PLoS one

MycoRRdb: A Database of Computationally Identified Regulatory Regions within Intergenic Sequences in Mycobacterial Genomes

Mohit Midha, Nirmal K. Prasad, Vaibhav Vindal*

Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India

Abstract

The identification of regulatory regions for a gene is an important step towards deciphering the gene regulation. Regulatory regions tend to be conserved under evolution that facilitates the application of comparative genomics to identify such regions. The present study is an attempt to make use of this attribute to identify regulatory regions in the *Mycobacterium species* followed by the development of a database, MycoRRdb. It consist the regulatory regions identified within the intergenic distances of 25 mycobacterial species. MycoRRdb allows to retrieve the identified intergenic regulatory elements in the mycobacterial genomes. In addition to the predicted motifs, it also allows user to retrieve the Reciprocal Best BLAST Hits across the mycobacterial genomes. It is a useful resource to understand the transcriptional regulatory mechanism of mycobacterial species. This database is first of its kind which specifically addresses *cis*-regulatory regions and also comprehensive to the mycobacterial species. Database URL: http://mycorrdb.uohbif.in.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: vvls@uohyd.ernet.in

Introduction

Over the past few years the genomic sequence repertoire of mycobacterial sequences has increased tremendously. The availability of complete genome sequences makes it possible to efficiently employ computational approaches to understand the genome function and its complexity [1]. One of the important aspects to compare genome sequences is to find orthologous proteins among the existing species [2,3]. The identification of orthologs is important not only to assist the functional annotation of a gene but also to identify its regulatory region. These regions are known to evolve at a slower rate than non-functional elements, and therefore finding the conserved DNA motifs within non coding region is an efficient method to predict these regions [4,5]. Different approaches have been used to find the regulatory regions [6–8]. Generally, identification of these DNA elements relies on an extensive set of known target genes [4,9]. Therefore, identification of regulatory region for a novel transcriptional regulator remains a challenging task.

Extensive research on mycobacteria has produced a number of online resources, providing information on pathogenicity, cellular physiology, operon arrangement, microarray, etc. [10–15]. These resources also include a database, MtbRegList, which contains the reported regulatory regions in *Mycobacterium tuberculosis* [16]. Nevertheless, there is still need to document the putative regulatory regions for all the mycobacterial genomes. Our present study addresses this issue, as it identifies the putative *cis*- regulatory sequences within the intergenic regions of mycobacterial species and also the similar DNA motif in a genome. In addition to the predicted regulatory regions, the database includes list of

Reciprocal Best BLAST Hits (RBBHs) for all 25 mycobacterial species. The database also has a search feature to identify the sequences similar to a query DNA motif. This database can assist in the characterization of gene regulation in all the mycobacterial species.

Methods

Retrieval and filtering the genome sequences

The complete genome sequences of 25 *Mycobacterium species* were downloaded from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Some of the proteins were found to be present in more than one copy, identical in sequence, in certain species. In present study, such Multiple Identical Proteins (MIPs) were identified and replaced with only one representative protein sequence for further analysis.

Identification of orthologs

Reciprocal Best BLAST Hit (RBBH) method was used to predict orthologous proteins in mycobacterial proteomes. Pairs of proteins, from two mycobacterial species, covering the at least 50% sequence length of both the proteins in alignment and E-values lower than of 10^{-20} for both directions using BLASTP program with all other parameters at default values were selected as RBBHs [17–19].

Retrieval of operons and the intergenic sequences

Information for all mycobacterial operons in genomes of all 25 species were retrieved from the DOOR database (version2) [20,21] Intergenic sequence upstream of the first gene of each
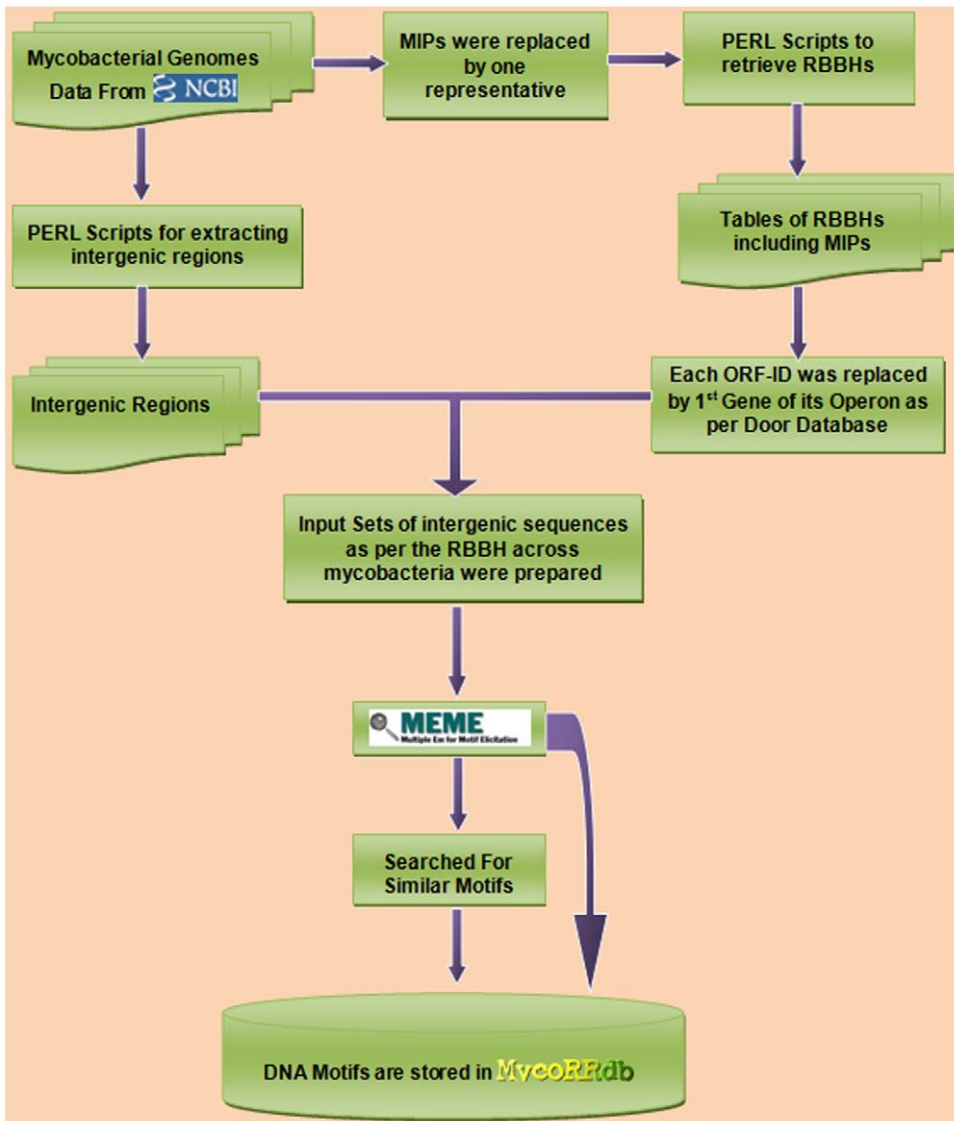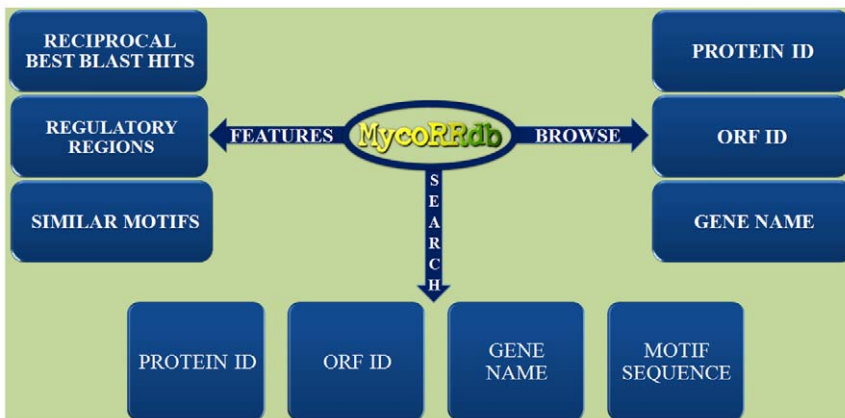
Figure 1. Flowchart of the methodology.

Figure 2. Schematic architecture of MycoRRdb.

Home   Reciprocal Best Hits   Search Motifs   Motifs By Sequence   Help   Submit Motif to MycoRRdb   Contact Us

Retrieve Reciprocal Best Blast Hits of Rv0025

About
M. abscessus ATCC 19977
M. africanum GM041182
M. avium 104
M. avium paratuberculosis
M. bovis AF2122/97
M. bovis BCG Pasteur 1173P2
M. bovis BCG Tokyo 172
M. canettii
M. gilvum
M. JDM601
M. leprae TN
M. leprae Br4923
M. marinum M
M. smegmatis
M. Spyr1
M. tuberculosis H37Rv
M. tuberculosis CDC1551
M. tuberculosis F11
M. tuberculosis H37Ra
M. tuberculosis KZN 1435
M. ulcerans
M. vanbaalenii
M. sp. JLS
M. sp. KMS
M. sp. MCS

Mycobacterium tuberculosis H37Rv

Protein Id	Synonym	Gene	Strand	Start	End	Length
15607153	Rv0011c	-	-	13714	13995	93
15607154	Rv0012	-	+	14089	14877	262
57116682	Rv0013	trpG	+	14914	15612	232
15607156	Rv0014c	pknB	-	15590	17470	626
15607157	Rv0015c	pknA	-	17467	18762	431
15607158	Rv0016c	pbpA	-	18759	20234	491
15607159	Rv0017c	rodA	-	20231	21640	469
15607160	Rv0018c	ppp	-	21637	23181	514
15607161	Rv0019c	-	-	23270	23737	155
15607162	Rv0020c	TB39.8	-	23861	25444	527
15607163	Rv0021c	-	-	25913	26881	322
15607164	Rv0022c	whiB5	-	27023	27442	139
15607165	Rv0023	-	+	27595	28365	256
15607166	Rv0024	-	+	28362	29207	281
15607167	Rv0025	-	+	29245	29607	120
15607168	Rv0026	-	+	29722	31068	448
15607169	Rv0027	-	+	31189	31506	105
15607170	Rv0028	-	+	31514	31819	101
15607171	Rv0029	-	+	32057	33154	365
15607172	Rv0030	-	+	33224	33553	109

Organism	RBBH
Mycobacterium tuberculosis CDC1551	MT0028
Mycobacterium tuberculosis H37Ra	MRA_0028
Mycobacterium tuberculosis F11	TBFG_10025
Mycobacterium tuberculosis KZN	TBMG_00025
Mycobacterium bovis AF2122 97	Mb0026
Mycobacterium bovis BCG Pasteur	BCG_0056
Mycobacterium bovis BCG Tokyo	JTY_0026
Mycobacterium leprae TN	----------
Mycobacterium leprae Br4923	----------
Mycobacterium avium paratuberculosis K 10	MAP0037
Mycobacterium avium	MAV_0043
Mycobacterium smegmatis	----------
Mycobacterium ulcerans	MUL_0043
Mycobacterium vanbaalenii	----------
Mycobacterium gilvum	----------
Mycobacterium abscessus ATCC 19977	----------
Mycobacterium sp. MCS	----------
Mycobacterium sp. KMS	----------
Mycobacterium sp. JLS	----------
Mycobacterium marinum M	MMAR_0044
Mycobacterium Spyr1	----------
Mycobacterium africanum GM041182	MAF_00250
Mycobacterium canettii	MCAN_00241
Mycobacterium JDM601	----------

Putative Regulatory Motif Information of Rv0025

Orthologs

ORF-ID	Strand	Sequence	Start	End
Rv0025	+	CGCCCAATCAGACGGGCAGGATGAGAAGGTT	-36	-6
TBMG_00025	+	CGCCCAATCAGACGGGCAGGATGAGAAGGTT	-36	-6
MAF_00250	+	CGCCCAATCAGACGGGCAGGATGAGAAGGTT	-36	-6
Mb0026	+	CGCCCAATCAGACGGGCAGGATGAGAAGGTT	-36	-6
BCG_0056	+	CGCCCAATCAGACGGGCAGGATGAGAAGGTT	-36	-6
JTY_0026	+	CGCCCAATCAGACGGGCAGGATGAGAAGGTT	-36	-6
MCAN_00241	+	CGCCCAATCAGACGGGCAGGATGAGAAGGTT	-36	-6
MT0028	+	CGCCCAATCAGACGGGCAGGATGAGAAGGTT	-36	-6
TBFG_10025	+	CGCCCAATCAGACGGGCAGGATGAGAAGGTT	-36	-6
MRA_0028	+	CGCCCAATCAGACGGGCAGGATGAGAAGGTT	-36	-6

Similar Motifs

ORF-ID	Strand	Sequence	Start	End
Rv2005c	-	CCCAATCGGACGGCCA	-67	-82
Rv3075c	-	AATCGGATTGGCAGGATG	-62	-79
Rv3076	+	AATCGGATTGGCAGGATG	-37	-20
Rv2006	+	CCCAATCGGACGGCCA	-52	-37
Rv3125c	-	GCCCAATCAGCCGCCCAG	-11	-28

Click to Find References For Reported Motif.

Previous ORF   List of ORFs in *Mycobacterium tuberculosis H37Rv*   Next ORF

Figure 3. A browsable interface to retrieve RBBHS and DNA motifs.
doi:10.1371/journal.pone.0036094.g003

operon was retrieved using perl script. Sets of intergenic sequences were compiled for each orthologous gene. These sets of sequences were further subjected to the identification of a regulatory region.

Identification of regulatory regions

MEME suite was used to identify the conserved regulatory DNA elements from the set of sequences described earlier [4]. The DNA motif length, from minimum of 20 bases to maximum length of 30, was optimized using known DNA targets from *M. tuberculosis* [16]. DNA search was carried out to look for palindromes within the given strand as well as its complementary strand. Additionally, the top predicted DNA motifs observed associated with three or more orthologous sequences were selected as potential regulatory DNA element. All other parameters were kept on its default values. These DNA motifs were further searched in their respective genomes to identify the significantly similar motifs with minimum aligned length (L) of 16 bases (allowing N mismatch where $N <= 0.2L$; $L-N > 14$).

Database development

Subsequent to the identification of regulatory regions from all mycobacterial genomes, a web resource, MycoRRdb was developed. This database has been developed using MySQL. It is constructed to allow user to browse the outcome of study in an easy accessible mode. Web interface of the database is designed using PHP, HTML and Javascripts. Flow chart of the methodology followed in the study is depicted in Figure 1.

Results and Discussion

RBBHs across the mycobacterial species

The ortholog prediction is not only important to identify the regulatory region but also helps in functional annotation of a

Figure 4. A searchable mode to retrieve RBBHS and DNA motifs. A. Interface to retrieve the RBBHs; **B.** Interface to retrieve the regulatory DNA motifs; **C.** Interface to retrieve the similar DNA motifs to the desired DNA sequence.
doi:10.1371/journal.pone.0036094.g004

sequenced genome. Our study also began with the identification of the RBBHs which serves as potential ortholog. All the RBBHs from the mycobacterial species were identified using the methodology discussed. The identified lists of RBBHs for any mycobacterial gene across all 25 mycobacterial genome were used as a data source for the MycoRRdb.

Mycobacterial regulatory regions

Subsequently, DNA regulatory regions were identified across the all 25 *Mycobacterium species*. The total predicted regulatory motifs were 37101 in number for all 25 mycobacterial genomes. Further, the motifs predicted across the Mycobacterial species were compared with the known DNA motifs reported in the literature [3,5,16,22–41]. It was observed that 116 DNA motifs, out of 181 retrieved, were mapped in MycoRRdb and notified through the link given in the database. The comparative list of the predicted and the reported DNA motifs is given in Table S1. The maximum number of motifs was predicted from *Mycobacterium tuberculosis* H37Ra while the minimum number was from *Mycobacterium abscessus* ATCC 19977. These predicted DNA motifs are the putative Transcription Factor Binding Sites (TFBS). The TFBS identified, positioned at more than 400 nucleotide upstream to the translational start site, are highlighted with red colour font.

Further in view of over representation, similar DNA motifs were searched to find the similar motifs within the predicted list of intergenic regulatory region. All the identified motifs are displayed with the strand information and the position from translational start site.

Database access

MycoRRdb can be accessed through the database web interface at http://mycorrdb.uohbif.in. There are two kind of data that has been stored in MycoRRdb:(i)Reciprocal Best BLAST Hits (RBBHs), and (ii) Predicted Regulatory Region for each transcription unit (Figure 2). This information for any mycobacterial gene can be retrieved from MycoRRdb in either browsable or searchable fashion. Homepage of the database provides links for the mycobacterial genome which further leads to complete list of genes/protein id/ORF id of a particular species. From the list one can proceed to find the RBBHs of any gene across other mycobacterial species and associated regulatory DNA motifs along with its occurrence in the orthologous intergenic sequences. It also gives link, to facilitate user, to the retrieve the known motif reported in literature. In addition to this list of similar DNA motifs in a genome is also available (Figure 3).

Besides browsing data from complete genes list, separate links have also been made available on web interface to quickly retrieve RBBHs or regulatory DNA motifs by gene name/protein id/ORF id. A searchable interface to retrieve RBBHs is shown in Figure 4A. The predicted regulatory regions and the similar sequences present in that genome can be also be retrieved by searchable interface using gene name/protein id/ORF id (Figure 4B). Moreover, user can scan the availability of its desired DNA sequence, if it exists in any Mycobacterial species, in the identified DNA motifs set of the Database (Figure 4C).

This database is under constant development to gather the experimentally validated DNA motifs to incorporate in the database. It also provides link for biologist to put forward the experimentally validated mycobacterial regulatory regions, if any.

Conclusions

The availability of whole genome sequences makes Mycobacterium one of the highly sequenced genera. This wealth of sequence data provides unique opportunity to extract the genome information in order to address cellular physiology and to develop better intervention strategies for pathogenic species. This study is a systematic approach to reveal the putative regulatory regions and RBBHs across the mycobacterial species. On the one hand, the identified regulatory regions will help to understand the transcriptional regulation of the mycobacterial genes, and on the other hand, the identified RBBHs will assist to impart the functional knowledge of one gene to another. The availability of all the identified regulatory regions and RBBHs from the mycobacterial species at a websource, MycoRRdb, will help to access the data and will have potential implications to unravel the genomic complexity of the mycobacteria.

Supporting Information

Table S1 DNA motifs in MycoRRdb mapped with regulatory regions reported in literature. (Available at: http://mycorrdb. uohbif.in/links.php).
(XLS)

Author Contributions

Conceived and designed the experiments: VV. Performed the experiments: MM. Analyzed the data: MM NKP VV. Wrote the paper: VV. Database design: MM.

References

1. Young DB (2001) A post-genomic perspective. Nature medicine 7: 11–13.
2. Lee SA, Chan CH, Tsai CH, Lai JM, Wang FS, et al. (2008) Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. BMC bioinformatics 9 Suppl 12: S11.
3. Vindal V, Ashwantha Kumar E, Ranjan A (2008) Identification of operator sites within the upstream region of the putative *mce2R* gene from mycobacteria. FEBS letters 582: 1117–1122.
4. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic acids research 34: W369–373.
5. Vindal V, Suma K, Ranjan A (2007) GntR family of regulators in *Mycobacterium smegmatis*: a sequence and structure based characterization. BMC genomics 8: 289.
6. Thieffry D, Salgado H, Huerta AM, Collado-Vides J (1998) Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12. Bioinformatics 14: 391–400.
7. Bulyk ML, McGuire AM, Masuda N, Church GM (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. Genome research 14: 201–208.
8. Siddharthan R, Siggia ED, van Nimwegen E (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. PLoS computational biology 1: e67.
9. Ranjan S, Seshadri J, Vindal V, Yellaboina S, Ranjan A (2006) iCR: a web tool to identify conserved targets of a regulatory protein across the multiple related prokaryotic species. Nucleic acids research 34: W584–587.
10. Bergh S, Cole ST (1994) MycDB: an integrated mycobacterial database. Molecular microbiology 12: 517–534.
11. Catanho M, Mascarenhas D, Degrave W, Miranda AB (2006) GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. Genetics and molecular research : GMR 5: 115–126.
12. Ranjan S, Gundu RK, Ranjan A (2006) MycoperonDB: a database of computationally identified operons and transcriptional units in Mycobacteria. BMC bioinformatics 7 Suppl 5: S9.
13. Reddy TB, Riley R, Wymore F, Montgomery P, DeCaprio D, et al. (2009) TB database: an integrated platform for tuberculosis research. Nucleic acids research 37: D499–508.
14. Vishnoi A, Srivastava A, Roy R, Bhattacharya A (2008) MGDD: *Mycobacterium tuberculosis* genome divergence database. BMC genomics 9: 373.
15. Zhu X, Chang S, Fang K, Cui S, Liu J, et al. (2009) MyBASE: a database for genome polymorphism and gene function studies of Mycobacterium. BMC microbiology 9: 40.
16. Jacques PE, Gervais AL, Cantin M, Lucier JF, Dallaire G, et al. (2005) MtbRegList, a database dedicated to the analysis of transcriptional regulation in *Mycobacterium tuberculosis*. Bioinformatics 21: 2563–2565.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215: 403–410.
18. Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24: 319–324.
19. Fulton DL, Li YY, Laird MR, Horsman BG, Roche FM, et al. (2006) Improving the specificity of high-throughput ortholog prediction. BMC bioinformatics 7: 270.
20. Mao F, Dam P, Chou J, Olman V, Xu Y (2009) DOOR: a database for prokaryotic operons. Nucleic acids research 37: D459–463.
21. Dam P, Olman V, Harris K, Su Z, Xu Y (2007) Operon prediction using both genome-specific and general genomic information. Nucleic acids research 35: 288–298.
22. Rand L, Hinds J, Springer B, Sander P, Buxton RS, et al. (2003) The majority of inducible DNA repair genes in *Mycobacterium tuberculosis* are induced independently of RecA. Mol Microbiol 50: 1031–1042.
23. Davis EO, Dullaghan EM, Rand L (2002) Definition of the mycobacterial SOS box and use to identify LexA-regulated genes in *Mycobacterium tuberculosis*. J Bacteriol 184: 3287–3295.
24. Brooks PC, Movahedzadeh F, Davis EO (2001) Identification of some DNA damage-inducible genes of *Mycobacterium tuberculosis*: apparent lack of correlation with LexA binding. J Bacteriol 183: 4459–4467.
25. Movahedzadeh F, Colston MJ, Davis EO (1997) Determination of DNA sequences required for regulated *Mycobacterium tuberculosis* RecA expression in response to DNA-damaging agents suggests that two modes of regulation exist. J Bacteriol 179: 3509–3518.
26. Prakash P, Yellaboina S, Ranjan A, Hasnain SE (2005) Computational prediction and experimental verification of novel IdeR binding sites in the upstream sequences of *Mycobacterium tuberculosis* open reading frames. Bioinformatics 21: 2161–2166.
27. Rodriguez GM, Gold B, Gomez M, Dussurget O, Smith I (1999) Identification and characterization of two divergently transcribed iron regulated genes in *Mycobacterium tuberculosis*. Tuber Lung Dis 79: 287–298.
28. Gold B, Rodriguez GM, Marras SA, Pentecost M, Smith I (2001) The *Mycobacterium tuberculosis* IdeR is a dual functional regulator that controls transcription of genes involved in iron acquisition, iron storage and survival in macrophages. Mol Microbiol 42: 851–865.
29. Rodriguez GM, Voskuil MI, Gold B, Schoolnik GK, Smith I (2002) *ideR*, An essential gene in *Mycobacterium tuberculosis*: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. Infect Immun 70: 3371–3381.
30. Haydel SE, Benjamin WH, Jr., Dunlap NE, Clark-Curtiss JE (2002) Expression, autoregulation, and DNA binding properties of the *Mycobacterium tuberculosis* TrcR response regulator. J Bacteriol 184: 2192–2203.
31. Cavet JS, Meng W, Pennella MA, Appelhoff RJ, Giedroc DP, et al. (2002) A nickel-cobalt-sensing ArsR-SmtB family repressor. Contributions of cytosol and effector binding sites to metal selectivity. J Biol Chem 277: 38441–38448.
32. Stewart GR, Wernisch L, Stabler R, Mangan JA, Hinds J, et al. (2002) Dissection of the heat-shock response in *Mycobacterium tuberculosis* using mutants and microarrays. Microbiology 148: 3129–3138.
33. Dullaghan EM, Brooks PC, Davis EO (2002) The role of multiple SOS boxes upstream of the *Mycobacterium tuberculosis* lexA gene–identification of a novel DNA-damage-inducible gene. Microbiology 148: 3609–3615.

34. Recchi C, Sclavi B, Rauzier J, Gicquel B, Reyrat JM (2003) *Mycobacterium tuberculosis* Rv1395 is a class III transcriptional regulator of the AraC family involved in cytochrome P450 regulation. J Biol Chem 278: 33763–33773.

35. Sala C, Forti F, Di Florio E, Canneva F, Milano A, et al. (2003) *Mycobacterium tuberculosis* FurA autoregulates its own expression. J Bacteriol 185: 5357–5362.

36. Engohang-Ndong J, Baillat D, Aumercier M, Bellefontaine F, Besra GS, et al. (2004) EthR, a repressor of the TetR/CamR family implicated in ethionamide resistance in mycobacteria, octamerizes cooperatively on its operator. Mol Microbiol 51: 175–188.

37. He H, Zahrt TC (2005) Identification and characterization of a regulatory sequence recognized by *Mycobacterium tuberculosis* persistence regulator MprA. J Bacteriol 187: 202–212.

38. Rickman L, Scott C, Hunt DM, Hutchinson T, Menendez MC, et al. (2005) A member of the cAMP receptor protein family of transcription regulators in *Mycobacterium tuberculosis* is required for virulence in mice and controls transcription of the *rpfA* gene coding for a resuscitation promoting factor. Mol Microbiol 56: 1274–1286.

39. Kendall SL, Burgess P, Balhana R, Withers M, Ten Bokum A, et al. (2010) Cholesterol utilization in mycobacteria is controlled by two TetR-type transcriptional regulators: *kstR* and *kstR2*. Microbiology 156: 1362–1371.

40. Kendall SL, Withers M, Soffair CN, Moreland NJ, Gurcha S, et al. (2007) A highly conserved transcriptional repressor controls a large regulon involved in lipid degradation in *Mycobacterium smegmatis* and *Mycobacterium tuberculosis*. Mol Microbiol 65: 684–699.

41. Festa RA, Jones MB, Butler-Wu S, Sinsimer D, Gerads R, et al. (2011) A novel copper-responsive regulon in *Mycobacterium tuberculosis*. Mol Microbiol 79: 133–148.