Int J Pharm Bio Sci 2014 April ; 5 (2) : (B) 155 - 175

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

HYPOTHETICAL PROTEINS OF *MYCOBACTERIUM TUBERCULOSIS* H37Rv AS DRUG TARGETS

ZAHRA M. AL-KHAFAJI *

Institute of Genetic Engineering and Biotechnology for Postgraduate Studies ,University of Baghdad , IRAQ

ABSTRACT

Hypothetical proteins often with unknown functions . *Mycobacterium tuberculosis* H37Rv has a very large number of such proteins , which can be exploited as drug target as the bacterium became more virulent with developing antibiotic resistance to first line and second line of drugs . Some hypothetical proteins (290 proteins) were studied from different aspects related to find the druggable proteins . Results shown that some of it get good druggability score such as Rv2074, Rv2827c , Rv0295c which are worth to be subjected to further scheme of drug design such as docking or using other computational techniques .

KEY WORDS : Rv proteins , Drug targets , *M. tuberculosis* , Protein druggability , New drugs , Bioinformatics

ZAHRA M. AL-KHAFAJI
Institute of Genetic Engineering and Biotechnology for Postgraduate Studies ,
University of Baghdad , IRAQ

INTRODUCTION

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* remains a major public health *problem* around the world [1] . and global surveillance indicates that multidrug-drug resistant *M. tuberculosis* (MRD-TB) and extensively –drug resistance (XDR-TB) are spreading to many spots posing a major threat to TB eradication programme [2] . The success of *M. tuberculosis* is in part due to the discontinuity of its life cycle in the host system , owing to its dormant ability to enter and exit from different states in response to antimycobacterial host defense mechanisms , enabling it to infect , grow , persist and survive in human macrophages[1] . This means good and active immune response quarantine the bacteria in granuloma which halts replication of bacilli and suppresses the immediate threat of active infection[3,4] and cells will be in dormant phase under different types of stresses . Approximately 10% of latent infections reactivate upon flattering the immune system as in HIV infections , malnutrition , using immunosuppressive drugs and others[5] . It is known that only Rifampicin and Pyrazinamide show activity against persister bacilli that are in dormant phase , however , they do not eliminate all dormant population , and Pyrazinamide is likely to effect only those persisters that reside in acidic pH conditions[6] . No new antituberculosis drugs have been developed for over 20 years[7] . Identification of novel drug targets for diseases and development of new drugs have always been expensive and time-consuming , the latter could extend to 10-15 years[1] . Currently , technological developments in large-scale biological experiments , coupled with Bioinformatics tools can provide inexpensive strategies which shorten the length of time spent in drug discovery[1]. The drugs should be able to kill the intracellular organisms that are in a persistent state for the *M. tuberculosis*[1]. On the other hand, it has been suggested that targeting already known *M. tuberculosis* targets for drug development have limited success because of potential cross-resistance[6,8] . The aim of this study was to characterize some hypothetical proteins (Rv proteins) which could be good drug target of the most virulent strain *M. tuberculosis* H37Rv .

MATERIALS AND METHODS

- Retrieval of Rv gene sequences and their products (Rv protein sequences) :

Some hypothetical protein sequences were retrieved as in previous study[9] using TubercuList database (http://tuberculist.epfl.ch/)

- Host (Human) homology with bacterial proteins was carried out by subtractive genomics method using DEG database (http://tubic.tju.edu.cn/deg/)
- BLASTn (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) was used to find out the similarity.
- TDR database[10] (http://tdrtargets.org/) was used for different purposes.
  - Finding the expression level[3,8]
  - Proteins segments associated with membranes
  - Finding Pfam database entries with suggested functions
  - Classification of proteins according to gene ontology (GO)
  - Essentiality of proteins
  - Other characters
- MycoPrInt database (http://www.imtech.res.in/raghava/mycoprint/) was used to estimate protein interactions (Interactome).
- webTB (http://www.webtb.org/) was used to find out the proteins with pdb files .
- DoG Site Score package[11] (http://dogsite.zbh.uni-hamburg.de/) was used to find protein druggability .

RESULTS

M. tuberculosis and specially the virulent strain H37Rv got a lot of hypothetical proteins with unknown functions (or at least not proved yet) . Some of these proteins studied previously[9] was used in this study (290 proteins) , the first step is to find similarity with the host genome , this was done using subtractive genomics approach[12,13] Table 1 shows the degree of similarity of some hypothetical proteins and human essential genes .

Table 1
Similarity of hypothetical proteins with human genes

Rv ID	Human Gene ID	% Similarity
Rv0699	DEG20060001	38
Rv3612c	DEG20060002	32
Rv0008c	DEG20060008	32
Rv3718c	DEG200600022	29
Rv0636		30
Rv1961	DEG200600026	44
Rv3755c	DEG200600034	31
Rv1890c	DEG200600038	28
Rv0122	DEG200600045	34
Rv0241c	DEG200600048	39
Rv0636		39
Rv2020c	DEG200600064	38
Rv1961	DEG200600083	43
Rv2288	DEG200600092	33
Rv1535	DEG200600099	39
	DEG200600105	69

Some of proteins exhibited some similarity, those with similarity of 60% or more were excluded, the rest of proteins showed no similarity.
Degree of expression at dormant phase[3, 8] at different level are shown in Fig 1 (Suppl Table S1)

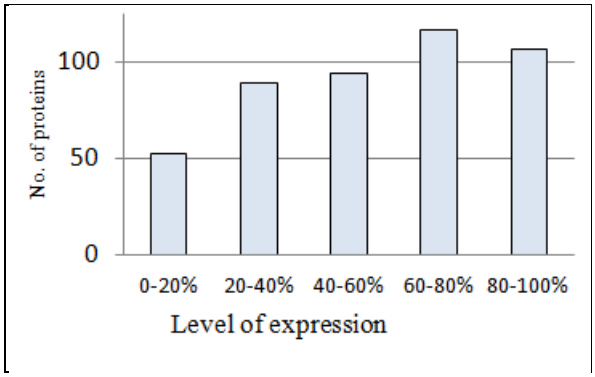

Figure 1
Distribution of proteins according to the level of expression in dormant phase.

It has been found that 45 protein out of 290 (15.5%) had some essentiality (Suppl Table S2).
Some of the hypothetical proteins under investigation showed different numbers of segments associated with membrane Fig 2.

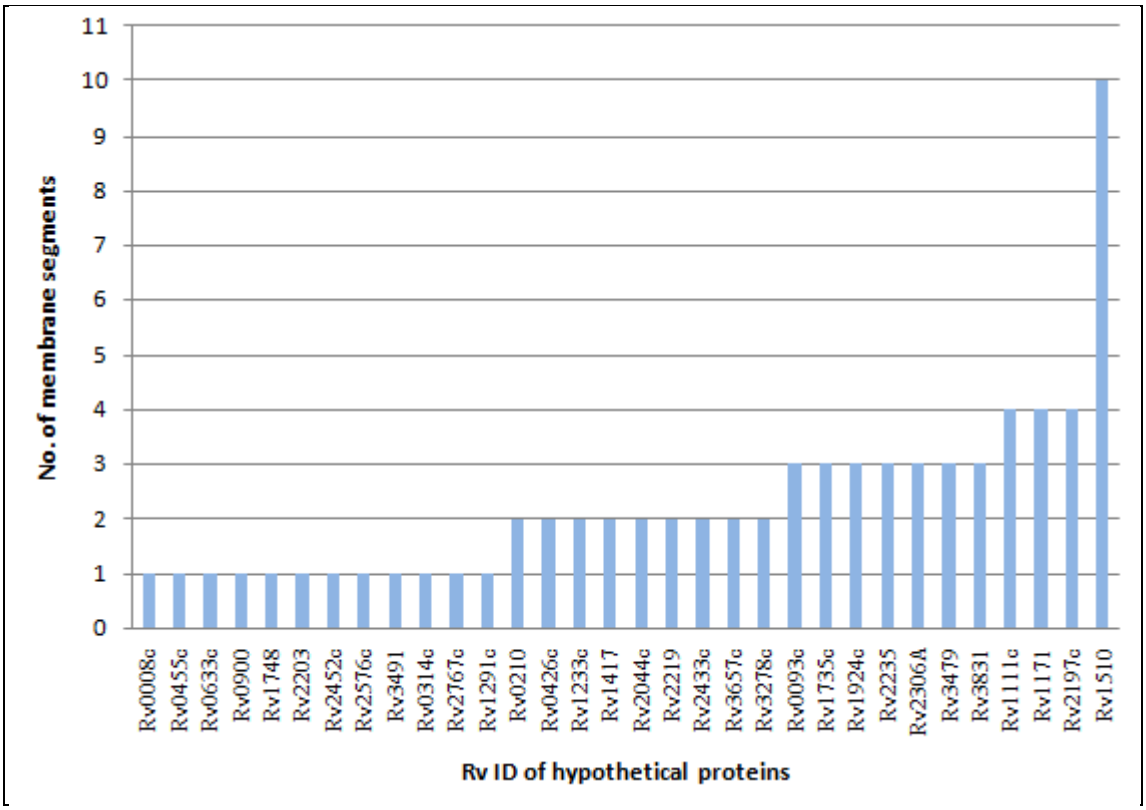

Figure 2
Proteins with membranous segments

Number of proteins got Pfam and GO entries (Suppl Table S3, suppl Table S4) Protein interactions has a very good strategy for targeting, this can be found by querying MycoPrInt database (Suppl Table S5). Some proteins interact with a large number of proteins as Rv0898c which interact with 908 proteins and Rv1233c interacts with 364 proteins , the rest ranged between undetected state and the degrees mentioned . Some proteins were distributed among different orthologous groups of OrthoMCL cluster (TDR database) (Suppl Table S6) . Structural clue is considered an integral issue in drug design and discovery, so the website (webTB) was screened to find out the proteins with pdb files, this shown in Table 2.

Table 2
Proteins with pdb structure(s)

Rv ID	Pdb ID
Rv1873	2D2Y 2JEK
Rv2074	2ASF
Rv2827c	1ZEL
Rv0295c	1TEX
Rv1155	1w9a, 1xxo, 1y30, 2aq6
Rv2740	2bng

Very little of hypothetical proteins were found with pdb structure, 9 out of 290 (3.1%). Those were used to find protein druggability using DoGSiteScore. It has been found that protein Rv0295c (pdb ID 1TEx) represents a protein with good druggability as 8 pockets get high score (0.78-0.8) as shown in Table 3

Table 3
Some of hypothetical proteins with potential druggability

Rv ID	No. of pockets suitable for molecule binding	Drug Score
Rv1873 (2D2Y 2JEK)	5 4	0.66 0.72
Rv2074 (2ASF)	3	0.85
Rv2827c (1ZEL)	19	0.84
Rv0295c (1TEX)	19	0.81
Rv1155 (1w9a, 1xxo, 1y30, 2aq6)	0	
Rv2740 (2bng)	0	

In addition, some hypothetical proteins were found to have enzymatic activity as in Rv0521c (Methyltrasnsferase , EC 2.1.1.-) , Rv0552 (Methyltransfearase , EC 2.1.1.-) and Rv0492c (Oxidoreductase , EC 1.-.-.-) .

DISCUSSION

Mycobacterium tuberculosis strain H37Rv draws attention of many research groups, therefore, databases and websites were developed for it[14]. Availability of genome sequence of pathogens has provided a tremendous amount of information that can be useful for drug target and vaccine target identification[12] . However , upon annotation of genomes a lot of hypothetical proteins with unknown functions appeared [8] , some of these proteins were studied to assign their function mainly with *in Silico* mode[9,15] . So these proteins can be exploited as drug targets using subtractive genomics method[12] . Genome subtraction approach has been successfully used to locate drug targets of *Pseudomonas aeruginosa* and other bacteria[12,13] . As target identification is the first step in drug design and discovery, this step on stage is complicated by the fact that the identified target must satisfy a variety of criteria to permit progression to the next step . Important factors in this context in clued homology between the target and host to avoid host toxicity, this homology must be low or non exist1[6]. In this study most of the hypothetical proteins were satisfied this criterion, except Rv1535 which showed more than 60% homology which has been used as a cutoff of acceptance (Table 1) . On the other hand the long treatment time of TB will encourage selection pressure for drug resistance in the natural flora[8] , this point was avoided in this study ,since the gut flora are in large numbers and highly variable among human populations and depends on several factors . The other parameter which can be considered is the essentiality, which included in DEG database, it has been reported that 570 genes could be essential recorded for *M. tuberculosis* under nutrient-rich conditions[8], such gene essentiality attracts attention as drug target[8,17,18] . The essentiality recorded for *M. tuberculosis* under certain conditions using transposon site hybridization[18,19] , which cannot give a full and complete picture about gene /protein essentiality . On the other hand in *M. tuberculosis* infection , dormancy represents an important phase , expression of genes under such conditions considers as a good targets for investigations , specially those with high expression in dormant phase (Fig 1, Suppl Table 1) , some genes expressed in latent infection (*in vivo*) could mean that it is required for survival during dormancy[8] , therefore , and only further laboratory knockout studies will show the gene essentiality .

It has been investigated that protein-protein interaction (PPI) as a potential approach for understanding the *M. tuberculosis* pathogenesis and building up networks[2,20] , so proteins with high level of interaction can be targeted such as Rv0898c and Rv1233c . Orthologous proteins are often functionally and structurally similar. They may be modulated by identical of similar molecules making orthology assessment a powerful tool to connect a known druggable target with a potential novel

targets[3,10] , and this approach has been used in TDR database , so drugs for some COGs can be borrowed as a start point for designing new drugs . Structural clues of proteins such as presence of pdb structure are helpful for designing drugs, 9 or the hypothetical proteins have pdb structure (Table 2) . The availability of a target crystal structure would aid in rational drug design and therefore provide a strong practical advantage in throughput docking and lead optimization[8]. So , proteins with pdb structure were used to measure the druggability of proteins which their activity has previously been shown to be modulated by binding small molecules and have binding affinity below 10 μM to compound pass through Lipinisky rule of 5[8,11] . DoGSiteScore software is an automated pocket detection and analysis tool which can be used for protein druggability assessment, it is based on calculated size, shape, and chemical features of automatically predicted pocket, incorporated into a support vector machine for druggability estimation . This software needs input pdb ID of the protein[11], using this software results showed that some proteins are worth to be pursued as drug target .

CONCLUSION

Novel strategies are essentially needed to overt the global catastrophic forecast by the WHO , therefore , computational approach for drug identification specifically for *M. tuberculosis* can produce a list of reliable targets very rapidly , as this approach has the advantage of speed and low cost[7] . It would be desirable if the target selected, is essential for both growing and dormant bacteria. So proteins of Table 3, Rv2827c, Rv0295c, Rv2074 and Rv1873 could be good targets as they are druggable and interact with other proteins ranging from 3 to 9 and at desirable level of expression and with recorded essentiality

REFERENCES

1. Mulder N., Mazandu G., Mulder H. Using host-pathogen functional interactions for filtering potential drug targets in *Mycobacterium tuberculosis* . J Mycobac Dis 2013; 3:1-8.
2. Raman K., Chandra N. *Mycobacterium tuberculosis* interactome analysis unravels potential pathways to drug resistance . BMC Microbiol 2008 ; 8:234 - 246 .
3. Dennis J Murphy D. Brown J. Identification of gene targets against dormant phase *Mycobacterium tuberculosis* infections . BMC Inf Dis 2007; 7: 84-99 .
4. Wayne G., Sohaskey C. Nonreplicating persistence of *Mycobacterium tuberculosis*. Annu Rev Microbiol 2001; 55:139-163 .
5. Tyagi J., Sharma D. Signal transduction systems of Mycobacteria with special reference to M. tuberculosis . Curr Sci 2004 ; 86: 93-102.
6. Chopra I., Hesse L.., Neill A. Exploiting current understanding of antibiotic action for discovery of new drugs. Symp Ser Soc Appl Microbiol 2002 ; 4S–15S.
7. Asif S., Asad A ., Faizan A., Anjali M., Arvind A, Neelesh K., Hirdesh K., Sanjay K. Dataset of potential targets for Mycobacterium tuberculosis H37Rv through comparative genome analysis . Bioinformation 2009 ; 4(6): 245-248 .
8. Hasan S., Daugelat S., Rao S., Schreiber M. Prioritizing Genomic Drug Targets in Pathogens: Application to *Mycobacterium tuberculosis .* PLoS Biol 2006; 2: e61
9. Al-Khafaji Z. *In Silico* Investigation of Rv hypothetical proteins of virulent strain *Mycobacterium tuberculosis* H37Rv. Indian J Pharm Biol Res 2013;1(4):81-88
10. Magarin M., Carmona S., Crowther G., Ralph S., Roos D., Shanmugam D., Voorhis W., Agu F., TDR Targets: a

chemogenomics resource for neglected diseases . Nuc Acids Res 2012 ; 40 : D1118–D1127 .
11. Volkamer A ., Kuhn D., Grombacher T., Rippmann F. , Rarey M. Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model*. 2012 ; 52: 360-372.
12. Dutta A. , Singh S., 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 : 43-47 .
13. Sakharkar K., Sakharkar M. Chow V. A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. *In Silico* Biol 2004; 4 : 0028.
14. Singla D., Tewari R. , Kumar A., Raghava G. Designing of inhibitors against drug tolerant Mycobacterium tuberculosis (H37Rv) . Chem Central J 2013 ; 7:49-60 .
15. Doerks T., van Noort V., Minguez P., Bork P. Annotation of the *M. tuberculosis* Hypothetical Orfeome: Adding Functional Information to More than Half of the Uncharacterized Proteins. PLoS ONE 2012 7(4): e34302.
16. Freiberg C. Novel computational methods in anti-microbial target identification. Drug Disco Today 2001; 6: S72–S80.
17. Freiberg C., Wieland B., Spaltmann F., Ehlert K., Brotz H., Identification of novel essential *Escherichia coli* genes conserved among pathogenic bacteria. J Mol Microbiol Biotechnol 2001; 3: 483–489.
18. Sassetti C., Boyd D., Rubin E. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol 2003; 48: 77–84.
19. Sassetti C., Boyd D., Rubin E . Comprehensive identification of conditionally essential genes in mycobacteria. Proc Natl Acad Sci USA 2001; 98 : 12712–12717.
20. Cui T., Zhang L., Wang X ., He Z. Uncovering new signaling proteins and potential drug targets through the interactome analysis of *Mycobacterium tuberculosis* . BMC Genomics 2009, 10:118 -127.

Supplement Table S1
Percent of expression of proteins in dormant phase

0-20%	20-40%	40-60%	60-80%	80=100%
	Rv0008c			
		Rv0057		
	Rv0028			
		Rv0030		
Rv0057				
		Rv0061		
				Rv0079
				Rv0080
				Rv0088
				Rv0093c
		Rv0098		
		Rv0108c		
				Rv0122
				Rv0123
			Rv0193c	
		Rv0209		
			Rv0210	
Rv0241c				
		Rv0257c		
			Rv0272c	
		Rv0295c		
	Rv0299			

			Rv0307c	
		Rv0311		
				Rv0313
Rv0332				
			Rv0333	
			Rv0340	
		Rv0358		
	Rv0360c			
	Rv0381c			
	Rv0390			
	Rv0395	Rv0395		
	Rv0471c	Rv0471c		
Rv0424c				
				Rv0426c
	Rv0455c			Rv0455c
			Rv0471c	
	Rv0499		Rv0499	
Rv0518			Rv0518	
				Rv0521c
Rv0558	Rv0558			
Rv0660c		Rv0660c		
	Rv0612	Rv0612		
Rv0613c				
			Rv0616c	Rv0616c
Rv0633c		Rv0633c		
	Rv0636		Rv0636	
	Rv0662c	Rv0662c		
		Rv0689c	Rv0689c	
		Rv0690c	Rv0690c	
	Rv0699		Rv0699	
Rv0743c				Rv0743c
Rv0756c	Rv0756c			
	Rv0787		Rv0787	
			Rv0789c	Rv0789c
				Rv0810c
		Rv0836c		Rv0836c
			Rv0837c	Rv0837c
			Rv0898c	Rv0898c
		Rv0900		
		Rv0909	Rv0909	
		Rv0942		Rv0942
	Rv0979c	Rv0979c		
			Rv1006	Rv1006
	Rv1012		Rv1012	
				Rv1045
		Rv1046c		Rv1046c
			Rv1048c	
	Rv1052	Rv1052		
		Rv1053c	Rv1053c	
	Rv1060		Rv1060	
	Rv1109c		Rv1109c	
Rv1111c		Rv1111c		
Rv1116A				
	Rv1118c			
		Rv1119c	Rv1119c	
Rv1125		Rv1125		
	Rv1134			Rv1134
			Rv1137c	
	Rv1154c		Rv1154c	
	Rv1155			Rv1155
Rv1171			Rv1171	
		Rv1203c		
				Rv1222

Rv1233c				
			Rv1265	Rv1265
	Rv1289			
	Rv1351			
		Rv1356c		
Rv1366			Rv1366	
			Rv1374c	Rv1374c
		Rv1417		
				Rv1419
	Rv1434	Rv1434		
			Rv1439c	
	Rv1443c		Rv1443c	
			Rv1444c	
			Rv1499	Rv1499
Rv1502	Rv1502			
			Rv1535	Rv1535
			Rv1545	
Rv1632c		Rv1632c		
			Rv1669	
	Rv1693	Rv1693		
		Rv1724c		
				Rv1735c
			Rv1742	
			Rv1748	
	Rv1752		Rv1752	
	Rv1762c	Rv1762c		
		Rv1772		Rv1772
			Rv1778c	Rv1778c
				Rv1805c
				Rv1831
		Rv1873		Rv1873
		Rv1890c	Rv1890c	
				Rv1907c
		Rv1914c		
	Rv1924c			Rv1924c
	Rv1927	Rv1927		
				Rv1929c
	Rv1944c		Rv1944c	
		Rv1947		
	Rv1948c		Rv1948c	
				Rv1954c
				Rv1955
				Rv1958c
			Rv1961	Rv1961
			Rv1975	
				Rv1989c
				Rv1995
			Rv2000	Rv2000
				Rv2016
			Rv2019	Rv2019
			Rv2020c	Rv2020c
			Rv2023c	Rv2023c
				Rv2035
	Rv2044c	Rv2044c		
Rv2049c	Rv2049c			
Rv2074		Rv2074		
	Rv2075c	Rv2075c		
		Rv2076c	Rv2076c	
			Rv2078	
	Rv2079	Rv2079		
Rv2084	Rv2084			
			Rv2160c	Rv2160c
Rv2189c	Rv2189c			

	Rv2197c			
		Rv2203		Rv2203
		Rv2219		
	Rv2227		Rv2227	
			Rv2235	Rv2235
		Rv2255c		
				Rv2269c
	Rv2271			
	Rv2275			Rv2275
			Rv2283	
Rv2288			Rv2288	
Rv2297			Rv2297	
		Rv2304c	Rv2304c	
				Rv2306A
		Rv2311	Rv2311	
		Rv2312		Rv2312
	Rv2313c		Rv2313c	
Rv2331A				
	Rv2336	Rv2336		
Rv2337c	Rv2337c			
	Rv2342			Rv2342
		Rv2348c	Rv2348c	
Rv2360c		Rv2360c		
			Rv2369c	Rv2369c
	Rv2401	Rv2401		
Rv2418c	Rv2418c			
	Rv2422		Rv2422	
		Rv2432c	Rv2432c	
	Rv2433c		Rv2433c	
			Rv2452c	
Rv2468c				
			Rv2472	
Rv2474c	Rv2474c			
	Rv2481c	Rv2481c		
				Rv2516c
				Rv2517c
		Rv2561	Rv2561	
		Rv2562		Rv2562
			Rv2570	Rv2570
Rv2574				
		Rv2576c	Rv2576c	
	Rv2598	Rv2598		
			Rv2616	Rv2616
				Rv2628
				Rv2629
				Rv2630
			Rv2633c	Rv2633c
Rv2635			Rv2635	
			Rv2644c	Rv2644c
	Rv2645		Rv2645	
		Rv2661c	Rv2661c	
				Rv2662
				Rv2663
				Rv2664
				Rv2705c
				Rv2706c
		Rv2722		
			Rv2730	
		Rv2735c	Rv2735c	
Rv2762c				Rv2762c
				Rv2774c
		Rv2803c		
			Rv2804c	

	Rv2809		Rv2809	
		Rv2826c	Rv2826c	
	Rv2826c		Rv2826c	
Rv2901c				
	Rv2960c			
			Rv2990c	Rv2990c
			Rv3033	Rv3033
			Rv3047c	
Rv3091				
	Rv3096		Rv3096	
		Rv3098c	Rv3098c	
Rv3099c		Rv3099c		
		Rv3108		Rv3108
	Rv3123			Rv3123
				Rv3126c
			Rv3142c	Rv3142c
			Rv3172c	
	Rv3190c		Rv3190c	
Rv3210c	Rv3210c			
		Rv3243c	Rv3243c	
		Rv3259	Rv3259	
			Rv3294c	
		Rv3378c		
		Rv3403c		Rv3403c
				Rv3424c
Rv3479		Rv3479		
	Rv3489	Rv3489		
		Rv3491	Rv3491	
			Rv3527	Rv3527
	Rv3528c		Rv3528c	
			Rv3572	Rv3572
Rv3599c			Rv3599c	
				Rv3612c
			Rv3613c	Rv3613c
				Rv3642c
		Rv3643	Rv3643	
Rv3651			Rv3651	
		Rv3654c		Rv3654c
	Rv3657c	Rv3657c		
Rv3662c	Rv3662c			
	Rv3705c		Rv3705c	
Rv3718c	Rv3718c			
Rv3722c		Rv3722c		
		Rv3749c		Rv3749c
Rv3753c	Rv3753c			
	Rv3755c			
Rv3766				Rv3766
	Rv3768			
Rv3780	Rv3780			
			Rv3819	Rv3819
	Rv3831	Rv3831		
		Rv3839		Rv3839
		Rv3845		Rv3845
	Rv3847			Rv3847
	Rv3850	Rv3850		
	Rv3861		Rv3861	
				Rv1817
			Rv0314c	
	Rv0492c		Rv0492c	
			Rv1103c	Rv1103c
		Rv1179c	Rv1179c	
	Rv1424c			
				Rv1510

	Rv2274c	Rv2274c		
	Rv2740			Rv2740
Rv2767c		Rv2767c		
			Rv1576c	Rv1576c
			Rv3840	Rv3840
			Rv1291c	Rv1291c
	Rv2811		Rv2811	
		Rv3654c		Rv3654c
			Rv2369c	Rv2369c
	Rv2227		Rv2227	
				Rv2269c
Rv2513	Rv2513			
		Rv3278c		
Rv2513	Rv2513			
				Rv2660c
	Rv0029	Rv0029		
Rv0740				Rv0740
Rv1125		Rv1125		
53	90	95	116	108

Supplement Table S2
Essentiality of proteins

Rv ID	Essentiality
Rv0029	Yes
Rv0030	Yes
Rv0241c	Yes
Rv0257c	Yes
Rv0314c	Yes
Rv0390	Yes
Rv0492c	yes
Rv0499	yes
Rv0558	yes
Rv0636	yes
Rv0740	yes
Rv0756c	yes
Rv0836c	yes
Rv0837c	yes
Rv1116A	yes
Rv1154c	yes
Rv1155	yes
Rv1233c	yes
Rv1291c	yes
Rv1510	yes
Rv1752	yes
Rv1817	Yes
Rv1873	yes
Rv1927	yes
Rv1995	yes
Rv2079	yes
Rv2235	yes
Rv2306A	yes
Rv2616	yes
Rv2633c	yes
Rv2705c	yes
Rv2740	yes
Rv2803c	yes
Rv2811	yes
Rv2990c	yes
Rv3096	yes
Rv3278c	yes
Rv3294c	yes
Rv3403c	yes
Rv3479	yes
Rv3722c	yes
Rv3749c	yes
Rv3755c	yes
Rv3768	yes

Supplement Table S3
Pfam entries of some proteins

Rv ID	Pfam ID
Rv0088	PF10604 Polyketide_cyc2
Rv0241c	PF01575 MaoC_dehydratas
Rv0332	PF07398 DUF1503
Rv0499	PF00781 DAGK_cat
Rv0558	PF01209 Ubie_methyltran
Rv0613c	PF02810 SEC-C
Rv0636	PF01575 MaoC_dehydratas
Rv0837c	PF09952 DUF2186
Rv1045	PF08843 DUF1814
Rv1060	PF10604 Polyketide_cyc2
Rv1125	PF03007 UPF0089
Rv1154c	PF08002 DUF1697
Rv1155	PF01243 Pyridox_oxidase
Rv1366	PF04607 RelA_SpoT
Rv1417	PF10756 DUF2581
Rv1419	PF00652 Ricin_B_lectin
Rv1443c	PF10604 Polyketide_cyc2
Rv1632c	PF04167 DUF402
Rv1772	PF03861 ANTAR
Rv1927	PF09844 DUF2071
Rv1944c	PF02810 SEC-C
Rv1955	PF05973 Gp49
Rv1989c	PF08808 RES
Rv1995	PF01814 Hemerythrin
Rv2035	PF08327 AHSA1
Rv2074	PF01243 Pyridox_oxidase
Rv2079	PF06259 DUF1023
Rv2227	PF09859 DUF2086
Rv2570	PF04944 DUF661
Rv2574	PF10604 Polyketide_cyc2
Rv2616	PF09348 DUF1990
Rv2633c	PF01814 Hemerythrin
Rv2705c	PF06108 DUF952
Rv2762c	PF06224 DUF1006
Rv2803c	PF08681 DUF1778
Rv2826c	PF08843 DUF1814
Rv2827c	PF09407 DUF2005
Rv2901c	PF10611 DUF2469
Rv3091	PF01734 Patatin
Rv3479	PF01734 Patatin
Rv3657c	PF00482 GSPII_F
Rv3718c	PF10604 Polyketide_cyc2
Rv3755c	PF06475 DUF1089
Rv3780	PF10759 DUF2587
Rv1817	PF00890 FAD_binding_2
Rv0492c	PF00732GMC_oxred_N PF01266DAO PF05199 GMC_oxred_C
Rv1179c	PF04851 ResIII
Rv2740	PF07858 LEH
Rv3278c	PF03703 DUF304
Rv1576c	PF05065 Phage_capsid
Rv1291c	PF05305 DUF732

Supplement Table *S4 Gene Ontology functions*		
Rv ID	**GO**	**Functional Class**
Rv0123	**GO:0045449 regulation of transcription**	
Rv0241c	**GO:0005835 fatty acid synthase complex** **GO:0004312 fatty-acid synthase activity** **GO:0016491 oxidoreductase activity** **GO:0008152 metabolic process** **GO:0006633 fatty acid biosynthetic process**	**GO:0008150** biological_process **GO:0009987** cellular process **GO:0016491** oxidoreductase activity **GO:0016740** transferase activity
Rv0499	**GO:0004143 diacylglycerol kinase activity** **GO:0007205 activation of protein kinase C activity**	**GO:0007165** signal transduction **GO:0016301** kinase activity **GO:0016772** transferase activity, transferring phosphorus-containing groups
Rv0518	**GO:0016787 hydrolase activity** **GO:0006629 lipid metabolic process**	**GO:0008150** biological_process **GO:0016787** hydrolase activity
Rv05580c	**GO:0008168 methyltransferase activity**	
Rv0636	**GO:0016491 oxidoreductase activity** **GO:0008152 metabolic process**	**GO:0008150** biological_process **GO:0009987** cellular process **GO:0016491** oxidoreductase activity **GO:0016740** transferase activity
Rv0662c	**GO:0045449 regulation of transcription**	
Rv1155	**GO:0010181 FMN binding**	**GO:0005488** binding
Rv1366	**GO:0015969 guanosine tetraphosphate metabolic process**	**GO:0009987** cellular process
Rv2035	**GO:0006950 response to stress**	
Rv2074	**GO:0010181 FMN binding**	**GO:0005488** binding
Rv2075c	**GO:0004629 phospholipase C activity** **GO:0008081 phosphoric diester hydrolase activity** **GO:0005488 binding** **GO:0007242 intracellular signaling cascade** **GO:0006629 lipid metabolic process**	**GO:0007165** signal transduction **GO:0016787** hydrolase activity
Rv2160c	**GO:0030528 transcription regulator activity** **GO:0045449 regulation of**	

	transcription	
Rv2516c	**GO:0000166 nucleotide binding**	
Rv2660c		**GO:0003676** nucleic acid binding **GO:0009987** cellular process **GO:0030528** transcription regulator activity
Rv3091	**GO:0006629 lipid metabolic process** **GO:0008152 metabolic process**	**GO:0008150** biological_process
Rv3096	**GO:0003824 catalytic activity** **GO:0043169 cation binding** **GO:0005975 carbohydrate metabolic process**	
Rv3210c	**GO:0003700 transcription factor activity** **GO:0046914 transition metal ion binding** **GO:0016491 oxidoreductase activity**	**GO:0003676** nucleic acid binding **GO:0009987** cellular process **GO:0030528** transcription regulator activity
Rv3479	**GO:0006629 lipid metabolic process** **GO:0008152 metabolic process**	**GO:0008150** biological_process
Rv3722c	**GO:0003824 catalytic activity** **GO:0030170 pyridoxal phosphate binding**	
Rv1817	**GO:0009055 electron carrier activity** **GO:0016491 oxidoreductase activity**	**GO:0009987** cellular process **GO:0016491** oxidoreductase activity
Rv0492c	**GO:0016491 oxidoreductase activity** **GO:0050660 FAD binding**	**GO:0005488** binding **GO:0009987** cellular process **GO:0016491** oxidoreductase activity
Rv1103c	**GO:0045449 regulation of transcription**	
Rv1179c	**GO:0003677 DNA binding** **GO:0016787 hydrolase activity** **GO:0005524 ATP binding**	**GO:0003676** nucleic acid binding **GO:0004386** helicase activity **GO:0005488** binding

Table S5
Rvs interactome

Rv ID	Genes Interact with
Rv 0008c	not found
Rv0093c	not found
Rv0210	not found
Rv0426c	not found
Rv0455c	not found
Rv0900	not found
Rv1111c	not found
Rv1171	not found
Rv1233c	364 proteins
Rv1417	not found
Rv1735c	not found
Rv0633c	Rv0054
Rv1748	not found
Rv1924c	not found
Rv2044c	not found
Rv2197c	2 proteins
Rv2203	5 proteins
Rv2219	not found
Rv2235	not found
Rv2306A	not found
Rv2433c	not found
Rv2452c	not found
Rv2576c	54 proteins
Rv3479	Rv3763
Rv3491	29 proteins
Rv3657c	not found
Rv3831	not found
Rv0314c	112 proteins
Rv1510	not found
Rv2767c	not found
Rv3278c	not found
Rv1291c	7 proteins
Rv0088	3 proteins
Rv0241c	21 proteins
Rv0332	2 proteins
Rv0499	11 proteins
Rv0558	10 proteins
Rv0613c	3 proteins
Rv0636	38 proteins
Rv0837c	3 proteins
Rv1045	12 proteins
Rv1060	not found
Rv1125	not found
Rv1154c	30 proteins
Rv1155	39 proteins
Rv1366	3 proteins
Rv1419	7 proteins
Rv1632c	not found
Rv1772	3 proteins
Rv1927	3 proteins
Rv1944c	3 proteins
Rv1955	7 proteins
Rv1989c	4 proteins
Rv1995	72 proteins
Rv2035	not found
Rv2074	4 proteins
Rv2079	34 proteins
Rv2227	not found
Rv2570	Rv3763
Rv2574	Rv3763
Rv2616	8 proteins
Rv2633c	5 proteins
Rv2705c	Rv3763
Rv2762c	not found
Rv2803c	not found
Rv2826c	not found
Rv2827c	not found
Rv2901c	Rv3763
Rv3091	4 proteins
Rv3718c	47 proteins
Rv3755c	not found
Rv3780	9 proteins
Rv0492c	not found
Rv1179c	3 proteins
Rv2740	4 proteins
Rv3840	not found
Rv0098	9 proteins
Rv0390	3 proteins
Rv1873	3 proteins
Rv0295c	7 proteins
Rv0699	not found
Rv3612c	not found
Rv1961	not found
Rv1890c	8 proteins
Rv0122	3 proteins
Rv0636	not found
Rv2020c	not found
Rv2288	not found
Rv0029	3 proteins
Rv0030	not found
Rv0257c	not found

Rv0756c	not found
Rv0836c	36 proteins
Rv1116A	not found
Rv1752	3 proteins
Rv2990c	not found
Rv3096	not found
Rv3294c	not found
Rv3403c	18 proteins
Rv3722c	11 proteins
Rv3749c	Rv3763
Rv3768	22 proteins
Rv1817	11 proteins
Rv0740	3 proteins
Rv2811	not found
Rv0123	16 proteins
Rv0518	not found
Rv0662c	3 proteins
Rv2075c	3 proteins
Rv2160c	not found
Rv2516c	25 proteins
Rv2660c	not found
Rv3210c	40 proteins
Rv0492c	not found
Rv1103c	not found
Rv0272c	3 proteins
Rv0690c	5 proteins
Rv0743c	Rv2507
Rv1502	3 proteins
Rv3337c	not found
Rv0028	not found
Rv0057	3 proteins
Rv0061	not found
Rv0079	not found
Rv0080	not found
Rv0108c	not found
Rv0193c	15 proteins
Rv0209	3 proteins
Rv0299	not found
Rv0307c	not found
Rv0311	47 proteins
Rv0313	18 proteins
Rv0333	Rv0054
Rv0340	11 proteins
Rv0358	not found
Rv0360c	not found
Rv0381c	4 proteins
Rv0390	not found
Rv0395	not found
Rv0471c	4 proteins
Rv0424c	3 proteins
Rv0660c	not found
Rv0612	not found
Rv0616c	not found
Rv0689c	not found
Rv0787	86 proteins
Rv0789c	7 proteins
Rv0810c	not found
Rv0898c	908 proteins
Rv0909	365 proteins
Rv0942	not found
Rv0979c	not found
Rv1006	8 proteins
Rv1012	Rv2507

Rv1046c	13 proteins
Rv1048c	8 proteins
Rv1052	75 proteins
Rv1053c	5 proteins
Rv1109c	4 proteins
Rv1118c	not found
Rv1119c	not found
Rv1134	100 proteins
Rv1137c	6 proteins
Rv1203c	not found
Rv1265	3 proteins
Rv1289	3 proteins
Rv1351	not found
Rv1374c	not found
Rv1434	not found
Rv1439c	not found
Rv1443c	3 proteins
Rv1444c	62 proteins
Rv1499	4 proteins
Rv1545	not found
Rv1693	not found
Rv1724c	16 proteins
Rv1742	3 proteins
Rv1762c	18 proteins
Rv1778c	3 proteins
Rv1805c	not found
Rv1831	3 proteins
Rv1907c	not found
Rv1914c	not found
Rv1947	25 proteins
Rv1948c	37 proteins
Rv1954c	3 proteins
Rv1958c	3 proteins
Rv1975	284 proteins
Rv2000	7 proteins
Rv2016	4 proteins
Rv2019	not found
Rv2023c	not found
Rv2049c	not found
Rv2076c	not found
Rv2078	2 proteins
Rv2084	not found
Rv2189c	not found
Rv2255c	2 proteins
Rv2269c	not found
Rv2271	not found
Rv2275	29 proteins
Rv2283	53 proteins
Rv2288	not found
Rv2297	11 proteins
Rv2304c	not found
Rv2311	not found
Rv2312	Rv2507
Rv2313c	26 proteins
Rv2331A	not found
Rv2336	78 proteins
Rv2337c	2 proteins
Rv2342	not found
Rv2348c	3 proteins
Rv2369c	not found
Rv2401	not found
Rv2418c	6 proteins
Rv2422	not found

Rv2432c	Rv2507
Rv2468c	
Rv2472	not found
Rv2474c	2 proteins
Rv2481c	238 roteins
Rv2517c	not found
Rv2561	not found
Rv2562	not found
Rv2598	not found
Rv2628	not found
Rv2629	9 proteins
Rv2630	131 proteins
Rv2635	not found
Rv2644c	not found
Rv2645	not found
Rv2661c	not found
Rv2662	323 roteins
Rv2663	5 proteins
Rv2664	3 proteins
Rv2706c	not found
Rv2722	not found
Rv2730	60 proteins
Rv2735c	Rv3763
Rv2774c	Rv3763
Rv2804c	2 proteins
Rv2809	Rv0054
Rv2826c	not found
Rv2960c	2 proteins
Rv3033	46 proteins
Rv3047c	4 proteins
Rv3098c	not found
Rv3099c	10 proteins
Rv3108	not found
Rv3123	not found

Rv3126c	27 proteins
Rv3142c	3 proteins
Rv3172c	Rv3763
Rv3190c	2 proteins
Rv3243c	8 proteins
Rv3259	2 proteins
Rv3378c	4 proteins
Rv3424c	not found
Rv3489	not found
Rv3527	2 proteins
Rv3528c	Rv3763
Rv3572	282 proteins
Rv3599c	not found
Rv3613c	not found
Rv3642c	not found
Rv3643	not found
Rv3651	6 proteins
Rv3654c	not found
Rv3662c	not found
Rv3705c	13 proteins
Rv3753c	57 proteins
Rv3766	205 proteins
Rv3819	not found
Rv3839	not found
Rv3845	not found
Rv3847	11 proteins
Rv3850	not found
Rv3861	not found
Rv1424c	10 proteins
Rv2274c	not found
Rv1576c	not found
Rv2513	29 proteins

Supplement Table S6
Categorization of some hypothetical proteins among different COGs

TubercuList ID	TDR ID	Functional Description	Orthologous group
Rv0241c	7249	Probable 3-Hydroxyacyl-Thioester Dehydratase Htdx	OG4_53689
Rv0272c	7265	Hypothetical Protein	OG4_89725
Rv0390	7344	Conserved Hypothetical	OG4_49855
Rv0499	7406	Conserved Hypothetical	OG4_13082
Rv0558	6449	PROBABLE UBIQUINONE/MENAQUINONE BIOSYNTHESIS METHYLTRANSFERASE MENH (2-Heptaprenyl-1,4-Naphthoquinone Methyltransferase)	OG4_10696
Rv0636	7495	(3r)-Hydroxyacyl-Acp Dehydratase Subunit Hadb	OG4_14471
Rv0690c	7522	Conserved Hypothetical Protein	OG4_71206
Rv0743c	7541	Conserved Hypothetical Protein	OG4_113959
Rv0756c	7551	Hypothetical Protein	OG4_83843
Rv0836c	7600	Hypothetical Protein	OG4_113963
Rv0837c	7601	Hypothetical Protein	OG4_113964
Rv1111c	7754	Conserved Hypothetical Protein	OG4_17718
Rv1116A	7760	Conserved Hypothetical Protein (Fragment)	OG4_15389
Rv1154c	7785	Hypothetical Protein	OG4_30558
Rv1155	7786	Possible Pyridoxamine 5'-Phosphate Oxidase (Pnp/Pmp Oxidase) (Pyridoxinephosphate Oxidase) (Pnpox) (Pyridoxine 5'-Phosphate Oxid	OG4_83845
Rv1233c	7827	Conserved Hypothetical Membrane Protein	OG4_11008
Rv1502	7977	Hypothetical Protein	OG4_113981
Rv1735c	8109	Hypothetical Membrane Protein	OG4_13946
Rv1752	8120	Conserved Hypothetical Protein	OG4_11530
Rv1873	8191	Conserved Hypothetical Protein	OG4_78781
Rv1927	8223	Conserved Hypothetical Protein	OG4_63041
Rv1995	8271	Hypothetical Protein	OG4_39793
Rv2075c	8327	Possible Hypothetical Exported Or Envelope Protein	OG4_35068
Rv2079	8332	Conserved Hypothetical Protein	OG4_42701
Rv2227	8416	Conserved Hypothetical Protein	OG4_60147
Rv2235	8421	Probable Conserved Transmembrane Protein	OG4_12804
Rv2306A	8471	Possible Conserved Membrane Protein	OG4_49877
Rv2616	8642	Conserved Hypothetical Protein	OG4_27623
Rv2633c	8658	Hypothetical Protein	OG4_44079
Rv2705c	8712	Conserved Hypothetical Protein	OG4_21474
Rv2803c	8775	Conserved Hypothetical Protein	OG4_54642
Rv2990c	8876	Hypothetical Protein	OG4_83862
Rv3096	8935	Conserved Hypothetical Protein	OG4_14476
Rv3294c	9051	Conserved Hypothetical Protein	OG4_37894
Rv3378c	9044	Probable Conserved Tansmembrane Protein	OG4_100228
Rv3403c	9110	Hypothetical Protein	OG4_33814
Rv3479	9147	Possible Tansmembrane Protein	OG4_94776
Rv3722c	9736	Conserved Hypothetical Protein	OG4_103827
Rv3749c	9757	Conserved Hypothetical Protein	OG4_71286
Rv3755c	9762	Conserved Hypothetical Protein	OG4_52096
Rv3768	9767	Hypothetical Protein	OG4_111106
Rv1817	8159	Possible Flavoprotein	OG4_40127
Rv0029	7117	Conserved Hypothetical Protein	OG4_59409
Rv0314c	7295	Possible Conserved Membrane Protein	OG4_83838
Rv0492c	7399	Probable Oxidoreductase	OG4_16996

Rv0740	7541	Conserved Hypothetical Protein	OG4_113959
Rv1510	7987	Conserved Probable Membrane Protein	OG4_113982
Rv2740	8738	Epoxide Hydrolase	OG4_83860
Rv3278c	9044	Probable Conserved Transmembrane Protein	OG4_100228
Rv1576c	8021	Probable Phirv1 Phage Protein	OG4_30549
Rv1576c	8579	Hypothetical Protein	OG4_30549
Rv3840	9816	Possible Transcriptional Regulatory Protein	OG4_104178
Rv1291c	7862	Conserved Hypothetical Secreted Protein	OG4_42702
Rv2811	8783	Conserved Hypothetical Protein	OG4_83859