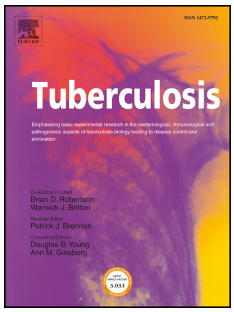
Accepted Manuscript



Single nucleotide polymorphisms may explain the contrasting phenotypes of two variants of a multidrug-resistant *Mycobacterium tuberculosis* strain

María Mercedes Bigi, Beatriz Lopez, Federico Carlos Blanco, María del Carmen Sasiain, Silvia De la Barrera, Marcelo A. Marti, Ezequiel Jorge Sosa, Darío Augusto Fernández Do Porto, Viviana Ritacco, Fabiana Bigi, Marcelo Abel Soria

PII: S1472-9792(16)30366-3

DOI: 10.1016/j.tube.2016.12.007

Reference: YTUBE 1552

To appear in: *Tuberculosis*

Received Date: 4 September 2016

Revised Date: 27 December 2016

Accepted Date: 30 December 2016

Please cite this article as: Bigi MM, Lopez B, Blanco FC, Sasiain MdC, De la Barrera S, Marti MA, Sosa EJ, Fernández Do Porto DA, Ritacco V, Bigi F, Soria MA, Single nucleotide polymorphisms may explain the contrasting phenotypes of two variants of a multidrug-resistant *Mycobacterium tuberculosis* strain, *Tuberculosis* (2017), doi: 10.1016/j.tube.2016.12.007.

Single nucleotide polymorphisms may explain the contrasting phenotypes of two variants of a multidrug-resistant *Mycobacterium tuberculosis* strain

María Mercedes Bigi[1], Beatriz Lopez[2], Federico Carlos Blanco[3], María del Carmen Sasiain[4], Silvia De la Barrera[4], Marcelo A. Marti[5], Ezequiel Jorge Sosa[6], Darío Augusto Fernández Do Porto[6], Viviana Ritacco[2], Fabiana Bigi[3]*, Marcelo Abel Soria[1]*

[1] Universidad de Buenos Aires. Facultad de Agronomía. Cátedra de Microbiolog[ia Agrícola.INBA-CONICET. Buenos Aires, Argentina.

[2] Instituto Nacional de Enfermedades Infecciosas-ANLIS Carlos Malbrán, Buenos Aires, Argentina.

[3] Instituto de Biotecnología, Instituto Nacional de Tecnología Agropecuaria (INTA), Argentina.

[4] IMEX-CONICET, Academia Nacional de Medicina, Buenos Aires, Argentina.

[5] Departamento de Química Biológica, e IQUIBICEN-CONICET, FCEyN, UBA, Buenos Aires, Argentina.

[6] Plataforma de Bioinformática Argentina, Instituto de Cálculo, FCEyN, UBA, Buenos Aires, Argentina.

* corresponding authors

E-mail: bigi.fabiana@inta.gob.ar

N. Repetto and De los Reseros, Hurlingham, 1686; Buenos Aires, Argentina

Tel: +54 11 4621 1447 ext 172

E-mail: soria@agro.uba.ar

Av. San Martín 4453, C1417DSE, Buenos Aires, Argentina

Tel: +54 11 4524 8000

Author details:

MMB: mebigi@hotmail.com

Av. San Martín 4453, C1417DSE, Buenos Aires, Argentina

Tel: +54 11 4524 8000

FCB: blanco.federico@inta.gob.ar

N. Repetto and De los Reseros, Hurlingham, 1686; Buenos Aires, Argentina

Tel: +54 11 4621 1447 ext 172

BL: bealopez@anlis.gov.ar
VR: vivianaritacco@gmail.com
Av. Vélez Sarsfield 563, C1282AFF, Buenos Aires. Argentina
Tel: +54 11 4303 1809

MCS: mariadelcarmen.sasiain@hotmail.com
SDLB: sildelabar@gmail.com
José Andrés Pacheco de Melo 3081, C1425AUM, Buenos Aires, Argentina
+54 11 4805 5759
MAM: marti.marcelo@gmail.com
EJS: ezequieljsosa@gmail.com
DAFDP: dariofd@gmail.com
Intendente Güiraldes 2160, C1428EGA, Buenos Aires, Argentina
+54 11 4576 3380

Abstract

Globally, about 4.5% of new tuberculosis (TB) cases are multi-drug-resistant (MDR), i.e. resistant to the two most powerful first-line anti-TB drugs. Indeed, 480,000 people developed MDR-TB in 2015 and 190,000 people died because of MDR-TB. The MDR *Mycobacterium tuberculosis* M family, which belongs to the Haarlem lineage, is highly prosperous in Argentina and capable of building up further drug resistance without impairing its ability to spread. In this study, we sequenced the whole genomes of a highly prosperous M-family strain (Mp) and its contemporary variant, strain 410, which produced only one recorded tuberculosis case in the last two decades. Previous reports have demonstrated that Mp induced dysfunctional CD8+ cytotoxic T cell activity, suggesting that this strain has the ability to evade the immune response against *M. tuberculosis*. Comparative analysis of Mp and 410 genomes revealed non-synonymous polymorphisms in eleven genes and five intergenic regions with polymorphisms between both strains. Some of these genes and promoter regions are involved in the metabolism of cell wall components, others in drug resistance and a SNP in *Rv1861, a gene* encoding a putative transglycosylase that produces a truncated protein in Mp. The mutation in Rv3787c, a putative S-adenosyl-L-methionine-dependent methyltransferase, is conserved in all of the other prosperous M strains here analysed and absent in non-prosperous M strains. Remarkably, three polymorphic promoter regions displayed differential transcriptional activity between Mp and 410. We speculate that the observed

mutations/polymorphisms are associated with the reported higher capacity of Mp for modulating the host's immune response.

Key words: *Mycobacterium tuberculosis*, MDR, Haarlem, genome sequencing, polymorphisms

Introduction

TB development depends on the host's natural resistance/susceptibility to *Mycobacterium tuberculosis* (*Mtb*) infection and differences in transmissibility, virulence, and immunogenicity among *Mtb* strains. In turn, these latter bacterial factors are determined by the genetic background of the organisms. In this line, certain strains of *Mtb* with special transmission potential are able to manipulate host immunity, which could impact on the evolution and/or outcome of the disease [1][2]. Multidrug-resistant tuberculosis (MDR-TB) poses a threat to the control and elimination of TB, which, despite being a curable and preventable disease, is still a major public health problem.

In Argentina, 9,605 new cases of TB were reported in 2014, with an incidence of 22.5 per 100,000 inhabitants and MDR-TB was documented in 101 cases. MDR-TB outbreaks emerged in the early 1990s in AIDS patients and thereafter disseminated to immunocompetent individuals [3]. Epidemiological, bacteriological, and genotyping data allowed the identification of certain MDR *Mtb* outbreak strains, such as the M strains of the Haarlem family and the Ra strain of Latin America and the Mediterranean (LAM). Each of these two strains managed to perpetuate in their geographical niches, the metropolitan areas of Buenos Aires and Rosario cities, respectively [4]. In particular, the M strains are still prosperous in the country and can build up further drug resistance without impairing their ability to spread [5]. In contrast, its clonal variant 410 has caused a single tuberculosis case since the onset of the outbreak [6].

Previous studies have demonstrated that Mp, a highly prosperous M family member, and 410 strains modulate the host immune response in different ways [7][8]. Particularly, Mp induced dysfunctional CD8+ cytotoxic T cell (CTL) activity, while 410 elicited a CTL response similar to that of the *M. tuberculosis* reference strain, H37Rv [9][10]. CTL activity has been associated with lysis of *Mtb*-infected macrophages [11] and with reduction in *Mtb* viability [12]. Therefore, the reduced CTL activity of Mp could be considered as part of an evasion mechanism to leave the bacterial niche intact, thus allowing the persistence and its successful spreading to the community. Despite the increased cost in public health importance of MDR-TB, the bacterial factors involved in the immune evasion mechanisms and the consequent ability to cause outbreaks are still unknown.

Relevant pathogenic differences between *Mtb* strains may be revealed by genomic comparison. With this premises in mind, we performed whole genome sequencing of two variants of the M strains: the highly prosperous Mp strain, and a non-prosperous close relative, the 410 strain.

The sequence alignment of both genomes against the reference Haarlem strain allowed us to identify genomic differences between Mp and 410 strains. These differences could explain the contrasting immune phenotypes that both strains induce *in vitro* and likely the higher capacity of Mp to perpetuate in the community, compared to 410.

Results and discussion

Genome sequencing and assembly

Whole-genome sequence reads from Mp and 410 were assembled using as reference the genomic sequence of Haarlem strain (see Materials and Methods for accession numbers). Specifically, we generated the genome assembly of strain 410 by aligning 7.91 million of reads producing a 169X average mapping coverage with a fraction on non-mapped reads of 6.14 %. With respect to strain Mp, we generated its genome assembly by aligning approximately 4.67 million reads, with an average coverage of 105X and 7.31 % of non-mapped reads.

Comparative analysis of the assembled genomes to that of reference Haarlem strain showed 345 and 362 single nucleotide mismatches in 410 and Mp, respectively. In addition, 48 insertions and deletions (INDELs) in 410 and 46 INDELs in Mp were detected in the open reading frames (ORFs). Besides, assembly gaps were also present in Mp and 410 (Figure 1). This finding suggests deletions, as previously demonstrated in the Haarlem lineage [13].

We identified 4,067 and 4,066 genes for Mp and 410 strains, respectively. In Mp strain the annotation pipeline recognized a small region of 171 nt that bears similarity to gene *rv2819c* (1128 nt) and annotated the smaller fragment as a full gene. This annotation error might explain the extra gene assigned to Mp compared to 410.

Two additional contigs for each genome were built *de novo* with the reads that did not map to Haarlem strain. One of these contigs, carrying 12 genes, was very similar in Mp and 410, with eleven genes having a high level of similarity. These genes, which are present in other *M. tuberculosis* complex genomes, encode possible transposases and proteins similar to Esat6 and PPE protein families.

Comparative analysis of predicted proteins and promoter regions in MDR *Mtb* strains

Comparing to the Haarlem reference strain, we found that sixteen genes showed one SNP in either Mp, 410 or both. We identified additional SNPs but they mapped in regions with abundant gaps and undetermined sequences and were discarded from this analysis.

Table 1 shows that three out of the 16 relevant polymorphisms were synonymous SNPs and 13 were non-synonymous (NS-SNPs). These polymorphisms were distributed in 11 genes, and confirmed by PCR with specific primers (SI 3).

Mp and 410 showed two different SNPs in *rpoB* (*rv0667*) and *pncA* (*rv2043c*). In total, five amino acid changes were present in Mp and five in 410: a nucleotide deletion produced a predicted reading frameshifting for the Rv1861 protein of Mp and the insertion of three nucleotides added a proline in Rv0668 of Mp. Finally, a SNP was present in the 16S RNA of Mp strain and this SNP was associated to kanamycin resistance.

The analysis of promoter regions showed five intergenic regions with at least one SNP exclusive of Mp compared to the Haarlem genome sequence (Figure 2). We identified -10 and -35 boxes of σ^{70} recognition sites in four of the five non-coding intergenic regions using the BPROM online service.

A SNP was identified upstream of *rv0010c*, *eccA3* (*rv0282*) and *rv1682*. In addition, the intergenic regions between *rv2172c*-*idsA2* (*rv2173*) showed two consecutive SNPs, while that of *rv3253c*-*rv3254* carries a 12-bp duplication.

To assess the importance of the SNPs in activity of the promoter regions, we compared the transcriptional level of the genes surrounding all polymorphic promoter regions. Figure 3 shows the expression of *rv0010, rv2172c, idsA2* and *rv3254* was upregulated in 410 strain as compared to Mp strain. The rest of the tested genes did not show significant difference in their expression between strains under the experimental conditions used in this study.

Polymorphisms in other isolates of the M strain family

We then extended the analysis of the polymorphic proteins and promoter regions (between Mp and 410) to other 16 M genomes obtained from the European Nucleotide Archive public database[5]. These 16 strains were MDR isolated from patients in Argentina from 1998 to 2008 [5], eight were prosperous (caused outbreaks) and eight were non-prosperous (caused isolated cases). Polymorphisms in the antibiotic resistant-related genes *rpoB*, *pncA*, *embB* (*rv3795*), and *16S-rRNA* gene were conserved in all prosperous strains but absent in non-prosperous strains (Figure 4). Interestingly, the mutation in *rpoB* of 410 strain differs from that of the other non-prosperous M strains. This result suggests that this mutation may be responsible for the slower replication rate observed in vitro for this strain (data not shown).

Regarding to the genes not reported as involved in antibiotic resistant mechanism, we found that the polymorphism in *rv3787c* was conserved in all prosperous strains (9/9), while polymorphisms in

rv1861 were present in four out of nine prosperous strains. Mutations in *cmaA2* (*rv0503c*),, *xseA* (*rv1108*), *fadD5* (*rv2187*) and *rv3193* were exclusive of either Mp or 410 and not conserved in the other M strains investigated. The polymorphism detected in the intergenic regions of *rv2172c-idsA2* was conserved in 5/9 prosperous strains. The SNP at the -35 box of *eccA3* was also present in 4/9 prosperous strains. Only one non-prosperous strains, apart from 410, conserved the SNP mapped upstream of *rv1682.*

To better understand how the Mp and 410 strains relate to the above described 16 prosperous and non-prosperous M strains isolated in Argentina , we scanned them all for SNPs. It is important to note here that since we scanned the 18 genomes, not all of SNPs were polymorphic between Mp and 410. This allowed us to analyze the polymorphism patterns across isolates without restricting to those polymorphisms that only separated Mp and 410.

We found 165 SNPs that were present in all 18 strains. One hundred parsimonious trees were built with kSNP3.0 and the consensus tree is shown in figure 5. The Mp strain clustered in a group with a very high level of support (95%) that includes the eight prosperous strains. Within this cluster, the Mp strain forms a high-support (72%) subcluster with four other strains. On the other hand, six non-prosperous strains form a cluster with a very high level of support (94%). The 410 strain joins this group with a quite lower level of support (64%). Two other non-prosperous strains join this combined group at greater distances, and comparable low support levels.

These results taken together suggest that the prosperous strains form more compact groups of related strains, probably determined by selection pressure. While some non-prosperous are similar, in general these strains seem to be less genetically related.

Description of polymorphic promoters and genes

Secondary mutations in genes encoding the RNA polymerase subunits RpoA and RpoC (Rv0668) alleviate the fitness cost incurred by mutations in the region of *rpoB* determining rifampicin resistance[14]. Such compensatory mutations have been most frequently described in association with *rpoB* S450L, which is the predominant rifampicin resistance-conferring mutation worldwide [15] and previously identified in the M outbreak [5]. However, not all secondary mutations in *rpo* genes are exclusively associated to fitness restoration. In our study, we found that in addition to *rpoB* S450L (S456L for Haarlem annotation), Mp carries a secondary *rpoC* mutation (INS739P); this mutation was neither detected in the other strains analysed in this study (Table 1, Figure 4) nor in any other *Mtb* strains. Therefore, a role in transmission for this secondary mutation cannot be assigned. On the other hand, we found that strain 410 carries *rpoB* mutation H445L (H451L for Haarlem annotation) (Table 1, Figure 4), which was absent from all M outbreak strains, and lacks any secondary *rpo* mutation.

We also found that PncA (Rv2043c), which is a pyrazinamidase/nicotinamidase involved in susceptibility or resistance to antituberculous drug pyrazinamide, was mutated in different positions in both strains (Table 1, Figure 4).

Most of the polymorphisms identified in this study localized in proteins involved in cell wall biogenesis such as EmbB (Rv3795), which is a membrane indolylacetylinositol arabinosyltransferase that is involved in the transfer of arabinose to the cell wall acceptor [16]. Mutations in *embB* have been related to ethambutol resistance [17]. Mp carried mutations in other cell wall-related proteins such as Rv3787c, Rv1861 and Rv3193c (Table 1, Figure 4). Rv3787c has an S-adenosylmethionine (SAM)-dependent methyltransferase domain that may be involved in the synthesis of methylbranched short-chain fatty acids [18]. Rv3193c and Rv1861 are transmembrane proteins of unknown functions with either ABC transporter or putative transglycosylase associated protein domains, respectively (http://tuberculist.epfl.ch/). Although there is no available information in the literature on Rv1861 function, we could speculate a role in glycosylation of lipids, proteins or other complex molecules and the deletion of a single base in *Rv1861* detected in Mp would translate a truncated protein.

Only 410 contained polymorphisms in FadD15 and CmaA2 (Table 1, Figure 4). Although mutations in *fadD* genes may have a compensatory role in drug resistance [19], no experimental evidence supports this hypothesis. CmaA2 is an S-adenosylmethionine (SAM)-dependent methyltransferase involved in the trans cyclopropanation of methoxy- and ketomycolates [20]. In this respect, clinical strains of *Mtb* have higher trans-cyclopropane content, as measured by the content of methoxymycolates, than strains that were extensively propagated in laboratories [21]. These results are consistent with a relevant role of CmaA2 during the infection of *Mtb* and may suggest that the mutation on this protein in 410 strain leads to an alteration in the enzyme functionality. It is remarkable that two proteins with S-adenosyl-l-methionine (SAM)-dependent methyltransferase domains (Rv3787c and CmaA2) were polymorphic. *Mtb* encodes several S-adenosylmethionine (SAM)-dependent methyltransferases with a potential or demonstrated role in mycolic acid modifications, four of which participate in the cyclopropanation of mycolic acids (MmaA2, MmaA1, CmaA2 and PcaA) [18]. Substantial evidence demonstrates that modified mycolic acids play specific and diverse roles in *Mtb* pathogenesis [22][23][24]. Indeed, the complete lack of mycolate cyclopropanation suppresses the host immune response [18]. Therefore, our findings suggest that the NS-SNPs identified in S-adenosyl-l-methionine (SAM)-dependent methyltransferase genes would have a role in the low CTL response that induces Mp *in vitro* [9][10].

As shown above (Figure 3), mutation on three promoter regions would potentially impair the expression of *rv0010c, rv2172c, idsA2* and *rv3254* in Mp strain. *rv0010c* encodes a conserved membrane protein with unknown functions with a probable role in cell wall and cell processes (tuberculist.epfl.ch/). *idsA2* and *rv2172c* express proteins likely involved in lipid biosynthesis and a conserved hypothetical protein (tuberculist.epfl.ch/), respectively, while *rv3254* encodes a conserved hypothetical protein. The polymorphism upstream of *rv3254* is a 12-bp duplication that maps at a potential sigma factor F (SigF) binding site. SigF is upregulated upon nutrient depletion of *Mtb* cultures and during infection of cultured human macrophages [25]. As mentioned previously, the polymorphism in the promoter region *rv2172c- idsA2* is conserved in 5/9 prosperous M strains and absent in the non-prosperous M strains included in this study Figure (4). This finding suggests that these down-regulated promoters may have a biological relevance in the contrasting phenotype reported for Mp and 410 strains [9][10]. By contrast, mutation in promoter regions of *rv0282* and *rv1682* do not affect the expression of their corresponding genes in the culture condition used in this study. *rv0282-eccA3* operon encodes components of an ESX-3 type VII secretion system as well as PE, PPE and ESAT6-like proteins [25]. This operon seems to be regulated by two transcriptional regulators, IdeR and Zur, and the mutation is localized in the promoter region that overlaps the IdeR binding site upstream of *rv0282* in the Mp genome. IdeR is an iron-dependent regulatory protein essential in *Mtb* that functions as a repressor [26].

Although *rv1682* encodes a probable coiled-coil structural protein, the function of Rv1682 is unknown. However, this gene is expressed within the macrophages in response to the short-term acidification occurring immediately upon bacterial invasion (2007) [27].

Conclusion

In this study, we found allelic variants in ten proteins, SNPs/INDELs in five promoter regions and in the 16S RNA gene, between Mp and 410 strains. Four polymorphic proteins (Rv3795 Rv3787c, Rv1861 and Rv3193c) and two polymorphic promoters (*rv0010* and *idsA2*) have potential or demonstrated role in the bacterial cell wall biogenesis. Noticeably, excluding drug-resistance associated mutations, only one protein polymorphism (Rv3787c) was conserved in Mp and all eight previously sequenced prosperous isolates of the M outbreak. This polymorphic protein is a putative methyltransferase with a possible role in cell wall biogenesis. Importantly, polymorphisms in three promoter regions seem to have an impact on the expression of their downstream genes, being the deletion in the *rv2172c-idsA2* region conserved in other four M prosperous strains. Given the relevance of the bacterial cell wall components in the interaction of *Mtb* with the host, the presence of polymorphisms in promoters and genes involved in its biogenesis should help to improve our

comprehension of host–pathogen interactions in tuberculosis. The study of these polymorphisms could also help to explore the role of the cell wall components in the transmission/evasion of immune response mechanisms of *Mtb*.

In summary, we identified few functional polymorphisms in two closely related *Mtb* strains with contrasting biological characteristics. These mutations/polymorphisms could be associated with the impaired CTL response induced by Mp that may in turn contribute to the successful spreading of this strain into the Argentinean population. However, further experimental studies are required to individually determine the impact of these genomic polymorphisms in the differential phenotype of Mp and 410.

Materials & Methods

Characteristics of the *M. tuberculosis* isolates

Mtb Mp and 410 strains were obtained from the collection kept at the Reference Laboratory for Mycobacteria at the INEI-ANLIS ‘‘Carlos G. Malbrán’’ in Buenos Aires. Both strains were isolated from an HIV positive and a negative patient, respectively, in Buenos Aires. They belong to the Haarlem H2 genotype (SIT 2) and are resistant to isoniazid, rifampicin, streptomycin and pyrazinamide. Additionally, Mp is resistant to ethambutol and kanamycin. Genotyping studies have shown that both strains are genetically close to each other with identical spoligotype, similar RFLP patterns (one additional band in 410) and only one difference in VNTR-MIRU 15 loci (254533233433537 for Mp and 254633233433537 for 410).

Other 16 isolates of the M outbreak (as assessed by IS6110 RFLP) publicly available with access numbers **ERR760925**, **ERR760768**, **ERR760778**, **ERR760819**, **ERR760817**, **ERR760753**, **ERR760759,** **ERR760882**, **ERR760843**, **ERR760849**, **ERR760749**, **ERR757170**, **ERR760762**, **ERR760820**, **ERR760842** and **ERR760785** were included in the study (see Eldholm 2015 for a complete list of isolates and their sources).

DNA extraction

Mycobacterial cultures were grown for 20 days at 37 °C on Middlebrook 7H10 Agar supplemented with 0.5% glycerol, 0.4% glucose and 0.5% albumin. Chromosomal DNA samples were obtained from loops of cultures following the protocol described in Van Soolingen et al (1991) [28] and then checked by agarose gel electrophoresis.

Genome sequencing and bioinformatics analysis

The genome sequences of *Mtb* strains Mp and 410 were generated by combining Roche 454 (Roche GS FLX+ system, v2.9) and Illumina HiSeq 2000 paired-ended (2x100 bp) technologies. The

MIRA 4.0 [29](Mimicking Intelligent Read Assembly) software was used for sequence assembling. The program Prokka (Rapid prokaryotic genome annotation, version 1.10) [30] was used to annotate the M and 410 genomes. Individual gene and protein alignment were done in a local installation of blast [31]. The open reading frames (ORFs) and protein sequences, of the reference strains H37Rv and Haarlem were retrieved from the NCBI's RefSeq collection with accession numbers NC_000962 and NC_022350 respectively and stored in a MySQL database. This database was used to search for single nucleotide polymorphisms (SNP) using blast. The sequencing reads of the other six M isolates were downloaded from the European Nucleotide Archive PRJEB7669 (http://www.ebi.ac.uk/ena/data/view/PRJEB7669) and were aligned to the Haarlem genome with MIRA 4.0. Protein sequences were analysed with the TuberQ database [32].

Accession codes: genome sequences of *Mtb* strains Mp and 410 were deposited in National Center for Biotechnology Information database under accession code PRJNA317008 and PRJNA317008 respectively.

For the analysis of intergenic regions, a custom Perl script was used to extract the sequences, followed by blast to search for differences. In addition, the BPROM on-line service (http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb) was used to find the -35 and -10 boxed of σ^{70} recognition sites.

kSNP 3.0 was used to scan SNPs across 18 genomes and to build consensus parsimonious tree. One hundred equivalent parsimonious trees were constructed and support values and consensus trees were derived from them [33].

PCR and Sanger sequencing of polymorphic regions

DNA fragment encompassing polymorphic regions were PCR amplified with the primers listed in Supplementary material 1 following a standard method. Amplicons were then purified from PCR reactions and sequenced by Sanger method at the Unit of Genomic and Bioinformatic (UGB) of INTA (http://inta.gob.ar/documentos/formularios-del-servicio-secuenciacion-de-acidos-nucleicos-adn-por-electroforesis-capilar/). DNA sequences were visualized and analysed with 4Peaks and Strider software, respectively.

Expression analysis

RNA extraction: 50 ml of three independent cultures of Mp and 410 were harvested at the exponential phase of growth (Optical density 600nm: 0.3-0.4). The cell pellets were immediately resuspended in 1 ml of Trizol (Sigma) and transferred to a 2-ml screw-cap microcentrifuge tube containing 0.1 mm silica glass beads (Sigma-Aldrich). Cells were disrupted with a Fastprep FP120 bead-beater for 20 s at a speed of 6 m s −1. The samples were treated twice with 200 µl of

chloroform, centrifuged at 9.000 g for 5 min and the nucleic acids present in the upper phases (aqueous phases) were precipitated with isopropanol. The RNA/DNA pellets were washed up with ethanol 70% and resuspended in Rnase-free water. Finally, the samples were cleaned up with RNeasy MinElute Cleanup Kit (Qiagen) and treated with DnaseI Ambion (LifeTechnology) following the manufacture's specifications.

RT-qPCR: RT-qPCR reactions were performed as previously described [34] using specific primers (Supplementary material 1) and DNA-free RNA (1µg) extracted from mid-exponential growth-phase cultures of M and 410. Briefly, RNA (1 µg) was mixed with 50 ng of random primers (Invitrogen) in 20 µl of final volume and reverse-transcribed to total cDNA with SuperScript II reverse transcriptase (Invitrogen, Life Technologies) following the manufacturer's instructions. Control reactions without reverse transcriptase were included.

The cDNA (0.5 µl) was used as template for each RT-qPCR reaction. All primers were designed using Primer 3 Software (bioinfo.ut.ee/primer3-0.4.0/) (Supplementary material 1). The qPCR reactions were performed with Taq Platinum DNA polymerase (Invitrogen, Life Technologies) and SYBR reagent (Thermo Fisher, Life Technologies) following the manufacturer's instructions.

All qPCR reactions were performed in duplicate and average values of duplicates were analysed using the LinRegPCR software [35] with default settings. All samples without plateau or amplification and with very low Cq value were excluded for mean efficiency calculation. A strictly continuous log-linear setting was used for baseline estimation and the excluded samples were analysed individually and corrected with the manual correction baseline option. The fold change was calculated using *sigA* as the reference gene. The final results and permutation statistical analysis were assessed with fg statistic software, which is part of the Infostat software package [36]. For the statistical test, the parameters were set to defaults with 5,000 permutations at random.

Legends to the figures

Figure 1: Global comparison of Mp and 410 strain genomes compared to the *Mtb* Haarlem sequence. We aligned each of the genome sequence assemblies against the *Mtb* Haarlem genome using BLASTN. The innermost ring indicates genomic coordinates on the Haarlem strain. The concentric coloured rings indicate the presence or absence of BLASTN hits at that position, with one ring corresponding to each genome assembly. Positions covered by BLASTN alignments are indicated with a solid colour. White gaps represent genomic regions not present in Haarlem. The features present in the other two strains that map on the gaps are indicated and their coordinates are coloured according to the strain (410, green; Mp, blue). The graphical view of the alignments was rendered using BLAST Ring Image Generator (BRIG) [37].

hp: hypothetical protein; PPE34: PPE family protein; LppA: Probable conserved lipoprotein; Cas2: CRISPR-associated endonuclease; Csm6:CRISPR-associated protein; Cas1:CRISPR-associated endonuclease; Csm5: CRISPR-associated protein; MoaX: Probable MoaD-MoaE fusion protein; MoaC3: Probable molybdenum cofactor biosynthesis protein C 3; MoaB3: Probable pterin-4-alpha-carbinolamine dehydratase; MoaA3: Probable molybdenum cofactor biosynthesis protein A; EmbR2: Probable transcriptional regulatory protein; PE-PGRS family protein

Figure 2: Schematic representation of polymorphic intergenic regions. Predicted -10 and -35 boxes in the intergenic regions and the surrounding genes are shown. Positions of SNPs from start codons are indicated and the mutated bases in Mp and 410 strains (regarding to Haarlem genome) are labelled in red. An insertion in the *Rv3253c-Rv3254* intergenic region is indicated with a triangle. Non-prosperous M strains are in bold.

Figure 3: Expression of genes surrounding polymorphic promoters between 410 and Mp strains obtained by RT-qPCR. Fold-changes were calculated using *sigA* mRNA expression as reference gene and Mp strain as calibrator. Data were analysed using a random permutation test (fg statistic software * $p<0.05$). Bars represent average expression ratios ± SD between Mp and 410 strains.

Figure 4: Allelic variants of polymorphic proteins (A) and intergenic regions (B) in prosperous and non-prosperous M strains. The mutations of each variant are described in the bottom chart.

Figure 5: Consensus parsimonious tree derived from 165 SNPs detected in 18 strains of *M. tuberculosis* (strains Mp and 410, this study; the other 16 were reported by Eldholm et al. (2015). The bar represents changes per number of SNPs.

Acknowledgments

We thank Dr Julia Sabio y García for critical reading of the manuscript. FCB, MCS, SDB, VR, MAM, DAFD and FB are CONICET fellows. This study was supported by PNBIO1131034-INTA and CONICET-NIH 2014-2016 Res. 2322.

Supporting information captions:

Supplementary information Table S1: Primer used in this study.

Funding statement

This study was funded by INTA project PNBIO1131034.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

There are not competing interests of authors

References

[1] Manca C, Reed MB, Freeman S, Mathema B, Kreiswirth B, Barry CE, et al. Differential monocyte activation underlies strain-specific Mycobacterium tuberculosis pathogenesis. Infect Immun 2004;72:5511–4. doi:10.1128/IAI.72.9.5511-5514.2004.

[2] Sharma MK, Al-Azem A, Wolfe J, Hershfield E, Kabani A. Identification of a predominant isolate of Mycobacterium tuberculosis using molecular and clinical epidemiology tools and in vitro cytokine responses. BMC Infect Dis 2003;3:3. doi:10.1186/1471-2334-3-3.

[3] Ritacco V, Di Lonardo M, Reniero A, Ambroggi M, Barrera L, Dambrosi A, et al. Nosocomial spread of human immunodeficiency virus-related multidrug-resistant tuberculosis in Buenos Aires. J Infect Dis 1997;176:637–42. doi:10.1086/514084.

[4] Ritacco V, Iglesias MJ, Ferrazoli L, Monteserin J, Dalla Costa ER, Cebollada A, et al. Conspicuous multidrug-resistant Mycobacterium tuberculosis cluster strains do not trespass country borders in Latin America and Spain. Infect Genet Evol 2012;12:711–7. doi:10.1016/j.meegid.2011.06.006.

[5] Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, et al. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. Nat Commun 2015;6:7119. doi:10.1038/ncomms8119.

[6] Yokobori N, López B, Geffner L, Sabio y García C, Schierloh P, Barrera L, et al. Two genetically-related multidrug-resistant Mycobacterium tuberculosis strains induce divergent outcomes of infection in two human macrophage models. Infect Genet Evol 2013;16:151–6. doi:10.1016/j.meegid.2013.01.007.

[7] Basile JI, Geffner LJ, Romero MM, Balboa L, Sabio Y García C, Ritacco V, et al. Outbreaks of Mycobacterium Tuberculosis MDR Strains Induce High IL-17 T-Cell Response in Patients With MDR Tuberculosis That Is Closely Associated With High Antigen Load. J Infect Dis 2011;204:1054–64. doi:10.1093/infdis/jir460.

[8] Geffner L, Yokobori N, Basile J, Schierloh P, Balboa L, Romero MM, et al. Patients with multidrug-resistant tuberculosis display impaired Th1 responses and enhanced regulatory T-cell levels in response to an outbreak of multidrug-resistant Mycobacterium tuberculosis M and Ra strains. Infect Immun 2009;77:5025–34. doi:10.1128/IAI.00224-09.

[9] Geffner L, Basile JI, Yokobori N, Kviatcovsky D, Sabio y Garciá C, Ritacco V, et al. Mycobacterium tuberculosis multidrug resistant strain M induces an altered activation of cytotoxic CD8+ T cells. PLoS One 2014;9. doi:10.1371/journal.pone.0097837.

[10] Geffner L, Basile JI, Yokobori N, Sabio Y García C, Musella R, Castagnino J, et al. CD4(+) CD25(high) Foxp3(+) regulatory T lymphocytes suppress IFNγ and CD107 expression in CD4(+) and CD8(+) T cells from tuberculous pleural effusions. Clin Exp Immunol 2013;3:1–11. doi:10.1111/cei.12227.

[11] Canaday DH, Wilkinson RJ, Li Q, Harding C V., Silver RF, Boom WH. CD4(+) and CD8(+) T cells kill intracellular Mycobacterium tuberculosis by a perforin and Fas/Fas ligand-independent mechanism. J Immunol 2001;167:2734–42. doi:10.4049/jimmunol.167.5.2734.

[12] Cho S, Mehra V, Thoma-Uszynski S, Stenger S, Serbina N, Mazzaccaro RJ, et al. Antimicrobial activity of MHC class I-restricted CD8+ T cells in human tuberculosis. Proc Natl Acad Sci U S A 2000;97:12210–5. doi:10.1073/pnas.210391497.

[13] Cubillos-Ruiz A, Morales J, Zambrano MM. Analysis of the genetic variation in Mycobacterium tuberculosis strains by multiple genome alignments. BMC Res Notes 2008;1:110. doi:10.1186/1756-0500-1-110.

[14] De Vos M, Müller B, Borrell S, Black PA, Van Helden PD, Warren RM, et al. Putative compensatory mutations in the rpoc gene of rifampin-resistant mycobacterium tuberculosis are associated with ongoing transmission. Antimicrob Agents Chemother 2013;57:827–32. doi:10.1128/AAC.01541-12.

[15] Brandis G, Hughes D. Genetic characterization of compensatory evolution in strains carrying rpoB Ser531Leu, the rifampicin resistance mutation most frequently found in clinical isolates. J Antimicrob Chemother 2013;68:2493–7. doi:10.1093/jac/dkt224.

[16] Belanger AE, Besra GS, Ford ME, Mikusová K, Belisle JT, Brennan PJ, et al. The embAB genes of Mycobacterium avium encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol. Proc Natl Acad Sci U S A 1996;93:11919–24. doi:10.1073/pnas.93.21.11919.

[17] Telenti A, Philipp WJ, Sreevatsan S, Bernasconi C, Stockbauer KE, Wieles B, et al. The emb operon, a gene cluster of Mycobacterium tuberculosis involved in resistance to ethambutol. Nat Med 1997;3:567–70. doi:10.1038/nm0597-567.

[18] Barkan D, Hedhli D, Yan HG, Huygen K, Glickman MS. Mycobacterium tuberculosis lacking all mycolic acid cyclopropanation is viable but highly attenuated and hyperinflammatory in mice. Infect Immun 2012;80:1958–68. doi:10.1128/IAI.00021-12.

[19] Kuan CS, Chan CL, Yew SM, Toh YF, Khoo J-S, Chong J, et al. Genome Analysis of the First Extensively Drug-Resistant (XDR) Mycobacterium tuberculosis in Malaysia Provides Insights into the Genetic Basis of Its Biology and Drug Resistance. PLoS One 2015;10:e0131694. doi:10.1371/journal.pone.0131694.

[20] Barkan D, Rao V, Sukenick GD, Glickman MS. Redundant function of cmaA2 and mmaA2 in Mycobacterium tuberculosis cis cyclopropanation of oxygenated mycolates. J Bacteriol 2010;192:3661–8. doi:10.1128/JB.00312-10.

[21] Yuan Y, Crane DC, Musser JM, Sreevatsan S, Barry CE. MMAS-1, the branch point between cis- and trans-cyclopropane-containing oxygenated mycolates in Mycobacterium tuberculosis. J Biol Chem 1997;272:10041–9. doi:10.1074/jbc.272.15.10041.

[22] Dubnau E, Chan J, Raynaud C, Mohan VP, Lan??elle MA, Yu K, et al. Oxygenated mycolic acids are necessary for virulence of Mycobacterium tuberculosis in mice. Mol Microbiol 2000;36:630–7. doi:10.1046/j.1365-2958.2000.01882.x.

[23] Rao V, Gao F, Chen B, Jacobs WR, Glickman MS. Trans-cyclopropanation of mycolic acids on trehalose dimycolate suppresses Mycobacterium tuberculosis-induced inflammation and virulence. J Clin Invest 2006;116:1660–7. doi:10.1172/JCI27335.

[24] Rao V, Fujiwara N, Porcelli S a, Glickman MS. Mycobacterium tuberculosis controls host innate immune activation through cyclopropane modification of a glycolipid effector molecule. J Exp Med 2005;201:535–43. doi:10.1084/jem.20041668.

[25] Forrellad MA, Klepp LI, Gioffré A, Sabio Y García J, Morbidoni HR, Santangelo M de la P, et al. Virulence factors of the Mycobacterium tuberculosis complex. Virulence 2012;4.

[26] Forrellad MA, Klepp LI, Gioffré A, Sabio y García J, Morbidoni HR, de la Paz Santangelo M, et al. Virulence factors of the Mycobacterium tuberculosis complex. Virulence 2013;4:3–66. doi:10.4161/viru.22329.

[27] Rohde KH, Abramovitch RB, Russell DG. Mycobacterium tuberculosis Invasion of Macrophages: Linking Bacterial Gene Expression to Environmental Cues. Cell Host Microbe 2007;2:352–64. doi:10.1016/j.chom.2007.09.006.

[28] Van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol 1991;29:2578–86.

[29] Chevreux B, Wetter T, Suhai S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Comput Sci Biol Proc Ger Conf Bioinforma 1999:45–56. doi:10.1.1.23/7465.

[30] Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–9. doi:10.1093/bioinformatics/btu153.

[31] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421. doi:10.1186/1471-2105-10-421.

[32] Radusky L, Defelipe LA, Lanzarotti E, Luque J, Barril X, Marti MA, et al. TuberQ: a Mycobacterium tuberculosis protein druggability database. Database (Oxford) 2014;2014. doi:10.1093/database/bau035.

[33] Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics 2015;31:2877–8. doi:10.1093/bioinformatics/btv271.

[34] Blanco FC, Nunez-García J, García-Pelayo C, Soria M, Bianco MV, Zumárraga M, et al. Differential transcriptome profiles of attenuated and hypervirulent strains of Mycobacterium bovis. Microbes Infect 2009;11:956–63. doi:10.1016/j.micinf.2009.06.006.

[35] Ramakers C, Ruijter JM, Deprez RHLL, Moorman AF. M. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett 2003;339:62–6.

[36] JA di Rienzo, F Casanoves, Balzarini M, Gonzalez L, Tablada M, Robledo C. Infostat - Sofware estadístico. Universidad Nacional de Córdoba, Argentina. Univ Nac Córdoba, Argentina 2013. http://www.infostat.com.ar/.

[37] Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 2011;12:402. doi:10.1186/1471-2164-12-402.

Table 1: Polymorphisms in coding sequences of M strains

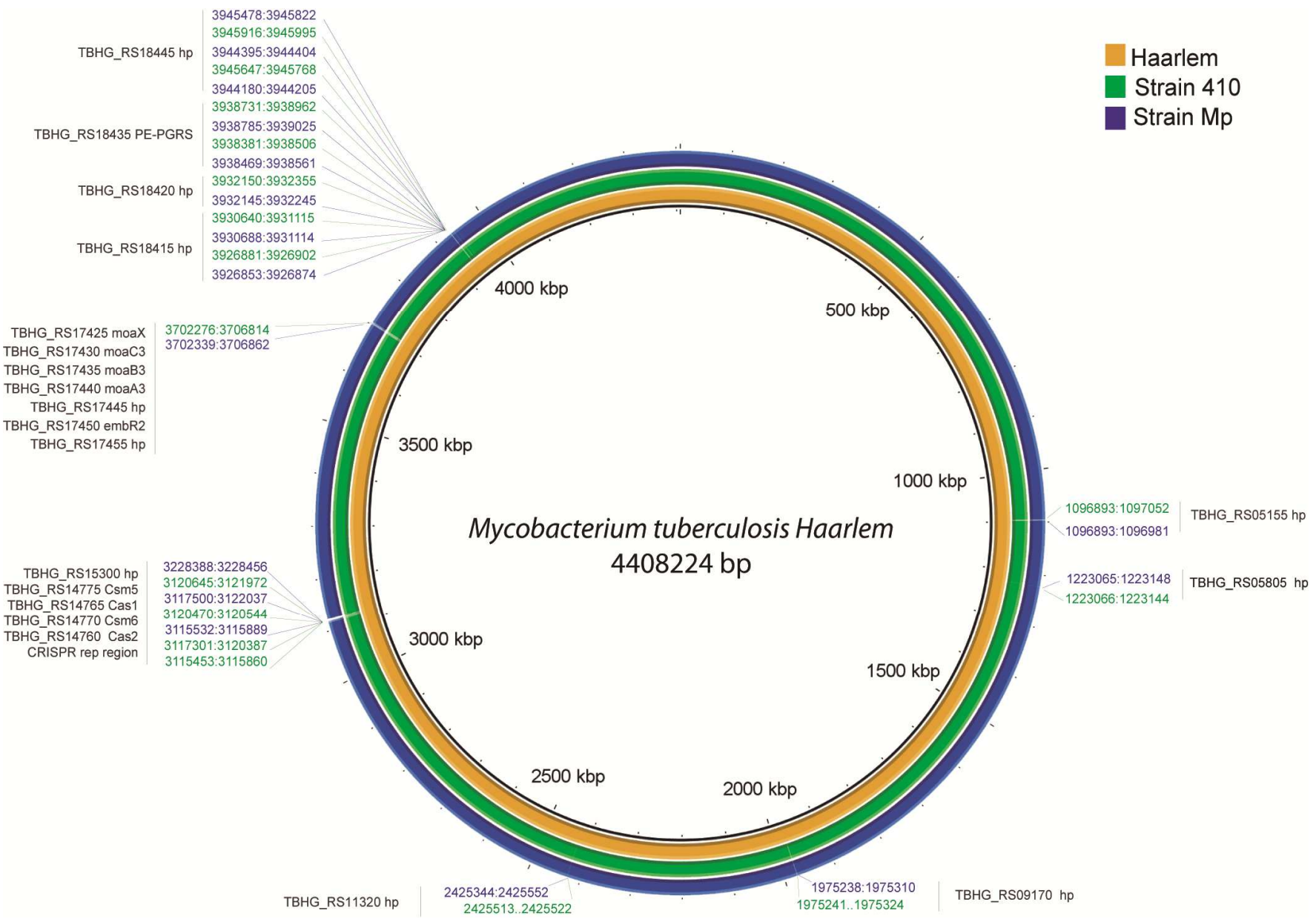
A) Non-Synonymous SNP

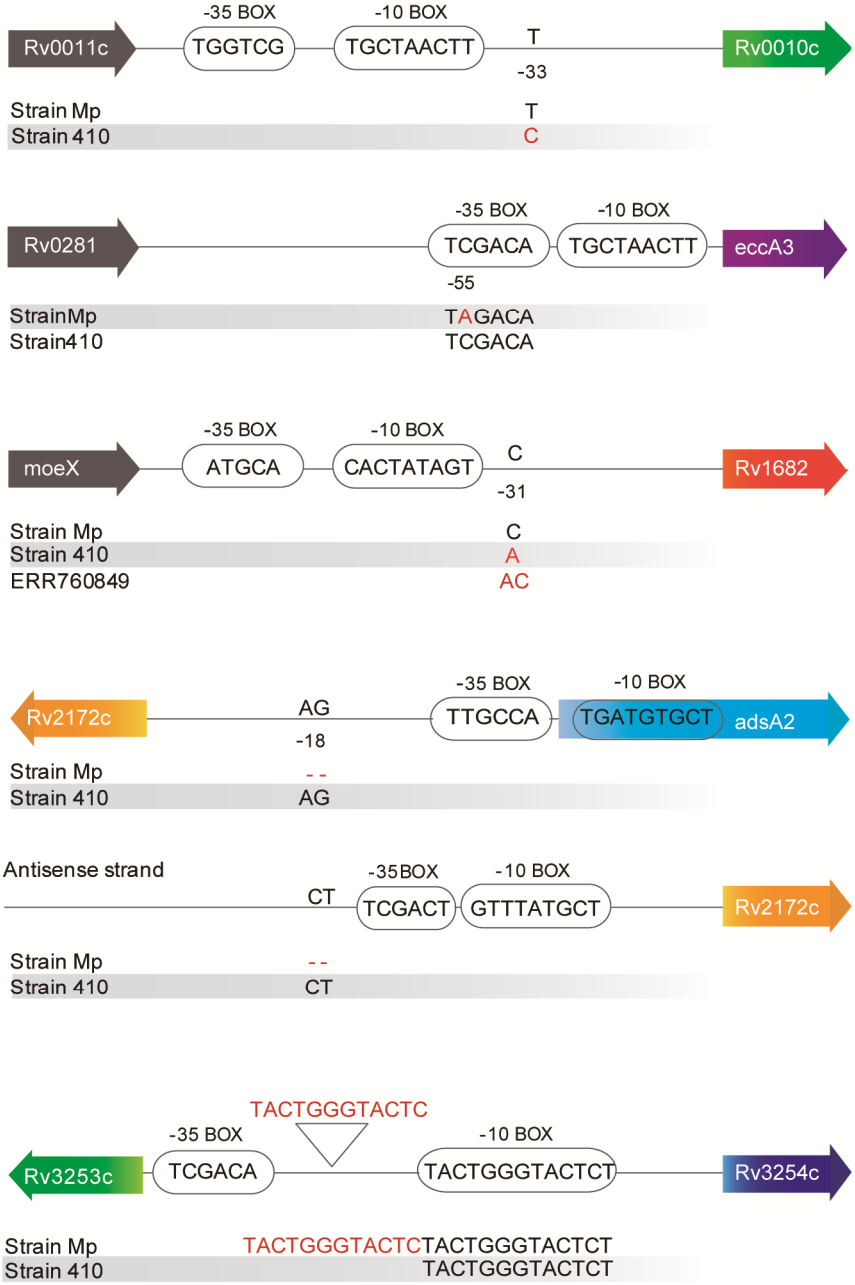
Rv number	Gene	Description	Strain Mp		Strain 410	
			SNP	AA change	SNP	AA change
Rv0503c	*CmaA2*	Cyclopropanemycolic acid synthase 2	-	-	a832c	M278L
Rv0667	*rpoB*	DNA-directed RNA polymerase subunit beta	c1368t	S456L	a1352t	H451L
Rv0668	*rpoC*	DNA-directed RNA polymerase subunit beta'	ins2213ccg	ins739P	-	-
Rv1108c	*xseA*	Exodeoxyribonucleaselargesubunit	-	-	c1126g	G376A
Rv1861		Transglycosylaseassociatedprotein	C303del	101fs	-	-
Rv2043c	*pncA*	Nicotinamidase /Pyrazinamidase	a29c	Q10P	t307g	Y103D
Rv2187	*FadD15*	Long-chain-fatty-acid--CoA ligase	-	-	g671t	G224V
Rv3193c		Transmembraneprotein	g2971a	A991T	-	-
Rv3787c		Putative S-adenosyl-L-methionine-dependent methyltransferase	c872t	A291V	-	-
Rv3795	*EmbB*	Integral membraneindolylacetylinositolarabinosyltransferase	g1217c	G406A	-	-
16s rRNA		16s rRNArrs	a1401g		-	-

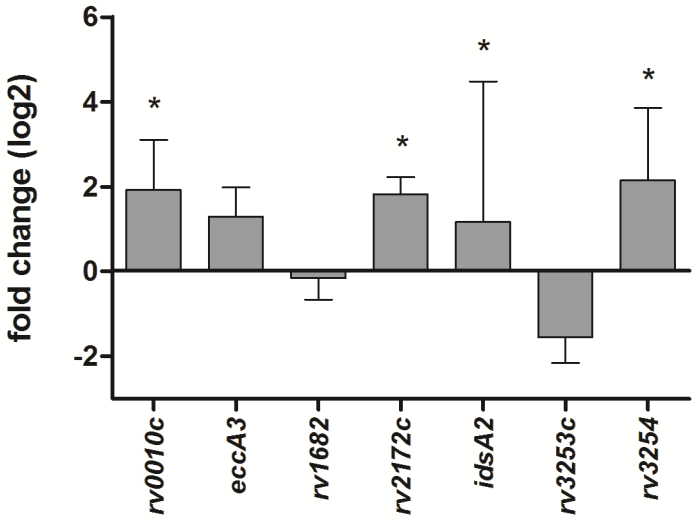
* H37Rv annotation: S450L (S456L), H445L (H451L), H445Y(H451Y) and H445R(H451R).

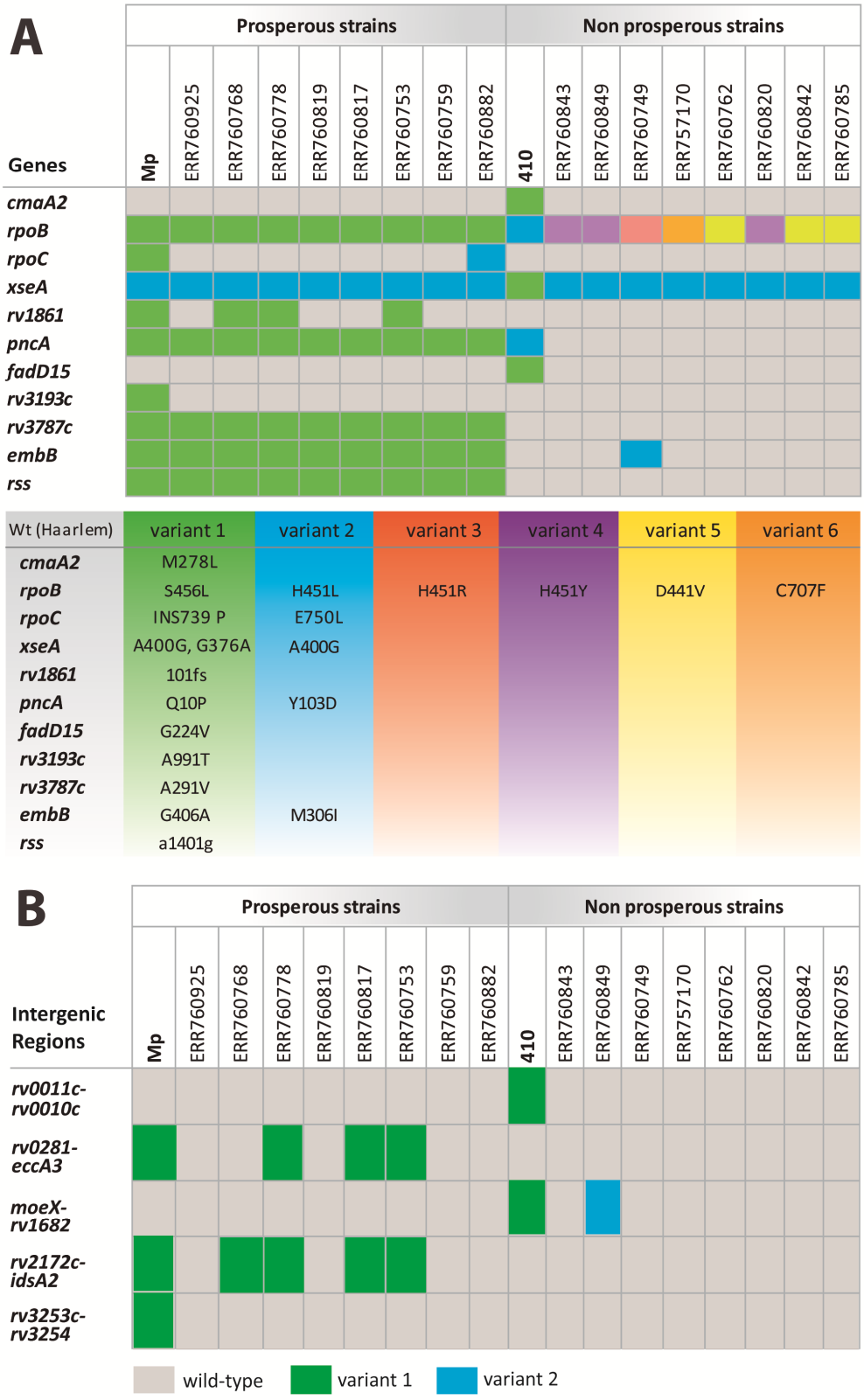
B) Synonymous SNP

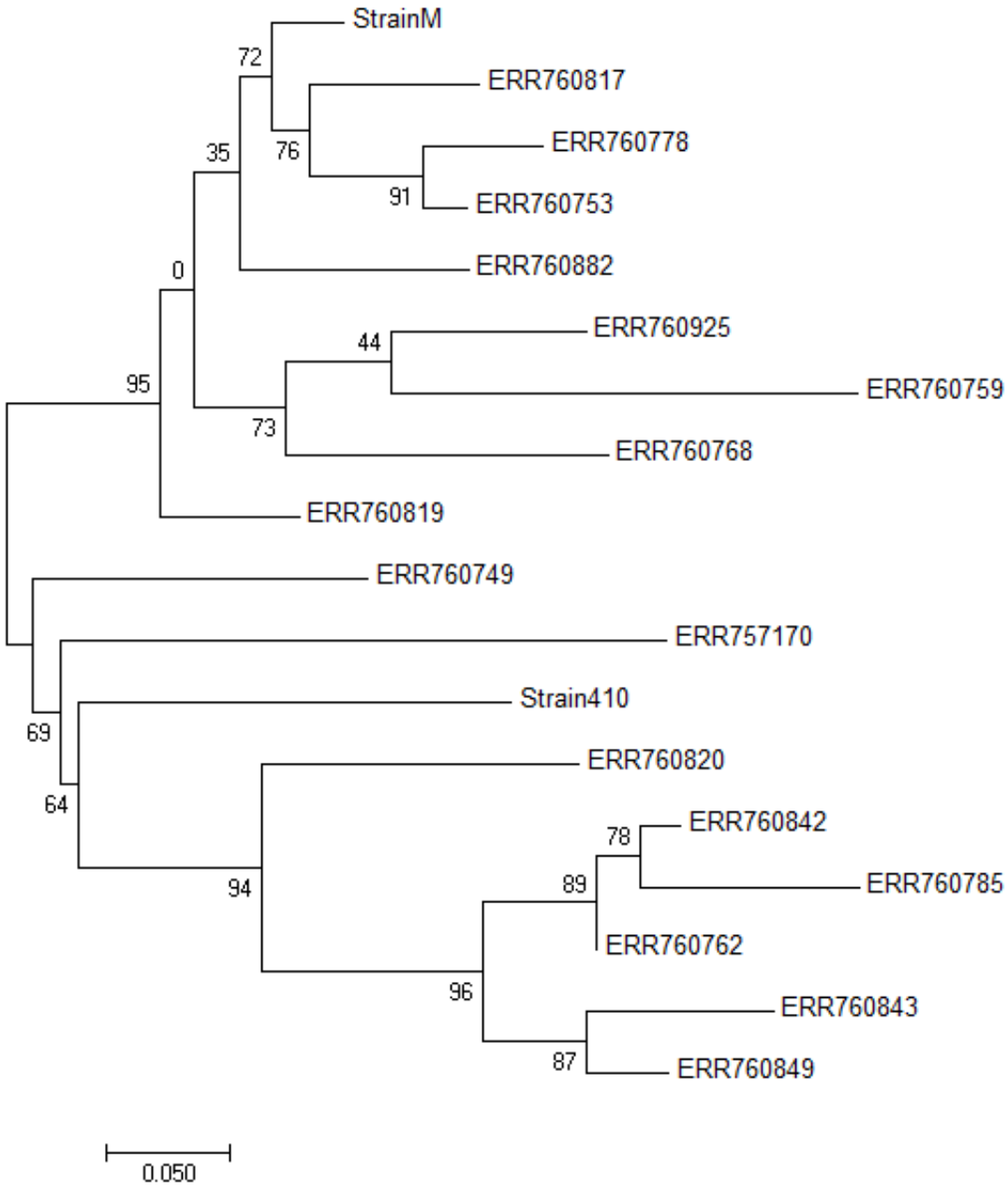
Rv number	Gene description	Location	Mp	410	Haarlem	Amino acid
Rv0650	Possible sugar kinase	346	TTG	CTG	CTG	L
Rv0266c	Probable 5-oxoprolinase opla	954	GTG	GTA	GTG	V
Rv0424c	Hypotheticalprotein	171	GAA	GAG	GAA	E




-35 BOX
-10 BOX
Rv0011c
TGGTCG
TGCTAACTT
T
-33
Rv0010c
Strain Mp
T
Strain 410
C
-35 BOX
-10 BOX
Rv0281
TCGACA
TGCTAACTT
eccA3
-55
StrainMp
TAGACA
Strain410
TCGACA
-35 BOX
-10 BOX
moeX
ATGCA
CACTATAGT
C
Rv1682
-31
Strain Mp
C
Strain 410
A
ERR760849
AC
-35 BOX
-10 BOX
Rv2172c
AG
TTGCCA
TGATGTGCT
adsA2
-18
Strain Mp
- -
Strain 410
AG
Antisense strand
-35BOX
-10 BOX
CT
TCGACT
GTTTATGCT
Rv2172c
Strain Mp
- -
Strain 410
CT
TACTGGGTACTC
-35 BOX
-10 BOX
Rv3253c
TCGACA
TACTGGGTACTCT
Rv3254c
Strain Mp
TACTGGGTACTCTACTGGGTACTCT
Strain 410
TACTGGGTACTCT




A
Prosperous strains
Non prosperous strains
Genes
Mp
ERR760925
ERR760768
ERR760778
ERR760819
ERR760817
ERR760753
ERR760759
ERR760882
410
ERR760843
ERR760849
ERR760749
ERR757170
ERR760762
ERR760820
ERR760842
ERR760785
cmaA2
rpoB
rpoC
xseA
rv1861
pncA
fadD15
rv3193c
rv3787c
embB
rss
Wt (Haarlem)
variant 1
variant 2
variant 3
variant 4
variant 5
variant 6
cmaA2 M278L
rpoB S456L H451L H451R H451Y D441V C707F
rpoC INS739 P E750L
xseA A400G, G376A A400G
rv1861 101fs
pncA Q10P Y103D
fadD15 G224V
rv3193c A991T
rv3787c A291V
embB G406A M306I
rss a1401g
B
Prosperous strains
Non prosperous strains
Intergenic Regions
Mp
ERR760925
ERR760768
ERR760778
ERR760819
ERR760817
ERR760753
ERR760759
ERR760882
410
ERR760843
ERR760849
ERR760749
ERR757170
ERR760762
ERR760820
ERR760842
ERR760785
rv0011c-rv0010c
rv0281-eccA3
moeX-rv1682
rv2172c-idsA2
rv3253c-rv3254
wild-type
variant 1
variant 2


StrainM
ERR760817
ERR760778
ERR760753
ERR760882
ERR760925
ERR760759
ERR760768
ERR760819
ERR760749
ERR757170
Strain410
ERR760820
ERR760842
ERR760785
ERR760762
ERR760843
ERR760849
72
76
35
91
0
44
95
73
69
64
78
89
94
96
87
0.050