essentiality data, a set of proteins that would be useful as drug targets is identified. The list is pruned by a series of filters to eliminate all those with a risk of causing side effects. Traditionally, drug safety has been addressed by modification of the drug molecule itself, but this paper reports how a careful choice of the target molecule can be made to achieve that goal, which could be used as a general strategy right in the beginning of the drug discovery process. To our knowledge, this is also the first study to carry out a comprehensive structural level analysis of identifying binding pockets and matching them so as to obtain a possible pharmacodynamic map of the administered drugs. In addition to the sequence and structure level filters, the final list of targets identified has also passed filters put in place to eliminate those similar to known anti-targets and the gut flora proteins. Finally, the list is further enriched by considering possible mechanisms in emergence of drug resistance. The pipeline developed provides rational

schema for drug target identification that are likely to have high rates of success, which should save enormous amounts of money, resources and time in the drug discovery process.

## Abbreviations
FBA: Flux Balance Analysis; HGT: Horizontal Gene Transfer; *Mtb*: *Mycobacterium tuberculosis*; PD: PocketDepth; PM: PocketMatch; TB: tuberculosis.

## Authors' contributions
NC generated the idea and closely supervised the project. KR performed most of the work in this study, especially that of systems and sequence-level analyses. YK performed the structural comparisons and validations. KR and NC wrote the manuscript and all authors read and approved the final manuscript.

## Additional material

### Additional file 1
***Validation of PocketMatch for predicted pockets.*** *Figure illustrating the variation of the XOR of PocketMatch matrix vs. SCOP matrix, for various PM thresholds.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-109-S1.pdf]

### Additional file 2
***Anti-target sequences.*** *Accession numbers of the 306 anti-target sequences considered.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-109-S2.txt]

### Additional file 3
***List of gut flora.*** *List of 95 organisms present in gut flora, as retrieved from the NCBI database.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-109-S3.txt]

### Additional file 4
***Lists of proteins passing various filters.*** *Detailed lists of proteins passing various filters. Also includes the final lists H, I, J and K.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-109-S4.xls]

### Additional file 5
***Passage of known and proposed targets in the targetTB pipeline.*** *An account of the passage of known targets (previously reported in literature) through the targetTB pipeline. The putative targets are classified based on their broad functional categories.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-109-S5.pdf]

### Additional file 6
***List of pathogenic genomes.*** *List of 228 pathogenic genomes considered.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-109-S6.txt]

### Additional file 7
***Comparison with Anishetty* et al *(2005).*** *A report of how the proteins proposed as targets in the study reported by Anishetty et al (2005) fare in the targetTB pipeline.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-109-S7.xls]

### Additional file 8
***Comparison with Hasan* et al *(2005).*** *A report of how the top 500 targets in each of the three lists proposed by Hasan et al (2006) fare in the targetTB pipeline.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-109-S8.xls]

## References
1.  World Health Organisation: *Global tuberculosis control: surveillance, planning, financing : WHO report 2008* World Health Organisation; 2008.
2.  Janin YL: **Antituberculosis drugs: ten years of research.** *Bioorg Med Chem* 2007, **15(7):**2479-2513.
3.  Lei B, Wei CJ, Tu SC: **Action mechanism of antitubercular isoniazid. Activation by *Mycobacterium tuberculosis* KatG, isolation, and characterization of *inhA* inhibitor.** *The Journal of Biological Chemistry* 2000, **275(4):**2520-6.
4.  Banerjee A, Dubnau E, Quémard A, Balasubramanian V, Um KS, Wilson T, Collins D, de Lisle G, Jacobs WR Jr: ***inhA*, a gene encoding a target for isoniazid and ethionamide in *Mycobacterium tuberculosis*.** *Science* 1994, **263:**227-230.
5.  Feng Z, Barletta RG: **Roles of *Mycobacterium smegmatis* D-alanine:D-alanine ligase and D-alanine racemase in the mechanisms of action of and resistance to the peptidoglycan inhibitor D-cycloserine.** *Antimicrob Agents Chemother* 2003, **47(1):**283-291.
6.  Deng L, Mikusova K, Robuck KG, Scherman M, Brennan PJ, McNeil MR: **Recognition of multiple effects of ethambutol on metabolism of mycobacterial cell envelope.** *Antimicrob Agents Chemother* 1995, **39(3):**694-701.
7.  Belanger AE, Besra GS, Ford ME, Mikusova K, Belisle JT, Brennan PJ, Inamine JM: **The *embAB* genes of *Mycobacterium avium* encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol.** *Proc Natl Acad Sci U S A* 1996, **93(21):**11919-11924.
8.  Telenti A, Imboden P, Marchesi F, Matter L, Schopfer K, Bodmer T, Lowrie D, Colston MJ, Cole ST: **Detection of rifampicin-resistance mutations in *Mycobacterium tuberculosis*.** *Lancet* 1993, **341(8846):**647-650.
9.  Busscher GF, Rutjes FPJT, van Delft FL: **2-Deoxystreptamine: Central Scaffold of Aminoglycoside Antibiotics.** *Chemical Reviews* 2005, **105(3):**775-792.

10. Maus CE, Plikaytis BB, Shinnick TM: **Molecular Analysis of Cross-Resistance to Capreomycin, Kanamycin, Amikacin, and Viomycin in *Mycobacterium tuberculosis*.** *Antimicrob Agents Chemother* 2005, **49(8):**3192-3197.

11. Nunn P, Williams B, Floyd K, Dye C, Elzinga G, Raviglione M: **Tuberculosis control in the era of HIV.** *Nat Rev Immunol* 2005, **5(10):**819-826.

12. Bonora S, Di Perri G: **Interactions between antiretroviral agents and those used to treat tuberculosis: Clinical pharmacology of antiretroviral drugs.** *Current Opinion in HIV & AIDS* 2008, **3:**306-312.

13. Raman K, Kalidas Y, Chandra N: *Biological Database Modeling, Artech House 2007 chap* Model Driven Drug Discovery: Principles and Practices:163-188.

14. Mdluli K, Spigelman M: **Novel targets for tuberculosis drug discovery.** *Curr Opin Pharmacol* 2006, **6(5):**459-467.

15. Anishetty S, Pulimi M, Pennathur G: **Potential drug targets in *Mycobacterium tuberculosis* through metabolic pathway analysis.** *Comput Biol Chem* 2005, **29(5):**368-378.

16. Hasan S, Daugelat S, Rao PS, Schreiber M: **Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*.** *PLoS Comput Biol* 2006, **2(6):**e61.

17. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P: **STRING 7-recent developments in the integration and prediction of protein interactions.** *Nucleic Acids Res* 2007, **35(Database issue):**D358-D362.

18. Strong M, Graeber TG, Beeby M, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps.** *Nucleic Acids Research* 2003, **31(24):**7099-7109.

19. Jamshidi N, Palsson BØ: **Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the *in silico* strain iNJ661 and proposing alternative drug targets.** *BMC Syst Biol* 2007, **1:**26.

20. Beste DJV, Hooper T, Stewart G, Bonde B, Avignone-Rossa C, Bushell ME, Wheeler P, Klamt S, Kierzek AM, McFadden J: **GSMN-TB: a web-based genome-scale network model of *Mycobacterium tuberculosis* metabolism.** *Genome Biol* 2007, **8:**R89.

21. Raman K, Rajagopalan P, Chandra N: **Flux Balance Analysis of Mycolic Acid Pathway: Targets for Anti-tubercular Drugs.** *PLoS Computational Biology* 2005, **1(5):**e46.

22. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox.** *Nat Protoc* 2007, **2(3):**727-738.

23. Sassetti CM, Boyd DH, Rubin EJ: **Genes required for mycobacterial growth defined by high density mutagenesis.** *Molecular Microbiology* 2003, **48:**77-84.

24. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997, **25(17):**3389-3402.

25. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A: **MODBASE: a database of annotated comparative protein structure models and associated resources.** *Nucleic Acids Res* 2006, **34(Database issue):**D291-D295.

26. Kalidas Y, Chandra N: **PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins.** *J Struct Biol* 2008, **161:**31-42.

27. Huang B, Schroeder M: **LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.** *BMC Struct Biol* 2006, **6:**19.

28. Yeturu K, Chandra N: **PocketMatch: A new algorithm to compare binding sites in protein structures.** *BMC Bioinformatics* 2008, **9(1):**543.

29. Wang R, Fang X, Lu Y, Yang CY, Wang S: **The PDBbind database: methodologies and updates.** *J Med Chem* 2005, **48(12):**4111-4119.

30. Gao Q, Kripke KE, Saldanha AJ, Yan W, Holmes S, Small PM: **Gene expression diversity among *Mycobacterium tuberculosis* clinical isolates.** *Microbiology* 2005, **151(Pt 1):**5-14.

31. Rachman H, Strong M, Ulrichs T, Grode L, Schuchhardt J, Mollenkopf H, Kosmiadi GA, Eisenberg D, Kaufmann SHE: **Unique transcrip-tome signature of *Mycobacterium tuberculosis* in pulmonary tuberculosis.** *Infect Immun* 2006, **74(2):**1233-1242.

32. Boshoff HI, Myers TG, Copp BR, McNeil MR, Wilson M, Barry CE III: **The transcriptional responses of *Mycobacterium tuberculosis* to inhibitors of metabolism: novel insights into drug mechanisms of action.** *The Journal of Biological Chemistry* 2004, **279(38):**40174-40184.

33. Recanatini M, Bottegoni G, Cavalli A: *In silico* antitarget screening. *Drug Discovery Today: Technologies* 2004, **1(3):**209-215.

34. Savage DC: **Microbial ecology of the gastrointestinal tract.** *Annu Rev Microbiol* 1977, **31:**107-133.

35. Gomez JE, McKinney JD: **M. tuberculosis persistence, latency, and drug tolerance.** *Tuberculosis* 2004, **84(1–2):**29-44.

36. McKinney JD, zu Bentrup KH, Muñoz-Elías EJ, Miczak A, Chen B, Chan WT, Swenson D, Sacchettini JC, Jacobs WR Jr, Russell DG: **Persistence of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase.** *Nature* 2000, **406:**735-738.

37. Glickman MS, Cox JS, Jacobs WR Jr: **A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of *Mycobacterium tuberculosis*.** *Mol Cell* 2000, **5(4):**717-727.

38. Betts JC, Lukey PT, Robb LC, McAdam RA, Duncan K: **Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling.** *Mol Microbiol* 2002, **43(3):**717-731.

39. Voskuil MI, Visconti KC, Schoolnik GK: *Mycobacterium tuberculosis* gene expression during adaptation to stationary phase and low-oxygen dormancy. *Tuberculosis* 2004, **84(3–4):**218-227.

40. Muttucumaru DGN, Roberts G, Hinds J, Stabler RA, Parish T: **Gene expression profile of *Mycobacterium tuberculosis* in a non-replicating state.** *Tuberculosis* 2004, **84(3–4):**239-246.

41. Hampshire T, Soneji S, Bacon J, James BW, Hinds J, Laing K, Stabler RA, Marsh PD, Butcher PD: **Stationary phase gene expression of *Mycobacterium tuberculosis* following a progressive nutrient depletion: a model for persistent organisms?** *Tuberculosis* 2004, **84(3–4):**228-238.

42. Date SV, Marcotte EM: **Protein function prediction using the Protein Link Explorer (PLEX).** *Bioinformatics* 2005, **21(10):**2558-2559.

43. Raman K, Chandra N: *Mycobacterium tuberculosis* interactome analysis unravels potential pathways to drug resistance. *BMC Microbiology* 2008, **8(1):**234.

44. Verkhedkar KD, Raman K, Chandra N, Vishveshwara S: **Metabolome based reaction graphs of *M. tuberculosis* and *M. leprae* : a comparative network analysis.** *PLoS ONE* 2007, **2(9):**e881.

45. Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD: **Global network analysis of phenotypic effects: protein networks and toxicity modulation in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2004, **101(52):**18006-18011.

46. Potapov AP, Goemann B, Wingender E: **The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks.** *BMC Bioinformatics* 2008, **9:**227.

47. Cohen-Gonsaud M, Ducasse S, Hoh F, Zerbib D, Labesse G, Quémard A: **Crystal structure of MabA from *Mycobacterium tuberculosis*, a reductase involved in long-chain fatty acid biosynthesis.** *Journal of Molecular Biology* 2002, **320(2):**249-261.

48. Marrakchi H, Ducasse S, Labesse G, Montrozier H, Margeat E, Emorine LJ, Charpentier X, Daffé M, Quémard A: **MabA (FabG1), a *Mycobacterium tuberculosis* protein involved in the long-chain fatty acid elongation system FAS-II.** *Microbiology* 2002, **148:**951-960.

49. Buchmeier NA, Fahey RC: **The *mshA* gene encoding the glycosyltransferase of mycothiol biosynthesis is essential in *Mycobacterium tuberculosis* Erdman.** *FEMS Microbiol Lett* 2006, **264:**74-79.

50. Terwilliger TC, Park MS, Waldo GS, Berendzen J, Hung LW, Kim CY, Smith CV, Sacchettini JC, Bellinzoni M, Bossi R, De Rossi E, Mattevi A, Milano A, Riccardi G, Rizzi M, Roberts MM, Coker AR, Fossati G, Mascagni P, Coates ARM, Wood SP, Goulding CW, Apostol MI, Anderson DH, Gill HS, Eisenberg DS, Taneja B, Mande S, Pohl E, Lamzin V, Tucker P, Wilmanns M, Colovos C, Meyer-Klaucke W, Munro AW, McLean KJ, Marshall KR, Leys D, Yang JK, Yoon HJ, Lee BI, Lee MG, Kwak JE, Han BW, Lee JY, Baek SH, Suh SW, Komen MM, Arcus VL, Baker EN, Lott JS, Jacobs W, Alber T, Rupp B: **The TB**

structural genomics consortium: a resource for *Mycobacterium tuberculosis* biology. *Tuberculosis* 2003, **83(4):**223-249.

51. Edwards A, Berman J, Sundström M: **Structural Genomics and Drug Discovery.** In *Annual Reports in Medicinal Chemistry Volume 40.* Edited by: Doherty AM, Bock MG, Desai MC, Overington J, Plattner JJ, Stamford A, Wustrow D, Young H. Academic Press; 2005:349-369.

52. Gileadi O, Knapp S, Lee W, Marsden B, Müller S, Niesen F, Kavanagh K, Ball L, von Delft F, Doyle D, Oppermann U, Sundström M: **The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins.** *J Struct Funct Genomics* 2007, **8(2-3):**107-119.

53. Wehenkel A, Bellinzoni M, Graña M, Duran R, Villarino A, Fernandez P, Andre-Leroux G, England P, Takiff H, Cerveñansky C, Cole ST, Alzari PM: **Mycobacterial Ser/Thr protein kinases and phosphatases: physiological roles and therapeutic potential.** *Biochim Biophys Acta* 2008, **1784:**193-202.

54. Levy J: **The effects of antibiotic use on gastrointestinal function.** *Am J Gastroenterol* 2000, **95(1 Suppl):**S8-10.

55. Nicholson JK, Wilson ID: **Understanding 'global' systems biology: metabonomics and the continuum of metabolism.** *Nat Rev Drug Discov* 2003, **2(8):**668-676.

56. Nicholson JK, Holmes E, Wilson ID: **Gut microorganisms, mammalian metabolism and personalized health care.** *Nat Rev Microbiol* 2005, **3(5):**431-438.

57. Wang Y, Xiao CL: **Isocitrate lyase, a new target in anti-tuberculosis drug research.** *Chinese Journal of Antibiotics* 2007, **32(7):**391-395.

58. Senaratne RH, Silva ADD, Williams SJ, Mougous JD, Reader JR, Zhang T, Chan S, Sidders B, Lee DH, Chan J, Bertozzi CR, Riley LW: **5'-Adenosinephosphosulphate reductase (CysH) protects Mycobacterium tuberculosis against free radicals during chronic infection phase in mice.** *Mol Microbiol* 2006, **59(6):**1744-1753.

59. Buetow L, Brown AC, Parish T, Hunter WN: **The structure of Mycobacteria 2C-methyl-D-erythritol-2,4-cyclodiphosphate synthase, an essential enzyme, provides a platform for drug discovery.** *BMC Struct Biol* 2007, **7:**68.

60. Cornish-Bowden A, Cárdenas ML: **Metabolic analysis in drug design.** *C R Biologies* 2003, **326:**509-515.

61. Raman K, Rajagopalan P, Chandra N: **Hallmarks of mycolic acid biosynthesis: a comparative genomics study.** *Proteins* 2007, **69(2):**358-368.

62. Voit EO: *Computational Analysis of Biochemical Systems* Cambridge, UK: Cambridge University Press; 2000.

63. Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E, Palsson BØ: **Metabolic modeling of microbial strains *in silico*.** *Trends in Biochemical Sciences* 2001, **26(3):**179-186.

64. Barabási AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5(2):**101-113.

65. Materi W, Wishart DS: **Computational systems biology in drug discovery and development: methods and applications.** *Drug Discov Today* 2007, **12(7–8):**295-303.

66. Raman K, Rajagopalan P, Chandra N: **Principles and Practices of Pathway Modelling.** *Current Bioinformatics* 2006, **1:**147-160.

67. Mdluli K, Slayden RA, Zhu Y, Ramaswamy S, Pan X, Mead D, Crane DD, Musser JM, Barry CE III: **Inhibition of a *Mycobacterium tuberculosis* -ketoacyl ACP Synthase by Isoniazid.** *Science* 1998, **280:**1607-1610.

68. Chong CE, Lim BS, Nathan S, Mohamed R: ***In silico* analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets.** In *Silico Biol* 2006, **6(4):**341-346.

69. Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, Bandyopadhyay D: ***In silico* identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*.** In *Silico Biol* 2006, **6(1–2):**43-47.

70. Perumal D, Lim CS, Sakharkar KR, Sakharkar MK: **Differential genome analyses of metabolic enzymes in *Pseudomonas aeruginosa* for drug target identification.** In *Silico Biol* 2007, **7(4–5):**453-465.

71. Sakharkar KR, Sakharkar MK, Chow VTK: **A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*.** In *Silico Biol* 2004, **4(3):**355-360.

72. Singh NK, Selvam SM, Chakravarthy P: **T-iDT : tool for identification of drug target in bacteria and validation by *Mycobacterium tuberculosis*.** In *Silico Biol* 2006, **6(6):**485-493.

73. Ramachandraiah G, Chandra N: **Sequence and structural determinants of mannose recognition.** *Proteins* 2000, **39(4):**358-364.

74. Vinod PK, Konkimalla B, Chandra N: ***In-silico* pharmacodynamics: correlation of adverse effects of H2-antihistamines with histamine N-methyl transferase binding potential.** *Appl Bioinformatics* 2006, **5(3):**141-150.

75. Escuyer VE, Lety MA, Torrelles JB, Khoo KH, Tang JB, Rithner CD, Frehel C, McNeil MR, Brennan PJ, Chatterjee D: **The Role of the embA and embB Gene Products in the Biosynthesis of the Terminal Hexaarabinofuranosyl Motif of *Mycobacterium smegmatis* Arabinogalactan.** *The Journal of Biological Chemistry* 2001, **276(52):**48854-48862.

76. Alderwick LJ, Seidel M, Sahm H, Besra GS, Eggeling L: **Identification of a novel arabinofuranosyltransferase (AftA) involved in cell wall arabinan biosynthesis in *Mycobacterium tuberculosis*.** *The Journal of Biological Chemistry* 2006, **281(23):**15653-15661.

77. Seidel M, Alderwick LJ, Birch HL, Sahm H, Eggeling L, Besra GS: **Identification of a novel arabinofuranosyltransferase AftB involved in a terminal step of cell wall arabinan biosynthesis in *Corynebacterianeae*, such as *Corynebacterium glutamicum* and *Mycobacterium tuberculosis*.** *J Biol Chem* 2007, **282(20):**14729-14740.

78. Hirano S, Ichikawa S, Matsuda A: **Design and synthesis of diketopiperazine and acyclic analogs related to the caprazamycins and liposidomycins as potential antibacterial agents.** *Bioorg Med Chem* 2008, **16:**428-436.

79. Choi KH, Kremer L, Besra GS, Rock CO: **Identification and substrate specificity of -ketoacyl (acyl carrier protein) synthase III (mtFabH) from *Mycobacterium tuberculosis*.** *The Journal of Biological Chemistry* 2000, **275:**28201-28207.

80. Musayev F, Sachdeva S, Scarsdale JN, Reynolds KA, Wright HT: **Crystal structure of a substrate complex of *Mycobacterium tuberculosis* -ketoacyl-acyl carrier protein synthase III (FabH) with lauroyl-coenzyme A.** *Journal of Molecular Biology* 2005, **346(5):**1313-1321.

81. Kremer L, Nampoothiri KM, Lesjean S, Dover LG, Graham S, Betts JC, Brennan PJ, Minnikin DE, Locht C, Besra GS: **Biochemical Characterization of Acyl Carrier Protein (AcpM) and Malonyl-CoA:AcpM Transacylase (mtFabD), Two Major Components of *Mycobacterium tuberculosis* Fatty Acid Synthase II.** *The Journal of Biological Chemistry* 2001, **276:**27967-27974.

82. Zhang YM, White SW, Rock CO: **Inhibiting bacterial fatty acid synthesis.** *J Biol Chem* 2006, **281(26):**17541-17544.

83. Ghadbane H, Brown AK, Kremer L, Besra GS, Fütterer K: **Structure of *Mycobacterium tuberculosis* mtFabD, a malonyl-CoA:acyl carrier protein transacylase (MCAT).** *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2007, **63(Pt 10):**831-835.

84. Wilson M, DeRisi J, Kristensen HH, Imboden P, Rane S, Brown PO, Schoolnik GK: **Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization.** *Proc Natl Acad Sci USA* 1999, **96:**12833-12838.

85. Portevin D, Sousa-D'Auria CD, Houssin C, Grimaldi C, Chami M, Daffé M, Guilhot C: **A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms.** *Proc Natl Acad Sci U S A* 2004, **101(1):**314-319.

86. Alahari A, Trivelli X, Guérardel Y, Dover LG, Besra GS, Sacchettini JC, Reynolds RC, Coxon GD, Kremer L: **Thiacetazone, an antitubercular drug that inhibits cyclopropanation of cell wall mycolic acids in mycobacteria.** *PLoS ONE* 2007, **2(12):**e1343.

87. Portevin D, de Sousa-D'Auria C, Montrozier H, Houssin C, Stella A, Lanéelle MA, Bardou F, Guilhot C, Daffé M: **The Acyl-AMP Ligase FadD32 and AccD4-containing Acyl-CoA Carboxylase Are Required for the Synthesis of Mycolic Acids and Essential for Mycobacterial Growth.** *The Journal of Biological Chemistry* 2005, **280(10):**8862-8874.

88. Phetsuksiri B, Jackson M, Scherman H, McNeil MR, Besra GS, Baulard AR, Slayden RA, DeBarber AE, Barry CE III, Baird MS, Crick DC, Brennan PJ: **Unique Mechanism of Action of the Thiourea Drug Isoxyl on *Mycobacterium tuberculosis*.** *The Journal of Biological Chemistry* 2003, **278(52):**53123-53130.

89. Zimhony O, Cox JS, Welch JT, Vilchèze C, Jacobs WR Jr: **Pyrazinamide inhibits the eukaryotic-like fatty acid synthetase I (FASI) of *Mycobacterium tuberculosis*.** *Nat Med* 2000, **6(9):**1043-1047.

90. Pavelka MS Jr, Chen B, Kelley CL, Collins FM, Jacobs WR Jr: **Vaccine efficacy of a lysine auxotroph of *Mycobacterium tuberculosis*.** *Infect Immun* 2003, **71(7):**4190-4192.

91. Gokulan K, Rupp B, Pavelka MS Jr, Jacobs WR Jr, Sacchettini JC: **Crystal structure of *Mycobacterium tuberculosis* diaminopimelate decarboxylase, an essential enzyme in bacterial lysine biosynthesis.** *The Journal of Biological Chemistry* 2003, **278(20):**18588-18596.

92. Koon N, Squire CJ, Baker EN: **Crystal structure of LeuA from *Mycobacterium tuberculosis*, a key enzyme in leucine biosynthesis.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101(22):**8295-8300.

93. Paiva AM, Vanderwall DE, Blanchard JS, Kozarich JW, Williamson JM, Kelly TM: **Inhibitors of dihydrodipicolinate reductase, a key enzyme of the diaminopimelate pathway of *Mycobacterium tuberculosis*.** *Biochim Biophys Acta* 2001, **1545(1–2):**67-77.

94. de Mendonça JD, Ely F, Palma MS, Frazzon J, Basso LA, Santos DS: **Functional characterization by genetic complementation of aroB-encoded dehydroquinate synthase from *Mycobacterium tuberculosis* H37Rv and its heterologous expression and purification.** *J Bacteriol* 2007, **189(17):**6246-6252.

95. Errey JC, Blanchard JS: **Functional characterization of a novel ArgA from *Mycobacterium tuberculosis*.** *Journal of Bacteriology* 2005, **187(9):**3039-3044.

96. Gerum AB, Ulmer JE, Jacobus DP, Jensen NP, Sherman DR, Sibley CH: **Novel *Saccharomyces cerevisiae* screen identifies WR99210 analogues that inhibit *Mycobacterium tuberculosis* dihydrofolate reductase.** *Antimicrob Agents Chemother* 2002, **46(11):**3362-3369.

97. Li R, Sirawaraporn R, Chitnumsub P, Sirawaraporn W, Wooden J, Athappilly F, Turley S, Hol WGJ: **Three-dimensional structure of *M. tuberculosis* dihydrofolate reductase reveals opportunities for the design of novel tuberculosis drugs.** *Journal of Molecular Biology* 2000, **295(2):**307-323.

98. Sambandamurthy VK, Wang X, Chen B, Russell RG, Derrick S, Collins FM, Morris SL, Jacobs WR Jr: **A pantothenate auxotroph of *Mycobacterium tuberculosis* is highly attenuated and protects mice against tuberculosis.** *Nature Medicine* 2002, **8(10):**1171-1174.

99. Velaparthi S, Brunsteiner M, Uddin R, Wan B, Franzblau SG, Petukhov PA: **5-tert-butyl-N-pyrazol-4-yl-4,5,6,7-tetrahydrobenzo[d]isoxazole-3-carboxamide derivatives as novel potent inhibitors of *Mycobacterium tuberculosis* pantothenate synthetase: initiating a quest for new antitubercular drugs.** *J Med Chem* 2008, **51(7):**1999-2002.

100. Gopalan G, Chopra S, Ranganathan A, Swaminathan K: **Crystal structure of uncleaved L-aspartate-alpha-decarboxylase from Mycobacterium tuberculosis.** *Proteins* 2006, **65(4):**796-802.

101. Kumar P, Chhibber M, Surolia A: **How pantothenol intervenes in Coenzyme-A biosynthesis of *Mycobacterium tuberculosis*.** *Biochem Biophys Res Commun* 2007, **361(4):**903-909.

102. Das S, Kumar P, Bhor V, Surolia A, Vijayan M: **Invariance and variability in bacterial PanK: A study based on the crystal structure of *Mycobacterium tuberculosis* PanK.** *Acta Crystallogr D Biol Crystallogr* 2006, **62(Pt 6):**628-638.

103. Schelle MW, Bertozzi CR: **Sulfate metabolism in mycobacteria.** *Chembiochem* 2006, **7(10):**1516-1524.

104. Carroll KS, Gao H, Chen H, Stout CD, Leary JA, Bertozzi CR: **A conserved mechanism for sulfonucleotide reduction.** *PLoS Biol* 2005, **3(8):**e250.

105. Mougous JD, Senaratne RH, Petzold CJ, Jain M, Lee DH, Schelle MW, Leavell MD, Cox JS, Leary JA, Riley LW, Bertozzi CR: **A sulfated metabolite produced by stf3 negatively regulates the virulence of Mycobacterium tuberculosis.** *Proc Natl Acad Sci USA* 2006, **103(11):**4258-4263.

106. Williams SJ, Senaratne RH, Mougous JD, Riley LW, Bertozzi CR: **5'-adenosinephosphosulfate lies at a metabolic branch point in mycobacteria.** *The Journal of Biological Chemistry* 2002, **277(36):**32606-32615.

107. Eisenreich W, Bacher A, Arigoni D, Rohdich F: **Biosynthesis of isoprenoids via the non-mevalonate pathway.** *Cell Mol Life Sci* 2004, **61(12):**1401-1426.

108. Muñoz-Elías EJ, McKinney JD: ***Mycobacterium tuberculosis* isocitrate lyases 1 and 2 are jointly required for *in vivo* growth and virulence.** *Nature Medicine* 2005, **11(6):**638-644.

109. Andries K, Verhasselt P, Guillemont J, Gohlmann HWH, Neefs JM, Winkler H, Van Gestel J, Timmerman P, Zhu M, Lee E, Williams P, de Chaffoy D, Huitric E, Hoffner S, Cambau E, Truffot-Pernot C, Lounis N, Jarlier V: **A Diarylquinoline Drug Active on the ATP Synthase of *Mycobacterium tuberculosis*.** *Science* 2005, **307(5707):**223-227.

110. McLean KJ, Dunford AJ, Neeli R, Driscoll MD, Munro AW: **Structure, function and drug targeting in *Mycobacterium tuberculosis* cytochrome P450 systems.** *Arch Biochem Biophys* 2007, **464(2):**228-240.

111. Betts JC, McLaren A, Lennon MG, Kelly FM, Lukey PT, Blakemore SJ, Duncan K: **Signature gene expression profiles discriminate between isoniazid-, thiolactomycin-, and triclosan-treated *Mycobacterium tuberculosis*.** *Antimicrob Agents Chemother* 2003, **47(9):**2903-2913.

112. Aubry A, Pan XS, Fisher LM, Jarlier V, Cambau E: ***Mycobacterium tuberculosis* DNA Gyrase: Interaction with Quinolones and Correlation with Antimycobacterial Drug Activity.** *Antimicrob Agents Chemother* 2004, **48(4):**1281-1288.

113. Mdluli K, Ma Z: ***Mycobacterium tuberculosis* DNA gyrase as a target for drug discovery.** *Infect Disord Drug Targets* 2007, **7(2):**159-168.

114. Sreevatsan S, Pan X, Stockbauer KE, Williams DL, Kreiswirth BN, Musser JM: **Characterization of *rpsL* and *rrs* mutations in streptomycin-resistant *Mycobacterium tuberculosis* isolates from diverse geographic localities.** *Antimicrob Agents Chemother* 1996, **40(4):**1024-1026.

115. Parish T, Stoker NG: **glnE is an essential gene in *Mycobacterium tuberculosis*.** *Journal of Bacteriology* 2000, **182(20):**5715-5720.

116. Zahrt TC, Deretic V: **An essential two-component signal transduction system in *Mycobacterium tuberculosis*.** *Journal of Bacteriology* 2000, **182(13):**3832-3838.

117. Saini DK, Tyagi JS: **High-throughput microplate phosphorylation assays based on DevR-DevS/Rv2027c 2-component signal transduction pathway to screen for novel antitubercular compounds.** *J Biomol Screen* 2005, **10(3):**215-224.

118. Boon C, Dick T: ***Mycobacterium bovis* BCG response regulator essential for hypoxic dormancy.** *Journal of Bacteriology* 2002, **184(24):**6760-6767.

119. Saini DK, Malhotra V, Dey D, Pant N, Das TK, Tyagi JS: **DevR-DevS is a bona fide two-component system of that is hypoxia-responsive in the absence of the DNA-binding domain of DevR.** *Microbiology* 2004, **150(Pt 4):**865-875.

120. Scherr N, Honnappa S, Kunz G, Mueller P, Jayachandran R, Winkler F, Pieters J, Steinmetz MO: **Structural basis for the specific inhibition of protein kinase G, a virulence factor of *Mycobacterium tuberculosis*.** *Proc Natl Acad Sci USA* 2007, **104(29):**12151-12156.

121. Qiao C, Gupte A, Boshoff HI, Wilson DJ, Bennett EM, Somu RV, Barry CE III, Aldrich CC: **5'-O-[(N-Acyl)sulfamoyl]adenosines as Antitubercular Agents that Inhibit MbtA: An Adenylation Enzyme Required for Siderophore Biosynthesis of the Mycobactins.** *Journal of Medicinal Chemistry* 2007, **50(24):**6080-6094.

122. Rodriguez GM, Voskuil MI, Gold B, Schoolnik GK, Smith I: **ideR, an essential gene in *Mycobacterium tuberculosis* : Role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response.** *Infect Immun* 2002, **70(7):**3371-3381.

123. Monfeli RR, Beeson C: **Targeting iron acquisition by *Mycobacterium tuberculosis*.** *Infect Disord Drug Targets* 2007, **7(3):**213-220.