Clin. Lab. 2018;64:XXX-XXX

ORIGINAL ARTICLE

Pathogenic Gene Screening of *Mycobacterium tuberculosis* by Literature Data Mining and Information Pathway Enrichment Analysis

Guangyu Xu [1, 2, *], Simin Wen [1, *], Yuchen Pan [1], Nan Zhang [3], Yuanyi Wang [1]

[*] *Guangyu Xu and Simin Wen contributed equally to this study*
[1] *Department of Clinical Research, The First Hospital of Jilin University, Changchun, China*
[2] *College of Pharmacy, Beihua University, Jilin, Jilin, China*
[3] *Department of Emergency Medicine, The First Hospital of Jilin University, Changchun, China*

SUMMARY

Background: **Recent studies have unraveled mutations which have led to changes in the original conformation of functional proteins targeted by frontline drugs against *Mycobacterium tuberculosis*. These mutations are likely responsible for the emergence of drug-resistant strains of *M. tuberculosis*. Identification of new therapeutic targets is fundamental to the development of novel anti-TB drugs.**
Methods: **Boost evolution analysis of interactome data with use of high-throughput biological experimental technologies provides opportunities for identification of pathogenic genes and for screening out novel therapeutic targets.**
Results: **In this study, we identified 584 proven pathogenic genes of *M. tuberculosis* and new pathogenic genes *via* bibliometrics and relevant websites such as PubMed, KEGG, and DOOR websites. We identified 13 new genes that are most likely to be pathogenic.**
Conclusions: **This study may contribute to the discovery of new pathogenic genes and help unravel new functions of known pathogenic genes of *M. tuberculosis*.**
(Clin. Lab. 2018;64:xx-xx. DOI: 10.7754/Clin.Lab.2018.170935)

Correspondence:
Nan Zhang
Department of Emergency Medicine
The First Hospital of Jilin University
Changchun, 130021
China
Phone/Fax: +86 43185619574
Email: zn0972@163.com

Yuanyi Wang
Department of Clinical Research
The First Hospital of Jilin University
Changchun, 130021
China
Phone/Fax: +86 43185619574
Email: tedwangyy@foxmail.com

Manuscript accepted January 1, 2018

KEY WORDS

Mycobacterium tuberculosis, pathogenic genes, pathway, operon analysis, bibliometrics

INTRODUCTION

Tuberculosis (TB) is one of the most frequently encountered health crises among human population, which can be traced back to ancient ages [1-3]. The development of combination therapies against TB has benefitted countless patients. However, the spread of drug resistant strains, such as multi-drug resistant TB and extensive-drug resistant TB [4-6], has challenged the traditional therapies. Since the discovery of rifampicin in 1963, new effective anti-TB drugs are rarely found. This can be attributed to the current functional genome of *Mycobacterium tuberculosis* as well as the lack of signal

transmission network annotation. This has led to lack of effective control of *Mycobacterium tuberculosis* and inadequate TB prevention and control [7]. Therefore, it is necessary to identify new targets in *M. tuberculosis* for developing new anti-TB drugs.

Generally, the genes responsible for the disease are called "pathogenic genes". Thus screening of pathogenic genes is very vital, as drugs generally target pathogenic genes [8]. Traditional screening methods are based on single-gene screening, a process which relies on the accuracy of detection equipment and is time and labor-consuming [8]. *Mycobacterium tuberculosis* H37Rv is the main pathogen of tuberculosis. With the development of molecular biology, the determination of *Mycobacterium tuberculosis* H37Rv gene sequence has been completed. Studying the pathogenic genes of *Mycobacterium tuberculosis* H37Rv will help to understand the molecular mechanism of tuberculosis and provide a more specific research direction on the clinical diagnosis and treatment of tuberculosis and the research of new drugs. With rapid development of high-throughput bio-experimental and bio-informatics technology, increasing numbers of potential pathogenic genes and their internal interactions have been identified based on the theoretical pathways. Further, functional analysis has helped confirm the relationship between the predicted pathogenic genes and diseases [9].

In the current study, we compiled all identified pathological genes of *M. tuberculosis* based on studies from PubMed and found 584 proven pathogenic genes in *M. tuberculosis.* We analyzed the relevance of pathogenic genes of *M. tuberculosis* with respect to pathways, operons, and gene functions on the website of KEGG and DOOR. Finally, we inferred 13 potential pathogenic genes of *M. tuberculosis*. This study may contribute to the discovery of new pathogenic genes and help unravel new functions of known pathogenic genes of *M. tuberculosis*.

MATERIALS AND METHODS

Bibliometrics

We searched the keyword "*M. tuberculosis* and pathogenic genes" on the online biomedical database PubMed. The reference period for the literature search was from 2001 to 2016. Duplicate publications and similar literatures were excluded using Epidata 3.1 software. A total of 1,527 relevant publications were retrieved, out of which only 232 publications showed a direct association and were eventually included in the analysis [10, 11]. A total of 751 literatures were retrieved and 232 literatures were used to analysis the pathogenic genes of H37Rv.

KEGG analysis and data resource

Each pathway was analyzed on KEGG (http://www.genome.jp/kegg/), in which 125 relevant pathways of *M. tuberculosis* were identified [12].

Operon analysis and data resource

The database of operons was downloaded (June 25, 2016) from http://csbl.bmb.uga.edu/DOOR. DOOR2 (database of prOkaryoticOpeRons, Version 2.0) is an operon database developed by calculation system biological laboratory (CSBL) of the University of Georgia. Operons of this database were predicted based on the fundamental genome characters. The algorithm of the operon database incorporates a data miner and a classifier; its characters include gene distance, neighborhood conservatism, phylogenetic distance, information from short DNA motifs, similarity scales of the gene ontology (GO) items between the gene pairs, and the length ratio of gene pairs [13].

Functional analysis based on Pfam database

In the present study, gene function was analyzed mainly through the Pfam database. The Pfam database (version 2.0) was downloaded (June 25, 2016) from http://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/. The Pfam database is an aggregation of massive protein families and consists of Pfamls, Pfamfs, and others.

RESULTS AND DISCUSSION

H37Rv Pathogenic genes and their associated pathways

We found 584 proven pathogenic genes of *M. tuberculosis* H37Rv using bibliometrics, which constitutes 15% of the whole genome (4,008). Among these, 125 genes were uncovered by KEGG website (Table 1) which made up 21% of all mining discoveries. Nine pathogenic genes correlated to ten pathways. *Rv0859*, *Rv3546*, *Rv1142c*, and *Rv1141c* correlated to 18 pathways (Table 1).

We found the involvement of these pathogenic genes in 79 pathways. Notably, at least 4 genes function in 27 pathways, 10 pathogenic genes in another 9 pathways, and 20 pathogenic genes in the last 5 pathways (Table 2), such as mtu01100 (57 genes), mtu02010 (28 genes), mtu01110 (23 genes), mtu01130 (21 genes), and mtu01120 (21 genes).

Mtu01110 and mtu01120 are common pathways of microbes. Mtu01110 is the pathway of biosynthesis of secondary metabolites, while mtu01120 pertains to microbial metabolism in diverse environments. Pathway mtu01100 correlates to 57 genes due to its association to glycan biosynthesis and metabolism, which exists in almost all kinds of microbes. This finding does explain the mechanism of pathogenesis and indicates that a large number of pathogenic genes are involved in the metabolism process. The proteins in Mtu02101 are related to ABC transporters, which are membrane proteins involved in transportation of compounds through membrane structures against the concentration gradient by utilizing the energy of ATP hydrolysis [14]. ABC proteins are able to transport ions, amino acids, carbohydrates, vitamins, peptides, polysaccharides, hormones,

Table 1. Pathogenic genes of *M. tuberculosis* H37Rv in more than 5 pathways.

Locus tag	Number	Pathway
Rv0859	18	mtu00071/mtu00072/mtu00280/mtu00310/mtu00362/mtu00380/mtu00620/mtu00630/mtu00640/mtu00650/mtu00900/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200/mtu01212/mtu02020
Rv3546	18	mtu00071/mtu00072/mtu00280/mtu00310/mtu00362/mtu00380/mtu00620/mtu00630/mtu00640/mtu00650/mtu00900/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200/mtu01212/mtu02020
Rv1142c	18	mtu00071/mtu00280/mtu00281/mtu00310/mtu00360/mtu00362/mtu00380/mtu00410/mtu00627/mtu00640/mtu00650/mtu00903/mtu00930/mtu01100/mtu01110/mtu01120/mtu01130/mtu01212
Rv1141c	18	mtu00071/mtu00280/mtu00281/mtu00310/mtu00360/mtu00362/mtu00380/mtu00410/mtu00627/mtu00640/mtu00650/mtu00903/mtu00930/mtu01100/mtu01110/mtu01120/mtu01130/mtu01212
Rv3303c	12	mtu00010/mtu00020/mtu00260/mtu00280/mtu00620/mtu00630/mtu00640/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200
Rv3285	11	mtu00061/mtu00280/mtu00620/mtu00630/mtu00640/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200/mtu01212
Rv2247	11	mtu00061/mtu00280/mtu00620/mtu00630/mtu00640/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200/mtu01212
Rv0650	10	mtu00010/mtu00052/mtu00500/mtu00520/mtu00521/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200
Rv0363c	10	mtu00010/mtu00030/mtu00051/mtu00680/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200/mtu01230
Rv1617	9	mtu00010/mtu00230/mtu00620/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200/mtu01230
Rv3339c	9	mtu00020/mtu00480/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200/mtu01210/mtu01230
Rv3509	9	mtu00290/mtu00650/mtu00660/mtu00770/mtu01100/mtu01110/mtu01130/mtu01210/mtu01230
Rv0904c	9	mtu00061/mtu00620/mtu00640/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200/mtu01212
Rv2220	8	mtu00220/mtu00250/mtu00630/mtu00910/mtu01100/mtu01120/mtu01230/mtu02020
Rv1449c	7	mtu00030/mtu01100/mtu01110/mtu01120/mtu01130/mtu01200/mtu01230
Rv1285	7	mtu00230/mtu00261/mtu00450/mtu00920/mtu01100/mtu01120/mtu01130
Rv3280	7	mtu00280/mtu00630/mtu00640/mtu01100/mtu01120/mtu01130/mtu01200
Rv2351c	6	mtu00562/mtu00564/mtu00565/mtu01100/mtu01110/mtu02024
Rv2350c	6	mtu00562/mtu00564/mtu00565/mtu01100/mtu01110/mtu02024
Rv2349c	6	mtu00562/mtu00564/mtu00565/mtu01100/mtu01110/mtu02024
Rv3711c	6	mtu00230/mtu00240/mtu01100/mtu03030/mtu03430/mtu03440
Rv2191	6	mtu00230/mtu00240/mtu01100/mtu03030/mtu03430/mtu03440
Rv3283	5	mtu00270/mtu00920/mtu01100/mtu01120/mtu04122
Rv0467	5	mtu00630/mtu01100/mtu01110/mtu01120/mtu01200
Rv1350	5	mtu00061/mtu00780/mtu01040/mtu01100/mtu01212
Rv0242c	5	mtu00061/mtu00780/mtu01040/mtu01100/mtu01212

lipids, and xenobiotic matters. As the cell gatekeepers, these can also maintain nutrients and eliminate toxicants from the cells, which otherwise may result in diseases. Pathway mtu01130 is the pathway for biosynthesis of antibiotics and may be closely related to resistance against the current drugs for treatment of tuberculosis. Based on the analysis of other relevant pathways of pathogenic genes, we found that fatty acid biosynthesis might be regulated by the mtu00061 pathway (Table 3). Multiple anti-TB therapeutic targets have been discovered (including the mycolic acid synthetic pathway) based on the genome contrastive analysis of human and biosynthesis of antibiotics [15]. Therefore, genes that modulate pathway mtu00061 are a potential therapeutic target, which implies that all related genes are most likely pathogenic genes. So far, 15 related genes have been identified in this pathway, of which 10 have been confirmed to be pathogenic (Figure 1); functional research on the other 5 genes (*Rv2501c, Rv3502c, Rv3559c, Rv0769, Rv2187*) is still a work in progress. Mtu00550 pathway was found to be associated with peptidoglycan synthesis. The pathogenesis of *M. tuberculosis* infection is closely related to the cell wall [3]. Peptidoglycan is an essential component of cell wall structure and is a therapeutic target of anti-TB drugs; thus, we inferred that all genes in this pathway are rele-

Table 2. Pathways containing more than four pathogenic genes of *M. tuberculosis* H37Rv.

Pathway	Number of genes	Genes
mtu01100	57	*Rv0859/Rv3546/Rv1142c/Rv1141c/Rv3303c/Rv3285/Rv2247/Rv0363c/Rv0650/Rv3339c/Rv0904c/Rv3509/Rv1617/Rv2220/Rv1449c/Rv1285/Rv3280/Rv2191/Rv3711c/Rv2351c/Rv2350c/Rv2349c/Rv1350/Rv1350/Rv3283/Rv0467/Rv2157c/Rv1908c/Rv3382c/Rv1110/Rv0806c/Rv2524c/Rv2243/Rv0649/Rv2780/Rv3341/Rv2152c/Rv2155c/Rv1315/Rv0482/Rv1018c/Rv2153c/Rv3229c/Rv2156c/Rv1412/Rv0260c/Rv2392/Rv3561/Rv2202c/Rv1338/Rv2911/Rv3804c/Rv0129c/Rv1886c/Rv1695/Rv2965c/Rv2002*
mtu02010	28	*Rv0928/Rv3499c/Rv0169/Rv0589/Rv1966/Rv0167/Rv0168/Rv0170/Rv0171/Rv0172/Rv0174/Rv0587/Rv0588/Rv0591/Rv0592/Rv0594/Rv1964/Rv1965/Rv1967/Rv1968/Rv1969/Rv1971/Rv3494c/Rv3496c/Rv3497c/Rv3498c/Rv3499c/Rv3500c*
mtu01110	23	*Rv0859/Rv3546/Rv1142c/Rv1141c/Rv3303c/Rv3285/Rv2247/Rv0363c/Rv0650/Rv3339c/Rv0904c/Rv3509/Rv1617/Rv1449c/Rv2351c/Rv2350c/Rv2349c/Rv0467/Rv1908c/Rv3382c/Rv1110/Rv1412/Rv0260c*
mtu01130	21	*Rv0859/Rv3546/Rv1142c/Rv1141c/Rv3303c/Rv3285/Rv2247/Rv0363c/Rv0650/Rv3339c/Rv0904c/Rv3509/Rv1617/Rv1449c/Rv1285/Rv3280/Rv3382c/Rv1110/Rv3341/Rv1018c/Rv3332*
mtu01120	21	*Rv0859/Rv3546/Rv1142c/Rv1141c/Rv3285/Rv3303c/Rv2247/Rv3561/Rv0363c/Rv0650/Rv3339c/Rv0904c/Rv1617/Rv2220/Rv1449c/Rv1285/Rv3280/Rv3283/Rv0467/Rv0373c/Rv2392*
mtu02020	19	*Rv0859/Rv3546/Rv2220/Rv0928/Rv1027c/Rv3501c/Rv3502c/Rv3503c/Rv3504c/Rv3505c/Rv3506c/Rv3507c/Rv3508c/Rv3509c/Rv3510c/Rv3511c/Rv3512c/Rv3513c/Rv3514c*
mtu01200	14	*Rv0859/Rv3546/Rv3303c/Rv3285/Rv2247/Rv0363c/Rv0650/Rv3339c/Rv0904c/Rv1617/Rv1449c/Rv3280/Rv0467 Rv0373c*
mtu01212	14	*Rv0859/Rv3546/Rv1142c/Rv1141c/Rv3285/Rv2247/Rv0904c/Rv1350/Rv0242c/Rv2524c/Rv0824c/Rv1094/Rv2243/Rv0649*
mtu00550	10	*Rv2157c/Rv2152c/Rv2155c/Rv1315/Rv0482/Rv2153c/Rv2156c/Rv2158c/Rv2911/Rv3627c*
mtu00061	10	*Rv3285/Rv2247/Rv0904c/Rv1350/Rv0242c/Rv2524c/Rv0824c/Rv1094/Rv2243/Rv0649*
mtu00640	9	*Rv0859/Rv3546/Rv1142c/Rv1141c/Rv3303c/Rv3285/Rv2247/Rv0904c/Rv3280*
mtu00520	9	*Rv0650/Rv0806c/Rv1315/Rv0482/Rv1018c/Rv0113/Rv3809c/Rv0112/Rv3332*
mtu00630	8	*Rv0859/Rv3546/Rv3303c/Rv3285/Rv2247/Rv2220/Rv3280/Rv0467*
mtu00280	8	*Rv0859/Rv3546/Rv1142c/Rv1141c/Rv3303c/Rv3285/Rv2247/Rv3280*
mtu00620	7	*Rv0859/Rv3546/Rv3303c/Rv3285/Rv2247/Rv0904c/Rv1617*
mtu00230	6	*Rv1617/Rv1285/Rv2191/Rv3711c/Rv2202c/Rv2583c*
mtu01040	6	*Rv1350/Rv0242c/Rv0824c/Rv1094/Rv1618/Rv2605c*
mtu05152	6	*Rv0928/Rv0440/Rv0350/Rv3875/Rv0410c/Rv3763*
mtu01230	6	*Rv0363c/Rv3339c/Rv3509/Rv1617/Rv2220/Rv1449c*
mtu02024	5	*Rv2351c/Rv2350c/Rv2349c/Rv1440/Rv1027c*
mtu00650	5	*Rv0859/Rv3546/Rv1142c/Rv1141c/Rv3509*
mtu00380	5	*Rv0859/Rv3546/Rv1142c/Rv1141c/Rv1908c*
mtu00010	4	*Rv3303c/Rv0650/Rv0363c/Rv1617*
mtu00900	4	*Rv0859/Rv3546/Rv3382c/Rv1110*
mtu00071	4	*Rv0859/Rv3546/Rv1142c/Rv1141c*
mtu00310	4	*Rv0859/Rv3546/Rv1142c/Rv1141c*
mtu00362	4	*Rv0859/Rv3546/Rv1142c/Rv1141c*

vant to the pathogenesis of *M. tuberculosis* and might be pathogenic genes. Sixteen genes have been identified including six new genes (*Rv0482, Rv2981c, Rv2163c, Rv3910, Rv0016c, Rv3330*) (Figure 2). Mtu01040 is an unsaturated fatty acid pathway. Isoniazid is a forefront anti-TB drug aimed to block the function of desaturase [16]. Isoniazid is able to block the translation of saturated fatty acids C24 and C26 into unsaturated fatty acids, which are possible precursors of mycolic acid that is critical in the formation of the cell wall. Inhibition of the synthesis of mycolic acid may potentially disturb the acid resistance of *M. tuberculosis*, which suggests

Table 3. Pathogenic and unknown genes on pathway analyses.

Pathway	Pathogenic genes	Unknown genes
mtu00061	***Rv3285/Rv2247/Rv0904c/Rv1350/Rv0242c/ Rv2524c/Rv0824c/Rv1094/Rv2243/Rv0649***	***Rv2501c/Rv3502c/Rv3559c/Rv0769/Rv2187***
mtu00550	***Rv2157c/Rv2152c/Rv2155c/Rv1315/Rv0482/ Rv2153c/Rv2156c/Rv2158c/Rv2911/Rv3627c***	***Rv0482/Rv2981c/Rv2163c/Rv3910/Rv0016c/Rv3330***
mtu01040	***Rv1350/Rv0242c/Rv0824c/Rv1094/Rv1618/ Rv2605c***	***Rv3502c/Rv3559c/Rv0769/Rv0860***

Table 4. Classification of virulence genes.

Lipid	PE/PPE	Outer membrane proteins	Kinase
Gene	**Gene**	**Gene**	**Gene**
Rv3875/Rv3874/Rv0288/ Rv3804/Rv0477/ Rv1661/Rv2934/Rv3823c/ Rv0899/Rv3875/Rv3874/ Rv0288/Rv3804c/Rv1661/ Rv2934/Rv0373c/Rv2485c/ Rv1695/Rv1412/Rv0650/ Rv2368c/Rv0757/Rv1169c/ Rv0282/Rv0283/Rv0284/ Rv0289/Rv0290/Rv0291/ Rv0292	***Rv3347c/Rv1768/Rv1983/Rv3367/Rv0 Rv0287/Rv3872/Rv3873/Rv1087/Rv3022A /Rv2770c/Rv1787/Rv1789/Rv1818c/ Rv1195/Rv0285/Rv1386/Rv1651c/Rv334 3c/Rv3344c/Rv3345c/Rv3350c/Rv3507 /Rv3508/Rv3511/Rv3514/Rv3597c/Rv3286c /Rv2710/Rv3414c/Rv2583c/Rv1168c /Rv2430c/Rv1917c/Rv2108/Rv3892c /Rv3893c/Rv0286/Rv1788/Rv1790/ Rv1791/Rv0442c/Rv3022c/Rv3136/Rv1361c /Rv0872c/Rv2162c/Rv2490c/Rv2591 /Rv1452c/Rv1450c/Rv1396c/Rv0152c/ Rv1441c/Rv2634c/Rv2853/Rv3135/ Rv3478/Rv1468c/Rv1067c/Rv0747/Rv2098c /Rv1803c/Rv0279c/Rv0160c/Rv2107 /Rv0335c/Rv0916c/Rv2519/Rv2769c/ Rv3477/Rv3622c/Rv0453/Rv3018c/Rv1387 /Rv3425/Rv3426/Rv3429/Rv1801/ Rv1809/Rv3621c/Rv3532/Rv2768c /Rv1705c/Rv3125c/Rv1548c/Rv0755c Rv0305c/Rv0355c/Rv0878c/Rv1135c/Rv1753c Rv1918c/Rv1800/Rv3159c/Rv3533c /Rv3144c/Rv3558/Rv2352c***	***Rv3877/Rv0841/Rv1028A /Rv2219A/Rv2401A/Rv3395A/ Rv1440/Rv2206/Rv3604c/Rv0928 /Rv0402c/Rv0450c/Rv1557 /Rv2339/Rv1183/Rv1522c /Rv0403c/Rv0506/Rv2198c /Rv0451c/Rv0677c/Rv2345/ Rv0899/Rv2091c/Rv3635/Rv1481 /Rv2969c/Rv3669/Rv0528 /Rv0497/Rv0051***	***Rv1908c/Rv1484/Rv2957/Rv2958c/Rv0447c /Rv0642c/Rv1449c/Rv3141/Rv3283/Rv1285/Rv3303c /Rv3868/Rv3883c/Rv2202c/Rv3825c/Rv1527c/Rv3298c /Rv3543c/Rv3544c/Rv0859/Rv0405/Rv1886c /Rv3487c/Rv2590/Rv2780/Rv0467/Rv2244/Rv2245 /Rv2246/Rv0645c/Rv0643c/Rv0503c/Rv0410c/Rv0757 /Rv2112c/Rv2097c/Rv2115c/Rv1169c/Rv2032/Rv3127 /Rv3627c/Rv1018c/Rv2460c/Rv3919c/Rv0904c /Rv2247/Rv2523c/Rv0649/Rv1618 /Rv2605c/Rv1722 /Rv3391/Rv3392c/Rv0824c/Rv1094/Rv3229c/Rv3538 /Rv0469/Rv3473/Rv1123c/Rv0554/Rv3617/Rv1938 /Rv1124/Rv2214c/Rv3670/Rv0134/Rv3171/Rv3846 /Rv0432/Rv1932/Rv3177/Rv3245c/Rv0220/Rv1923 /Rv3775/Rv0646c/Rv1399c/Rv1400c/Rv1900c/Rv2385 /Rv1497/Rv2284/Rv2970c/Rv1426c/Rv2463/Rv3084 /Rv3176c/Rv2045c/Rv1076/Rv3203/Rv0217c/Rv2351c /Rv2350c/Rv2349c/Rv0806c/Rv0112/Rv0113/Rv3782 /Rv2223c/Rv2224c/Rv2672/Rv3452/Rv0634c /Rv2581c/Rv0482/Rv2152c/Rv2155c/Rv2158c/Rv2157c /Rv2153c/Rv1338/Rv2156c/Rv2220/Rv2941/Rv1430 /Rv3332/Rv1315/Rv3809c /Rv1302/Rv3285***
Pathway	**Pathway**	**Pathway**	**Pathway**
mtu00010/ mtu00052/ mtu00500/ mtu00520/ mtu00521/ mtu01100/ mtu01110/ mtu01120/ mtu01130/ mtu01200/ mtu00740/ mtu00760/ mtu00633/ mtu00680/ mtu00561/ mtu05152	**mtu00230**	**mtu02024/mtu03060/ mtu03070/mtu02010/ mtu02020/mtu05152**	**mtu00360/mtu00380/mtu01100/mtu00030/ mtu01100/mtu01110/mtu01120/mtu01130/ mtu01200/mtu01230/mtu00270/mtu00920/ mtu04122/mtu00230/mtu00261/mtu00450/ mtu01130/mtu00010/mtu00020/mtu00260/ mtu00280/mtu00620/mtu00630/mtu00640/ mtu00230/mtu00071/mtu00072/mtu00310/ mtu00362/mtu00380/mtu00620/mtu00630/ mtu00650/mtu00900/mtu01212/mtu00561/ mtu00250/mtu00430/mtu05152/mtu03050/ mtu00550/mtu00520/mtu01130/mtu00620/ mtu00280/mtu00770/mtu00061/mtu01040/ mtu02020/mtu00562/mtu00564/mtu00565/ mtu02024/mtu00052/mtu00051/mtu00300/ mtu01502/mtu00471/mtu00220/mtu00910**

the functional selectivity of isoniazid towards the cell wall of *M. tuberculosis*. The functions of 6 genes in the pathway are not completely understood, while the role of other 4 potential pathogenic genes (*Rv3502c, Rv3559c, Rv0769, Rv0860*) remain unclear.

In this study, we found that most of the pathways associated with pathogenic genes were identified as basic metabolic pathways of microbes, which indicated that the pathogenesis of TB might be related to the general metabolism. The immune system plays a vital role in

Table 5. Operons contains three or more pathogenic genes.

Gene ID	Pathogenic genes number	All genes number	Pathogenic genes	Unknown genes	Percentage of pathogenic genes to all genes
6951	10	11	*Rv0282/Rv0283/Rv0284/Rv0286/ Rv0289/Rv0290/Rv0291/Rv0292/ Rv0288/Rv0287*	*Rv0285*	91%
6927	8	12	*Rv0167/Rv0168/Rv0169/Rv0170/ Rv0171/Rv0172/Rv0173/Rv0174*	*Rv0175/Rv0176/Rv0177/ Rv0178*	67%
7336	7	12	*Rv1964/Rv1965/Rv1967/Rv1968/ Rv1969/Rv1971/Rv1966*	*Rv1970/Rv1972/Rv1973/ Rv1974/Rv1975*	58%
7375	7	11	*Rv2152c/Rv2155c/Rv2153c/Rv2156c/ Rv2157c/Rv2158c/Rv2160c*	*Rv2151c/Rv2154c/Rv2159c/ Rv2160A*	64%
7694	7	10	*Rv3494c/Rv3496c/Rv3497c/Rv3498c/ Rv3500c/Rv3501c/Rv3499c*	*Rv3492c/Rv3493c/Rv3495c*	70%
7017	6	10	*Rv0587/Rv0588/Rv0591/Rv0592/ Rv0594/Rv0589*	*Rv0586/Rv0590/Rv0590A/ Rv0593*	60%
6915	6	6	*Rv0096/Rv0097/Rv0098/Rv0099/ Rv0100/Rv0101*		100%
7777	5	6	*Rv3870/Rv3869/Rv3868/Rv3866/ Rv3867*	*Rv3865*	83%
7577	4	5	*Rv3019c/Rv3020c/Rv3022A/Rv3022c*	*Rv3021c*	80%
7302	4	4	*Rv1795/Rv1796/Rv1797/Rv1798*		100%
7394	4	4	*Rv2244/Rv2245/Rv2246/Rv2247*		100%
7683	4	4	*Rv3444c/Rv3445c/Rv3446c/Rv3447c*		100%
7722	4	4	*Rv3619c/Rv3620c/Rv3621c/Rv3622c*		100%
6987	3	8	*Rv0450c/Rv0451c/Rv0447c*	*Rv0444c/Rv0445c/Rv0446c/ Rv0448c/Rv0449c*	38%
7590	3	7	*Rv3089/Rv3083/Rv3084*	*Rv3085/ Rv3086/ Rv3087/ Rv3088/*	43%
7279	3	6	*Rv1696/Rv1694/Rv1695*	*Rv1691/Rv1692/Rv1693*	50%
7527	3	3	*Rv2816c/Rv2817c/Rv2818c*		100%
7528	3	3	*Rv2819c/Rv2820c/Rv2821c*		100%
7529	3	3	*Rv2823c/Rv2822c/Rv2824c*		100%
7657	3	3	*Rv3343c/Rv3344c/Rv3345c*		100%
7696	3	3	*Rv3504/Rv3505/Rv3506*		100%
7761	3	3	*Rv3799c/Rv3800c/Rv3801c*		100%
7782	3	3	*Rv3885c/Rv3886c/Rv3887c*		100%
7784	3	3	*Rv3894c/Rv3895c/Rv3896c*		100%

metabolism as suggested by the majority of current research. Moreover, 3 new pathways (mtu00061, mtu00550, mtu01040) were considered to be involved in TB pathogenesis, and most genes in these pathways were pathogenic genes or inferred potential pathogenic genes.

Screening of H37Rv pathogenic genes of *M. tuberculosis* based on gene function analysis

We found 321 H37Rv genes with clear function on the Pfam website, and these genes were classified into 6 categories: PE/PPE family related genes (98 genes); lipid related genes (9 genes); metabolism related genes (21 genes); membrane protein related genes (31 genes); endocrine system related genes (41 genes); and enzyme related genes (121 genes) (Table 4). These functions are closely related to synthesis of the cell wall, whose structural components play a critical role in the pathogenesis of TB.

PE/PPE family proteins are mainly expressed in *M. tuberculosis*, which attracted the attention of researchers [17]. Almost all genes in the PE/PPE family are related

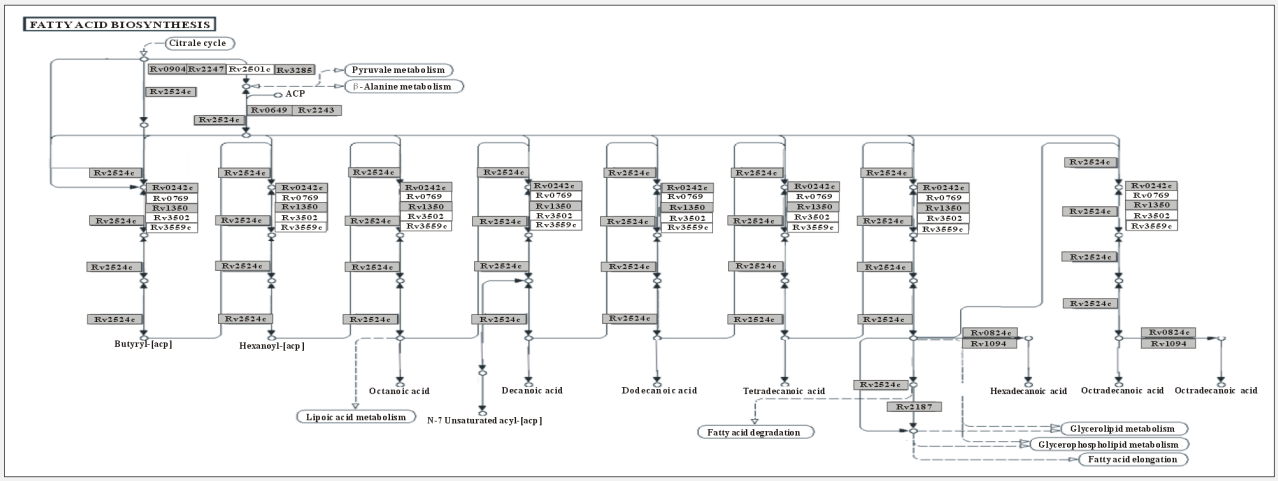

Figure 1. The mtu00061 pathway and its function in fatty acid biosynthesis.

The grey boxes indicate H37Rv pathogenic genes of *M. tuberculosis*, and the white boxes indicate other genes.

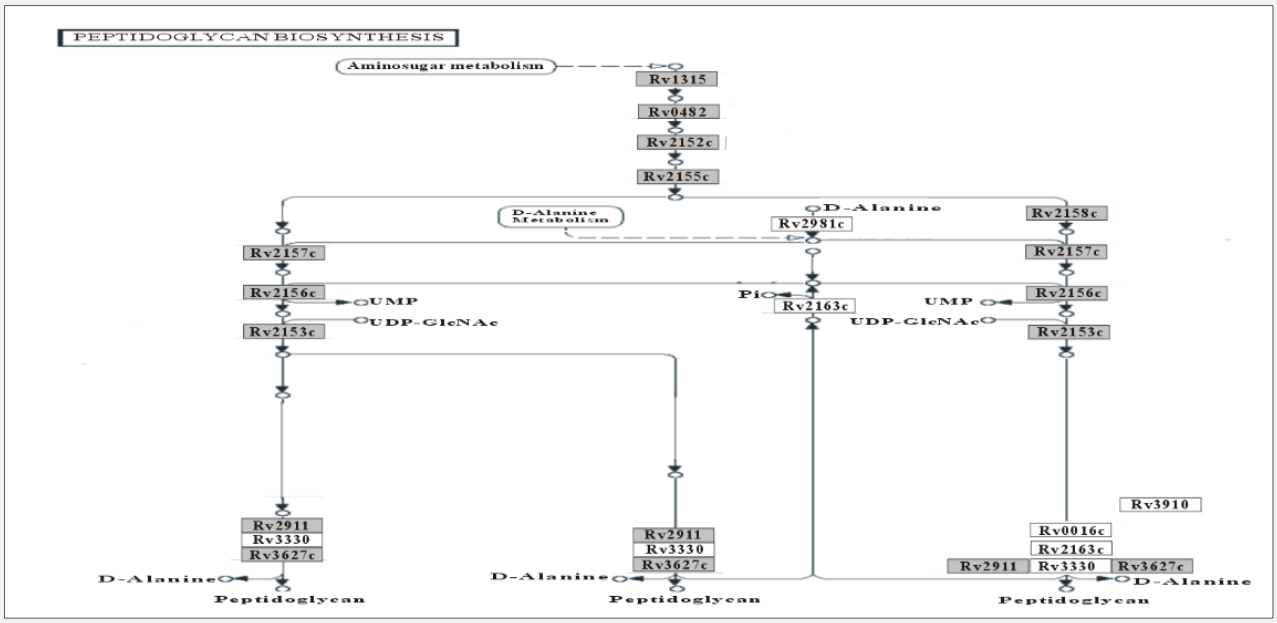

Figure 2. The mtu00550 pathway and its function in peptidoglycan biosynthesis.

The grey boxes indicate H37Rv pathogenic genes of *M. tuberculosis*, and the white boxes indicate other genes.

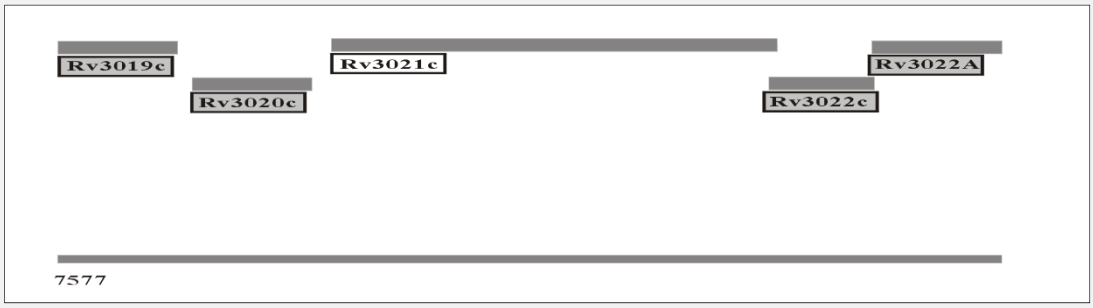

Figure 3. The *M. tuberculosis* H37Rv operon 7577.

Gray boxes indicate the pathogenic gene, white boxes indicate other genes.

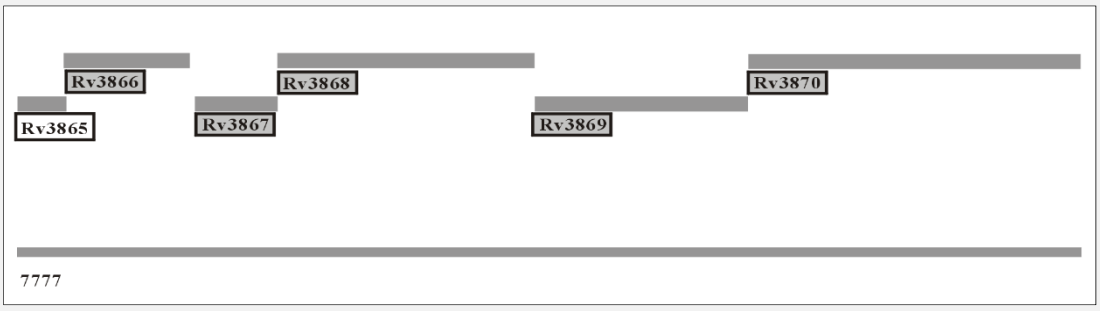

Figure 4. The *M. tuberculosis* H37Rv operon 7777.

Gray boxes indicate the pathogenic gene, white boxes indicate other genes.

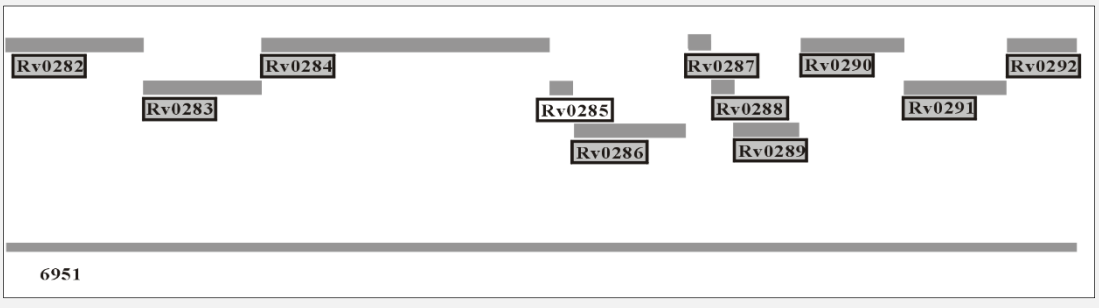

Figure 5. The *M. tuberculosis* H37Rv operon 6951.

Gray boxes indicate the pathogenic gene, white boxes indicate other genes.

to the pathogenesis of TB, and 98 pathogenic genes (58% of the genes of PE/PPE family) were proposed by us through literature mining. PE/PPE family proteins are located on the cell wall of *M. tuberculosi*s, which suggests that the proteins likely play an important role in the interaction between the pathogenic bacteria and the host. Then, pathway analysis was conducted for 98 genes. We found that only *Rv2583c* was involved in the pathway mtu00230, while the rest of the genes were not in any pathways.

As to *M. tuberculosis*, its pathogenesis mainly originates from the thallus, which consists of lipids, polysaccharides, and proteins which are abundantly expressed in the cell wall. The toxicity of *M. tuberculosis* is parallel to the capacity of lipids in the cell wall, which means the toxicity depends on the volume of lipids [18]. Lipids have large capacity (20% - 40%) in *M. tuberculosis* especially in the cell wall (60%). It confers effective hydrophobicity and high resistance to physical and chemical factors due to the large volume [19]. There are 30 pathogenic genes and 16 pathways related to lipids, in which mtu05152 and mtu00561 drew our attention. Mtu05152 pathway includes 13 genes, which also exist in lipid functional pathways that were classified by us (Table 4). Six genes of the pathway are pathogenic genes, and we hypothesized that the remaining seven (*Rv3818, Rv1411c, Rv1270c, Rv0932c, Rv0934, Rv3417c, Rv3310*) are pathogenic genes as well. Mtu00561 is involved in glycerin metabolism, which is a critical energy resource for the growth of *M. tuberculosis*, and the ketoplasia caused by fat metabolic disorder could elevate the vitalily of *M. tuberculosis*. In addition to the 3 known pathogenic genes, the remaining 12 of the 15 genes are also pathogenic genes in all probability.

Thirty-one pathogenic genes and 6 pathways (mtu02024, mtu03060, mtu03070, mtu02010, mtu02020, and mtu05152) were considered to be related to membrane proteins. Among these pathways, mtu02024, mtu02010, mtu02020, and mtu03060 are common protein transport and export pathways of microbes. Mtu03070 is a pathway associated to bacterial secretion system. According to previous research, this pathway contains a pathogenic gene of *M. tuberculosis*. Although 14 genes have been identified in the pathway, we are not able to define other genes as pathogenic genes based on the evidence of a single gene.

One hundred and twenty-one pathogenic genes and 60 pathways relate to enzymes. *M. tuberculosis* is able to express acid phosphatase [20], which prevents the merging of the phagocytotic vesicle and lysosome that express damaging free radicals like reactive oxygen and nitrogen. Thus, the phagocytosis and immune reaction of the host are evaded by *M. tuberculosis*, which results in long term *in vivo* incubation. Mtu00565 and mtu00562 are related to metabolic enzymes. Mtu00565 (5 genes) is a lipid metabolism pathway in which 3 genes are confirmed to be pathogenic genes, and the other 2 (*Rv2251, Rv3107c*) are probably pathogenic as well. Mtu00562 (8 genes) is a pathway of inositol phosphate metabolism in which 3 genes are pathogenic genes, and the other 5 (*Rv2701c, Rv1604, Rv0046c, Rv0753c, Rv1438*) have a high likelihood of being pathogenic.

The pathogenesis of *M. tuberculosis* is highly dependent on the unique structure of the cell wall, which suggests that the pathogenesis is closely related to PE/PPE family proteins, lipid metabolism, membrane proteins, and metabolic enzymes. We found that the majority of the genes from the single pathway are pathogenic genes, and that the rest of the genes can be pathogenic as well. Based on this theory, new potential pathogenic genes are proposed.

Screening for H37Rv pathogenic genes of *M. tuberculosis* through operon analysis

We searched the operons of the reported *M. tuberculosis* pathogenic genes based on literature mining on the website http://csbl.bmb.uga.edu/DOOR/index.php (Table 5). Four hundred and twenty-nine genes were found on operons by data analysis. Nine operons were related to at least five genes, and three of these were found to be related to up to 24 genes. At the same time, we found that all the genes were pathogenic genes in 13 operons (accounting for 54% of the total operon). Three operons (operon 6951, operon 7777, operon 7577) have only one gene for which evidence of being pathogenic has yet to be obtained (*Rv0285, Rv3865, Rv3021c*) (Figure 3, Figure 4, Figure 5), respectively. On the other operons (operon 6927, operon 7375, operon 7694, and operon 7017), as long as the proportion of pathogenic genes is greater than 60%, the remaining genes are likely to be pathogenic genes.

According to our analysis, 3 genes (*Rv0285, Rv3865, and Rv3021c*) have the highest possibility to be pathogenic genes, and 15 genes (*Rv0175, Rv0176, Rv0177, Rv0178, Rv2151c, Rv2154c, Rv2159c, Rv2160A, Rv3492c, Rv3493c, Rv3495c, Rv0586, Rv0590, Rv0590A, Rv0593*) are potential pathogenic genes.

CONCLUSION

We identified 584 pathogenic genes of *M. tuberculosis* H37Rv via bibliometrics. Also, the potential interaction of their pathways, functions, and operons are analyzed, and the specter of new pathogenic genes proposed. Our findings provide theoretical evidence for epidemiological research on prevention of TB, and afford a new index of diagnosis and prognosis of TB.

Acknowledgement:

This work was supported by National Natural Science Foundation of China (81401712), the Natural Science Foundation of Jilin Province (201603092YY), Jilin Province Chinese medicine science and technology projects (2017086), Jilin City Science and Technology Innovation and Development Project (20166017),

Beihua University Research and Development Innovation team of Animal and Plant Resources in Changbai Mountain.

Co-corresponding:
Nan Zhang and Yuanyi Wang are co-corresponding authors.

Declaration of Interest:
The authors declare that they have no competing interests.

References:

1. Dutta NK, Mehra S, Didier PJ, et al. Genetic requirements for the survival of tubercle bacilli in primates. J Infect Dis 2010;201(11): 1743-52 (PMID: 20394526).

2. Saramba MI, Zhao D. A Perspective of the Diagnosis and Management of Congenital Tuberculosis. J Pathog 2016;2016: 8623825 (PMID: 27999684).

3. Arcos J, Sasindran SJ, Moliva JI, et al. Mycobacterium tuberculosis cell wall released fragments by the action of the human lung mucosa modulate macrophages to control infection in an IL-10-dependent manner. Mucosal Immunol 2017;10(5):1248-58 (PMID: 28000679).

4. Kizilbash QF, Seaworth BJ. Multi-drug resistant tuberculous spondylitis: A review of the literature. Ann Thorac Med 2016; 11(4):233-36 (PMID: 27803747).

5. Ghimire S, Van't Boveneind-Vrubleuskaya N, Akkerman OW, et at. Pharmacokinetic/pharmacodynamic-based optimization of levofloxacin administration in the treatment of MDR-TB. J Antimicrob Chemother 2016;71(10):2691-703 (PMID: 27231277).

6. Walker TM, Merker M, Kohl TA, Crook DW, Niemann S, Peto TE. Whole genome sequencing for M/XDR tuberculosis surveillance and for resistance testing. Clin Microbiol Infect 2017;23(3): 161-6 (PMID: 27789378).

7. Maitre T, Aubry A, Jarlier V, Robert J, Veziris N; CNR-My-RMA. Multidrug and extensively drug-resistant tuberculosis. Med Mal Infect 2017;47(1):3-10 (PMID: 27637852).

8. Chen YL, Mo XQ, Huang GR, et al. Gene polymorphisms of pathogenic Helicobacter pylori in patients with different types of gastrointestinal diseases. World J Gastroenterol 2016;22(44): 9718-26 (PMID: 27956795).

9. Sun LY, Zhang YB, Jiang L, et al. Identification of the gene defect responsible for severe hypercholesterolaemia using whole-exome sequencing. Sci Rep 2015;5:11380 (PMID: 26077743).

10. Xu G, Liu B, Wang F, et al. High-throughput screen of essential gene modules in Mycobacterium tuberculosis: a bibliometric approach. BMC Infect Dis 2013;13:227 (PMID: 23687949).

11. Odongo CO, Bisaso RK, Byamugisha J, Obua C. Intermittent use of sulphadoxine-pyrimethamine for malaria prevention: a cross-sectional study of knowledge and practices among Ugandan women attending an urban antenatal clinic. Malar J 2014;13:399 (PMID: 25306431).

12. Xu G, Ni Z, Shi Y, et al. Screening essential genes of *Mycobacterium tuberculosis* with the pathway enrichment method. Mol Biol Rep 2014, 41(11):7639-44 (PMID: 25098602).

13. Mao X, Ma Q, Zhou C, Chen X, et al. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res 2013;42 (Database issue):D654-9 (PMID: 242 14966).

14. Snider J, Hanif A, Lee ME, et al. Mapping the functional yeast ABC transporter interactome. Nat Chem Biol 2013;9(9):565-72 (PMID: 23831759).

15. Singh V, Mani I, Chaudhary DK, Somvanshi P. The beta-ketoacyl-ACP synthase from Mycobacterium tuberculosis as potential drug targets. Curr Med Chem 2011;18(9):1318-24 (PMID: 21370 994).

16. Javad Nasiri M, Chirani AS, Amin M, Halabian R, Imani Fooladi AA. Isoniazid-resistant tuberculosis in Iran: A systematic review. Tuberculosis (Edinb) 2016;98:104-9 (PMID: 27156625).

17. Fishbein S, van Wyk N, Warren RM, Sampson SL. Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. Mol Microbiol 2015;96(5):901-16 (PMID: 25727695).

18. Aguilar-Ayala DA, Palomino JC, Vandamme P, Martin A, Gonzalez-Y-Merchand JA. "Genetic regulation of *Mycobacterium tuberculosis* in a lipid-rich environment". Infect Genet Evol. 2016 Nov;55:392-402 (PMID: 27771519).

19. Chalut C. MmpL transporter-mediated export of cell-wall associated lipids and siderophores in mycobacteria. Tuberculosis (Edinb) 2016;100:32-45 (PMID: 27553408).

20. Gonzalo-Asensio J, Mostowy S, Harders-Westerveen J, et al. PhoP: a missing piece in the intricate puzzle of *Mycobacterium tuberculosis* virulence. PloS One 2008;3(10):e3496 (PMID: 1894 6503).