Supplementary Information for

Multiple genetic paths including massive gene amplification allow *M. tuberculosis* to overcome loss of ESX-3 secretion system substrates

Lin Wang, Emmanuel Asare, Amol C. Shetty, Freddy Sanchez-Tumbaco, Megan R. Edwards, Rajagopalan Saranathan, Brian Weinrick, Jiayong Xu, Bing Chen, Angèle Bénard, Gordon Dougan, Daisy W. Leung, Gaya K. Amarasinghe, John Chan, Christopher F. Basler, William R. Jacobs Jr.*, and JoAnn M. Tufariello*
*correspondence: william.jacobs@einsteinmed.edu or joann.tufariello@mssm.edu

This pdf includes:

SI Materials and Methods

Figures S1 to S5

Tables S1 to S5

Legend for Dataset S1

Legend for Dataset S2

Legend for Dataset S3

SI references

Other supplementary materials for this manuscript include the following:

Dataset S1

Dataset S2

Dataset S3

SI Materials and Methods

Bacterial Strain Construction. Plasmids are detailed in **Table S3** and a list of strains is provided in **Table S4**.

Allelic exchange via specialized transduction was used to generate the deletion mutants (1, 2). In brief, the left and right flanks of each gene or gene pair were amplified by PCR from H37Rv genomic DNA and cloned into pYUB1471 using the indicated enzymes (for primers used in this study, see **Dataset S2**). Each allelic exchange substrate was then linearized with PacI, ligated into PacI-digested temperature-sensitive phAE159 and packaged *in vitro* into lambda phage particles (Maxplax, Epicentre), which were used to transduce *E. coli* HB101. The resulting

phasmids were transformed by electroporation into *M. smegmatis* mc^{2}155, and high-titer phage lysates, prepared as described (1-3), were used to transduce the indicated *Mtb* strains. Transductions were plated onto 7H10 agar supplemented with 10% OADC enrichment, 0.5% glycerol, 50-75 μg/ml hygromycin B, and 200 ng/ml mycobactin J (Allied Monitor) as required for the *esx-3* region deletion mutants. The deletions were confirmed by three-primer PCR using F, R, and universal tag primers specific for each deletion and were further verified by whole-genome sequencing using Illumina MiSeq technology (**Fig. S3**). The Δ*rv3058c* mutants used in aerosol infections of C57BL/6 mice and in transcriptomics experiments were unmarked at this locus by excising the *sacB*-hygromycin cassette using the phage phAE280 and sucrose selection as described (2). Successful unmarking was confirmed by PCR across the locus and sequencing of PCR products.

For complementation, the constructs were generated in pUC57-Kan by Genewiz (South Plainfield, NJ) and contained H37Rv *esxG-esxH* or *esxS-esxR* coding sequences, which extended from the ATG of *esxG* or *esxS* to, respectively, the stop codon of *esxH* or *esxR* and were flanked by an XmnI site-containing sequence (5' gaatcacttcgca 3') at the 5' end and by a HindIII site at the 3' end. The inserts were liberated by digestion with XmnI and HindIII, and subcloned downstream of the *hsp60* promoter in the *E. coli*–mycobacteria shuttle plasmid integrative vector pMV361. Complemented strains were prepared by introducing constructs into the deletion mutants by electroporation using standard protocols (3). Following electroporation, bacteria were inoculated into Middlebrook 7H9 medium and incubated overnight at 37°C before plating on 7H10 medium containing 25 μg/ml kanamycin and, for the *esx-3* region mutants, 200 ng/ml mycobactin J. The presence of the constructs in the complemented strains was verified by PCR.

Whole-genome sequencing (WGS). WGS of purified genomic DNA was conducted using Illumina and Nanopore MinION as described below.

Illumina WGS. Genomic DNA was extracted from H37Rv wild-type (WT) and mutant strains by the hexadecyltrimethylammonium bromide-lysozyme method using established protocols (3). Genomic libraries were prepared for sequencing using the Nextera XT kit per the manufacturer's instructions (Illumina, Inc., San Diego, CA). In various experiments, the DNA libraries were sequenced at the Wellcome Trust Sanger Institute (Cambridge, UK), at Albert Einstein College of Medicine (Bronx, NY) and at Georgia State University (Atlanta, GA) on the MiSeq platform using Illumina V2 or V3 chemistry according to manufacturer's protocols.

The WGS raw data were imported into CLC Genomics Workbench 10.0 (Qiagen, Hilden, Germany) and trimmed to remove low quality reads using default parameters. The trimmed sequencing data were mapped against the *M. tuberculosis* H37Rv reference genome (NCBI accession number NC000962.3), followed by local realignment with the default settings. More than 99% of Illumina raw reads (range 1.4 to 2.7 million) with an average mapped read length of ~150 bp mapped to the reference genome. Genome-wide single nucleotide polymorphism (SNP) variants were identified by comparison with the reference genome using the variant detection function of CLC Genomics Workbench passing variants with frequency higher than 20%, average base quality score over 20.0 and forward to reverse reads ratio over 0.05, and then the identified variants were compared between the H37Rv parental strain run in parallel and the various suppressor mutants with respect to the *M. tuberculosis* H37Rv reference genome. Because initial read mapping to the H37Rv reference genome revealed variants in our H37Rv WT parental strain, for each strain under analysis, the variant filtering tool was used to retain only variants with no exact match found in the track of H37Rv WT or in the parental mutant strain (when comparing, for example, Δ*esxG* recovered from an *in vivo* source to the Δ*esxG* parent) (**Dataset S1**).

Nanopore MinION (Oxford Nanopore Technologies, ONT) sequencing, assembly, and analysis. Approximately 1 μg of genomic DNA (gDNA) from strains used in this study were independently subjected to nanopore sequencing on a MinION sequencer. The Ligation Sequencing Kit (Oxford

Nanopore Technologies) was used for library preparation according to the manufacturer's protocol. Briefly, gDNA samples were independently treated with UltraII End repair enzymes in a 60 μl final volume reaction and incubated at 20°C on a thermal cycler for 5 minutes followed by an additional 5 minutes at 65°C to terminate the reaction. Treated gDNA was purified with AMPure XP magnetic beads and eluted with nuclease-free water. AMX adapter ligation reactions were performed with NEBNext Quick T4 DNA ligase (New England Biolabs) and supplied ONT ligation buffer in a 100 μl final volume reaction at room temperature for 10 minutes. AMX-tagged gDNA was incubated with AMPure XP magnetic beads, then washed with long fragment buffer supplied with the sequencing kit. Magnetic gDNA library complexes were eluted in 15 μl of elution buffer and eventually mixed with sequencing buffer and loading beads. The libraries were loaded onto FLO-MIN106D R9 flow cells for sequencing on a MinION sequencer (101B) with a 48 h run time.

For data processing, Guppy (v 3.4.2) was used to convert raw Nanopore fast5 files to FASTQ format, and the fast-calling option was selected on the MinKnown software. Combined FASTQ data were filtered using Trimmomatic (v. 36), applying a filter to retain files larger than the 10 kb read length for the Δ*esxGH*::361-EV M20 strain, and files larger than 50 kb read length for the Δ*esxGH*::361-EV M18 strain. Filtered files were submitted to the Albert Einstein College of Medicine Computing Cluster for chromosome assembly using Canu (v1.8). Standard Canu assembly parameters were used for Δ*esxGH*::361-EV M20, whereas for Δ*esxGH*::361-EV M18 (**Table S1**), the default parameters for CorMhapSensitivity and utgReAlign were changed to low and false, respectively, to allow the algorithm to capture the duplicated regions. Canu assembly resulted in a single contig that was predicted to be a circular chromosome for Δ*esxGH*::361-EV M20. The Δ*esxGH*::361-EV M18 assembly generated two contigs, which were over 3.94 Mb (tig001) and 661 kb (tig236); the two contigs had overlapping regions that were stitched together on BioMatters Geneious R11 software.

For contig polishing and annotation, Illumina reads from *Mtb* mutant strains were independently mapped to their respective corresponding Canu-generated contigs using BioMatters Geneious software (vR11) with standard parameters. The sequence accuracy of the MinION long-read contigs was improved by using Illumina-derived sequence data to "polish" the final genome assembly to 99.7% pairwise identity when compared to the reference assembly. The consensus polished contigs were extracted and annotated using the H37Rv reference genome. Annotations were performed by first annotating regions with perfect similarity, and incrementally decreasing the annotation similarity to 95 percent. It was found that Bowtie2 plugin on the Geneious software performed best in terms of distributing Illumina reads across the contigs, particularly on the Δ*esxGH*::361-EV M18 chromosome. Polished consensus contigs were then stitched together by removal of overlapping excess regions at both ends of the chromosome generated by the Canu assembler. The Δ*esxGH*::361-EV M20 chromosome was manually circularized on Geneious software by merging complementary regions at the ends of the contigs. However, the Δ*esxGH*::361-EV M18 chromosome remained linear due to the tandem duplicated regions at the ends of the chromosome.

Protein expression and purification. Rv3058c WT, R31H, and H22D/R31H/L33P mutants were expressed as maltose binding protein (MBP) fusion proteins in BL21(DE3) *E. coli* cells (Novagen). Recombinant protein expression was induced at OD_{600} of 0.6-0.7 with 0.5 mM IPTG for 12-14 h at 18˚C. Cells were harvested and resuspended in lysis buffer containing 20 mM Tris pH 7.5, 1 M NaCl, 20 mM imidazole, and 5 mM 2-mercaptoethanol. Cells were lysed using an EmulsiFlex-C5 homogenizer (Avestin) and clarified by centrifugation at 30,000 *g* at 4 ˚C for 40 min. Proteins were purified using a series of affinity and ion exchange chromatographic columns. The MBP tag was cleaved using TEV protease prior to final purification on a size exclusion column. Purity of proteins were determined by Coomassie staining of SDS-PAGE.

Secreted alkaline phosphatase (SEAP) reporter assay.

Construction of plasmids. Plasmids used in this study are listed in **Table S3**. Chimeric transcriptional regulator-expressing plasmids were constructed by fusing the coding sequence of H37Rv *rv3058c* (Genbank:NC_000962.3 nucleotides 3418726-3419376) or the point mutant R31H to the 78 amino acid transcription-activating domain of the herpes simplex virus type 1-derived VP16 protein followed by a carboxy-terminal hemagglutinin (HA) epitope tag. The resulting ORFs were codon-optimized for expression in human cells and cloned into pUC57 to generate pUC-3058c-VP16-HA and pUC-3058c-R31HVP16-HA (GenScript, Piscataway, NJ). Rv3058c-VP16-HA and R31H-VP16-HA were amplified by PCR from these vectors using Rv3058c forward and Rv3058c reverse primers to introduce EcoRI and XbaI sites at the 5' and 3' ends, respectively. PCR products were digested with EcoRI and XbaI and ligated into the corresponding sites of pSEAP2-control (Clontech, Takara Bio USA) to generate the expression vectors p3058c-VP16 and p3058c-R31H-VP16, in which the inserts replaced the SEAP coding sequences.

To test the DNA-binding capacity of the *Rv3058c*-VP16 chimeras using SEAP activity as a readout, ~200 nucleotide promoter regions from *rv0834c* and *pe29* were inserted upstream of a minimal variant of the human cytomegalovirus promoter ($P_{hCMVmin}$) and cloned into the pUC57 EcoRV site, generating vectors pUC-up-pe29 and pUC-up-0834c. The synthetic promoters were amplified from the constructs using the indicated primers (**Dataset S2**) to introduce XhoI and EcoRI restriction sites. The resulting PCR products were purified and cloned into pSEAP2-basic to generate pUP-pe29-pSEAP2 (pO_{pe29}-$P_{hCMVmin}$) and pUP-0834c-pSEAP2 ($pO_{Rv0834c}$-$P_{hCMVmin}$), which served as the SEAP reporter plasmids. All inserts were confirmed by sequencing (GenScript).

Cell culture. Human embryonic kidney (HEK) 293T cells (ATCC, CRL-3216) were maintained in Dulbecco's modified Eagle's medium (DMEM, Corning), supplemented with 10% fetal bovine serum (FBS) and 1% (v/v) penicillin/streptomycin and cultured at 37°C and 5% CO_2.

Transfections and SEAP assays. HEK293T cells were seeded at $1x10^5$ cells/well into Microlite luminescence microtiter 96-well plates (VWR, Radnor, PA). Cells were transfected using Lipofectamine 2000 (Thermo Fisher Scientific, MA) with 50 ng of a constitutively expressed *Renilla* luciferase reporter plasmid (pRL-tk) (Promega, Madison, WI) and 10 ng of indicated constructs. For control wells transfected with only a transcription factor-expressing construct or a SEAP reporter construct, 10 ng of pSEAP-Basic was included to adjust total amounts to 70 ng plasmid per transfection. The assay was performed in six replicates. At 48 h post-transfection, 50 μl supernatant of three replicates was transferred to tubes and centrifuged in a microcentrifuge at 4,000 rpm for 5 min; 25 μl of the resulting supernatant was transferred to a new luminescence 96-well plate, and SEAP activity was quantified with the Great EscAPe SEAP Chemiluminescence Kit 2.0 (Clontech). The corresponding cell monolayers were lysed in passive lysis buffer (Promega), and *Renilla* luciferase reporter assays (Promega) were performed, with SEAP values normalized to *Renilla* luciferase values. For the remaining three replicates, the cells were lysed with NP-40 lysis buffer (50 mM Tris [pH 7.5], 280 mM NaCl, 0.5% IGEPAL CA-630, 0.2 mM EDTA, 2 mM EGTA, 10% glycerol, protease inhibitor [cOmplete; Roche, Indianapolis, IN]); the lysates were combined and assayed by Western Blot to assess expression of the HA-tagged wild-type and mutant Rv3058c proteins.

EMSA. Electrophoretic mobility shift assays (EMSA) were performed to evaluate the DNA binding capacity of Rv3058c WT. DNA probes containing the ~200 bp sequences upstream of *rv0834c* and *pe29* were generated by PCR amplification using DNA template fragments synthesized by Integrated DNA Technologies as gBlocks® Gene Fragments, using primers cy5 M13F and cy5 M13R (**Dataset S2**). The EMSA reactions were prepared using the Invitrogen Molecular Probes

SYBR™ SYPRO™ Electrophoretic Mobility-Shift Assay (EMSA) Kit (Thermo Fisher Scientific, Waltham, MA) according to the manufacturer's instructions and contained 3 μl (30 ng/μl) DNA probe, 1 μl (1 ug/μl) Rv3058c recombinant purified protein, and 2 μl 5x binding buffer (750 mM KCl, 0.5 mM dithiothreitol, 0.5 mM EDTA, 50 mM Tris, pH 7.4). The binding reactions also included 3 μl (30 ng/μl) of non-specific DNA (**Dataset S2**) with dH_2O added to make 10 μl total reaction volume. Following incubation, the reaction mixtures were electrophoresed on 6% native polyacrylamide gels. Images of Cy5-labeled nucleic acids were captured using a Bio-Rad ChemiDoc Imager. After imaging the Cy5 signal, the gel was washed 2 x 5 minutes with 150 ml dH_2O and stained with Imperial™ Protein Stain (Thermo Fisher Scientific) for 2 h at room temperature on an orbital shaker. The gel was then washed 3 x 10 minutes in dH_2O at room temperature, and images were acquired with a Bio-Rad ChemiDoc Imager.

qPCR of *esxR* and *esxS* copy numbers in amplification strain. *EsxR* and *esxS* gene copy numbers in selected strains were determined using Evagreen® Dye and a qPCR Master Mix kit (Biotium, Fremont, CA) to perform a quantitative real-time PCR assay. Primers were designed to amplify an ~100 bp region of *esxR* and an ~120 bp region of *esxS*. The optimized final concentration of all primers was 0.5 μM. The ~5027 bp plasmid pMV361K-esxRS, containing the *esxS-esxR* coding sequences, was used as template to test primer efficiencies and to establish standard curves for determining copy numbers. Genomic DNA extracted from H37Rv WT and from the indicated Δ*esxGH* isolates before and after passage through mice was diluted to 50 ng/μl before performing the qPCR assay. Three microliters of each template and the plasmid standards were run in parallel on a CFX96 Touch Real-Time PCR Detection System (Bio-Rad, Hercules, CA) with the following thermal cycling conditions: 95°C for 2 min for enzyme activation, followed by 40 cycles of denaturation at 95°C for 5 s, annealing at 60°C for 30 s, and extension at 72°C for 30 s. The quantities of *esxR* and *esxS* copy numbers in unknown samples were calculated using CFX Manager Software 3.1 (Bio-Rad, Hercules, CA).

Southern blot of *esxRS* amplification strain. Genomic DNA extracted from the indicated strains was digested with BamHI, separated on a 0.7% agarose gel, transferred to Hybond N+ membrane (GE Healthcare Health Sciences) using an alkaline buffer (4), and Southern blot analysis was performed using standard procedures. The probe was a 537 bp portion of *esxS-esxR*, amplified using the indicated primers (**Dataset S2**) and labeled using the Biotin DecaLabel DNA Labeling kit (Thermo Fisher Scientific).

qRT-PCR analysis of *esxS* expression. All strains were grown to log phase in complete 7H9 medium supplemented with 100 μM hemin, washed three times with PBS+0.05% tyloxapol, inoculated into 7H9 medium lacking hemin at starting OD_{600} in the range of 0.015-0.045, and allowed to grow for 3 days until OD_{600} reached values between 0.4-0.5 for all strains. Samples from three biological replicates were harvested by centrifugation; cell pellets were resuspended in 1 ml Trizol, snap-frozen in a dry ice-ethanol bath, and stored at -80°C until processing using a Fast-Prep apparatus (MP Biomedicals). After removal of debris by brief centrifugation, RNA was prepared from the supernatant using the Direct-zol RNA MiniPrep Kit (Zymo Research) per manufacturer's instructions, including an on-column DNase I treatment. RNA was subjected to an additional DNase treatment with ezDNase (ThermoFisher, Waltham, MA, USA) and first strand synthesis carried out with SuperScript IV (ThermoFisher, Waltham, MA, USA) using the included random hexamer primers. qPCR was performed using PerfeCTa SYBR Green FastMix (VWR, Radnor, PA, USA) and primers for *Mtb* H37Rv *esxS*, and *sigA* as an internal control (see **Dataset S2**). Each assay was performed in triplicate with three technical replicates, and each assay contained RNA samples not subjected to reverse transcription as "no RT" controls. Data were analyzed by the $2^{-\Delta\Delta C_T}$ method with *sigA* serving as the housekeeping gene (5).

Transcriptome analysis. For RNA extraction, all strains were grown to log phase in complete 7H9 medium supplemented with 100 μM hemin, washed three times with PBS+0.05% tyloxapol, and inoculated into 7H9 medium lacking hemin at the desired OD_{600} values (0.01 for H37Rv and

H37Rv Δ*rv3058c*, 0.03 for Δ*esxG* Δ*rv3058c*, and 0.04 for Δ*esxG*). Next, 10 ml samples from three biological replicates were harvested on days 3 and 5, centrifuged, and then cell pellets were resuspended in 1 ml Trizol, snap-frozen in a dry ice-ethanol bath, and stored at -80°C until processing using a Fast-Prep apparatus (MP Biomedicals). After removal of debris by brief centrifugation, RNA was prepared from the supernatant using the Direct-zol RNA MiniPrep Kit (Zymo Research) per manufacturer's instructions, including an on-column DNase I treatment.

RNA-seq and bioinformatic analyses. RNA integrity was checked on a Bioanalyzer after which ribosomal RNA was removed using an NEB bacterial depletion kit. Libraries were prepared using an Illumina Stranded Total RNA Prep kit according to manufacturer's instructions. The RNA samples were sequenced at the Institute for Genome Sciences (IGS, Baltimore, MD) on the Illumina HiSeq sequencing platform using 75 bp paired-end reads. Raw sequencing reads generated for each sample were analyzed using the CAVERN transcriptomics analysis pipeline (6). Read quality was assessed using the FastQC toolkit (7). Reads were first aligned to the *Mtb* reference genome GCF_000195955.2_ASM19595v2 (available from NCBI RefSeq database) using Bowtie2 (8). Default parameters and the strand specific protocol were used to generate the alignment BAM files. Read alignments were assessed to compute gene expression counts for each gene using the HTSeq count tool (9) and the *Mtb* reference annotation (H37Rv ASM19595v2 37). The raw read counts were normalized for library size and utilized to assess differential gene expression between the wildtype and deletion mutant groups using the R package 'DESeq' (10). Normalized gene expression values were utilized for exploratory data analyses to assess reproducibility across replicates and identify potential outlier samples. Principal component analysis (PCA), commonly used for exploratory data analysis of multidimensional data, was implemented using R. The matrix of normalized gene expression values was used as input for PCA to generate a new set of orthogonal variables known as 'principal components'. These components were used to visualize variation in the dataset and assess sample-to-sample

distance. P-values were generated using a modified Fisher's exact test implemented in DESeq and then corrected for multiple hypothesis testing using the Benjamini-Hochberg correction method. Genes that were significantly differentially expressed between conditions were identified using a false discovery rate (FDR) of 5% and a minimum fold-change of 2. The sets of differentially expressed genes (DEGs) for individual comparisons were assessed for functional enrichment based on the knowledgebase of functional gene sets available from the KEGG database (11).

Mouse infections.

Female C57BL/6 mice of 6 weeks of age were obtained from the National Cancer Institute (data in **Fig. 1A**) or aged 5-6 weeks from Envigo (data in **Fig. 4**). Female CBA mice, aged 4-6 weeks, were obtained from the Jackson Laboratory. Female severe combined immunodeficiency (SCID) mice (C.B-17/IcrHsdPrkdc-scid), aged 4-6 weeks, were obtained from Harlan Sprague Dawley, Inc.

For aerosol infections, mycobacteria were grown in 7H9 medium with 100 μM hemin and antibiotics as appropriate. Bacteria were washed in PBS containing 0.05% Tween 80 (PBS-T) and sonicated, and then the bacterial suspension was supplemented with 0.04% antifoam Y-30 emulsion (Sigma) and delivered using a Madison chamber aerosol generation device calibrated to deliver ~100 colony-forming units (CFU) (12). Mice from each infection group were euthanized at day 1 post-aerosol exposure, and lung homogenates were plated onto 7H10 with 200 ng/ml mycobactin J to determine initial bacterial numbers delivered per strain. Bacterial burdens in lungs and spleen were determined by plating organ homogenates, prepared and diluted in PBS-T, from three to four mice per infection group onto 7H10 with 200 ng/ml mycobactin J.

For intravenous infections, mycobacterial strains were grown in 7H9 medium (data in **Fig. 1C**) or in 7H9 supplemented with 100 μM hemin and antibiotics as appropriate (data in **Fig. 3**). Bacteria were washed in PBS-T and sonicated, and mice were injected via the lateral tail vein with the indicated doses of bacteria in 200 μl PBS-T (target inoculum 2 x 10^6 CFU for experiment shown

in **Fig. 1C** and 1 x 10^5 CFU for experiment shown in **Fig. 3**). The actual doses administered were determined by serial dilution and plating onto 7H10 medium with 200 ng/ml mycobactin J. For the data shown in **Fig. 1C**, inocula were 3.1 x 10^6 for H37Rv WT and 2.7 x 10^6 for Δ*esxH*. For the data presented in **Fig. 3**, inocula were: 5.7 x 10^4, H37Rv WT; 1.27 x 10^5, H37Rv Δ*rv3058c*; 6.0 x 10^4, Δ*esxG;* 6.8 x 10^4, Δ*esxG* Δ*rv3058c*; 6.6 x 10^4, Δ*esxG* Rv3058c R31H; 5.04 x 10^4, Δ*esxG* Rv3058c H22D; 7.0 x 10^4, Δ*esxH*; 1.26 x 10^5, Δ*esxH* Δ*rv3058c*; 1.53 x 10^5, Δ*esxH* Rv3058c L33P; 6.0 x 10^4, Δ*esxH*::361-*esxGH*; and 4.9 x 10^4, Δ*esxH*::361-*esxRS*.

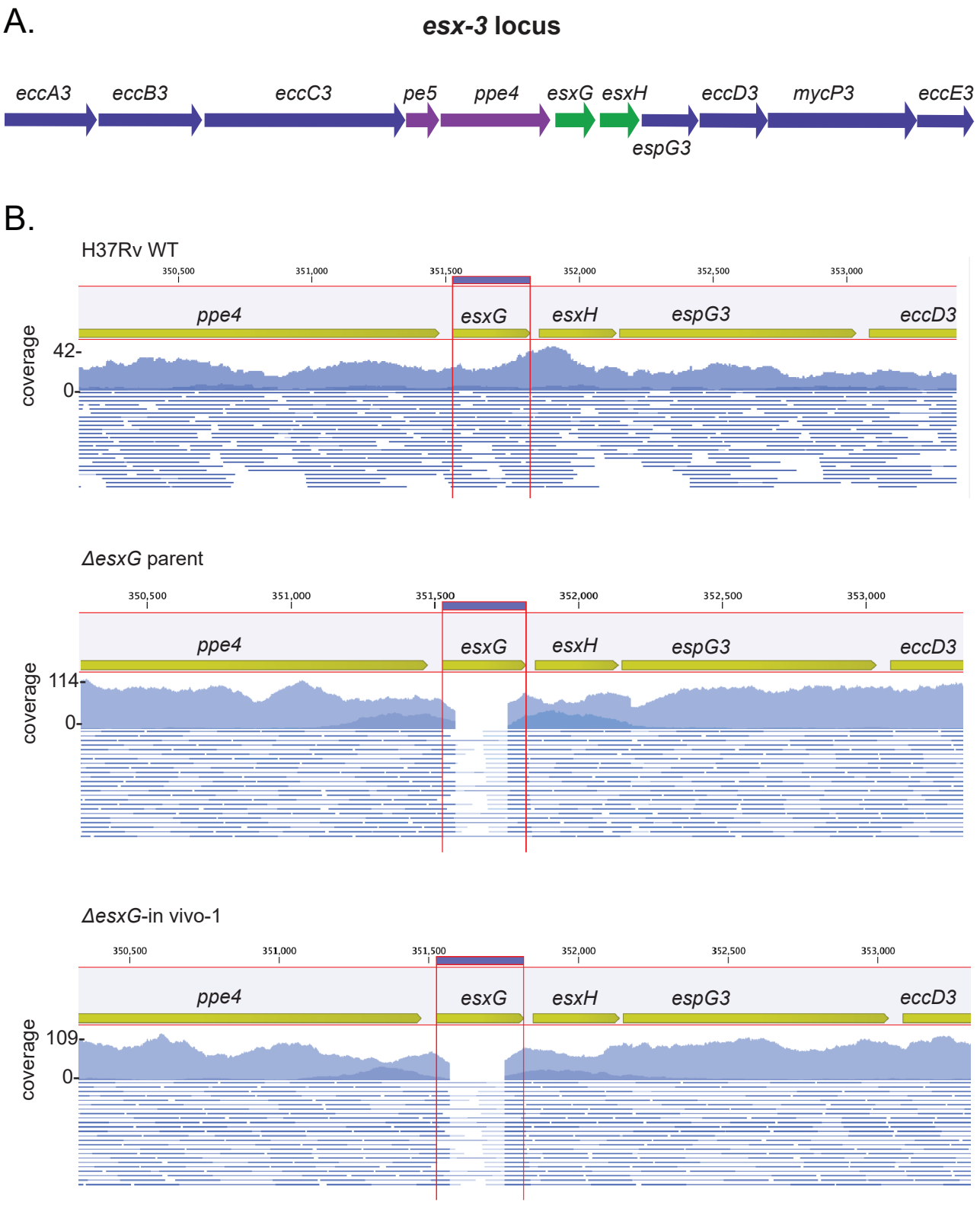


Figure S1. Confirmation of *esxG* deletion in isolate with suppressor phenotype. **A.** Schematic illustrating genomic organization at *M. tuberculosis* H37Rv *esx-3* locus (not to scale). The *pe-ppe* gene pair is indicated by purple arrows and the *esx* gene pair by green arrows. B. Read coverage maps from Illumina WGS of the H37Rv WT strain (top panel), the parental Δ*esxG* mutant strain (middle panel) and the in vivo isolate Δ*esxG*-in vivo-1 (bottom panel). Numbers above the coverage maps indicate genome position based on the standard NBCI annotation. The numbers on the y-axis indicate fold coverage.

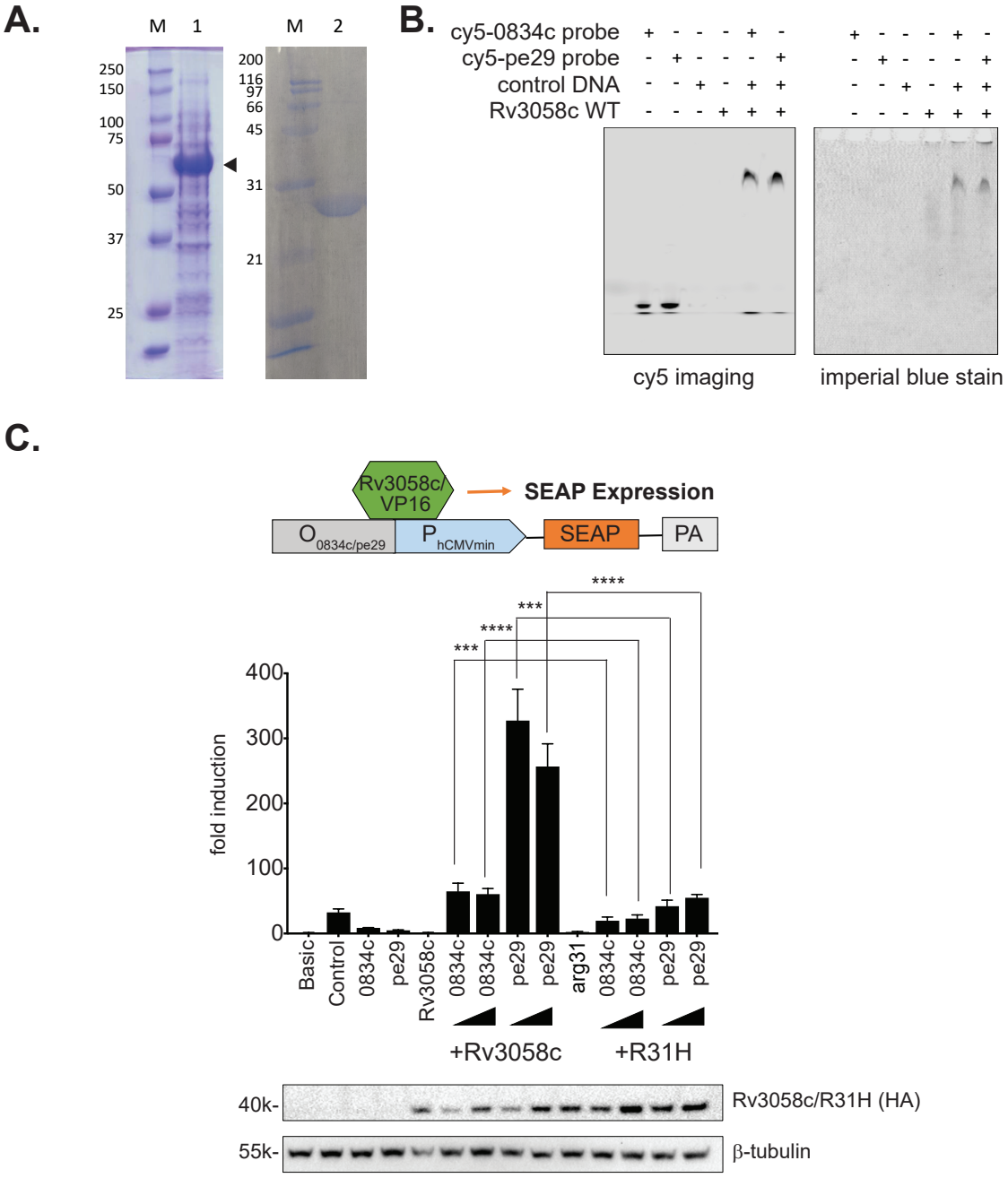


Figure S2. Interaction of rRv3058c repressor with target DNA sequences by electrophoretic mobility shift and reporter assays. A. Coomassie-stained SDS-PAGE of 1. Recombinant expression of MBP-tagged Rv3058c WT (arrowhead) and 2. Purified Rv3058c WT after cleavage of the MBP tag. M, molecular weight marker. **B.** Electrophoretic mobility shift assay (EMSA) showing binding of purified Rv3058c WT protein to target DNA sequences corresponding to the 200 bp upstream sequences of *rv0834c* (cy5-0834c probe) or of *pe29* (cy5-pe29 probe) in the presence of non-specific DNA sequence (control DNA). Rv3058c WT protein shifted both the 0834c and pe29 probes from the location of free probe near the bottom of the gel to slower migrating bands, as assessed by cy5 imaging. The association of Rv3058c WT protein with the probes was assessed by staining for the protein (imperial blue stain). **C.** Rv3058c WT and R31H coding sequences were fused to the HSV-1 VP16 transcriptional activating domain (Rv3058c/VP16) and assessed for their capacity to activate in HEK293T cells expression of a SEAP1 reporter gene. The SEAP1 gene was placed downstream of a minimal promoter (hCMVmin) preceded by 200 bp of sequence derived from the *rv0834c* or *pe29* presumptive promoter regions (O0834c/pe29). The protein expression plasmids were co-transfected with the SEAP reporter plasmids and a constitutively-expressing *Renilla* luciferase expression plasmid. As additional controls, the rv0834c and pe29 SEAP reporter plasmids were transfected in the absence of the Rv3058c WT or R31H plasmids. Also included was a positive control (Control) with SEAP1 expression driven by the SV40 early promoter and the SV40 enhancer. 48 hours post-transfection, SEAP and luciferase activities were measured and SEAP was normalized to *Renilla* luciferase activity and is presented as fold-induction relative to the signal from cells transfected with pSEAP-Basic (Basic), the parental plasmid lacking promoter and enhancer elements upstream of the SEAP gene. The reporter gene data represent the means and standard deviation (SD) of triplicates. Statistical significance was assessed between the indicated samples by using Student's t-test. ***P<0.01 and ****P<0.001. Samples transfected in parallel were used to assess expression of Rv3058c WT and R31H by western blot where anti-β-tubulin served as a loading control.

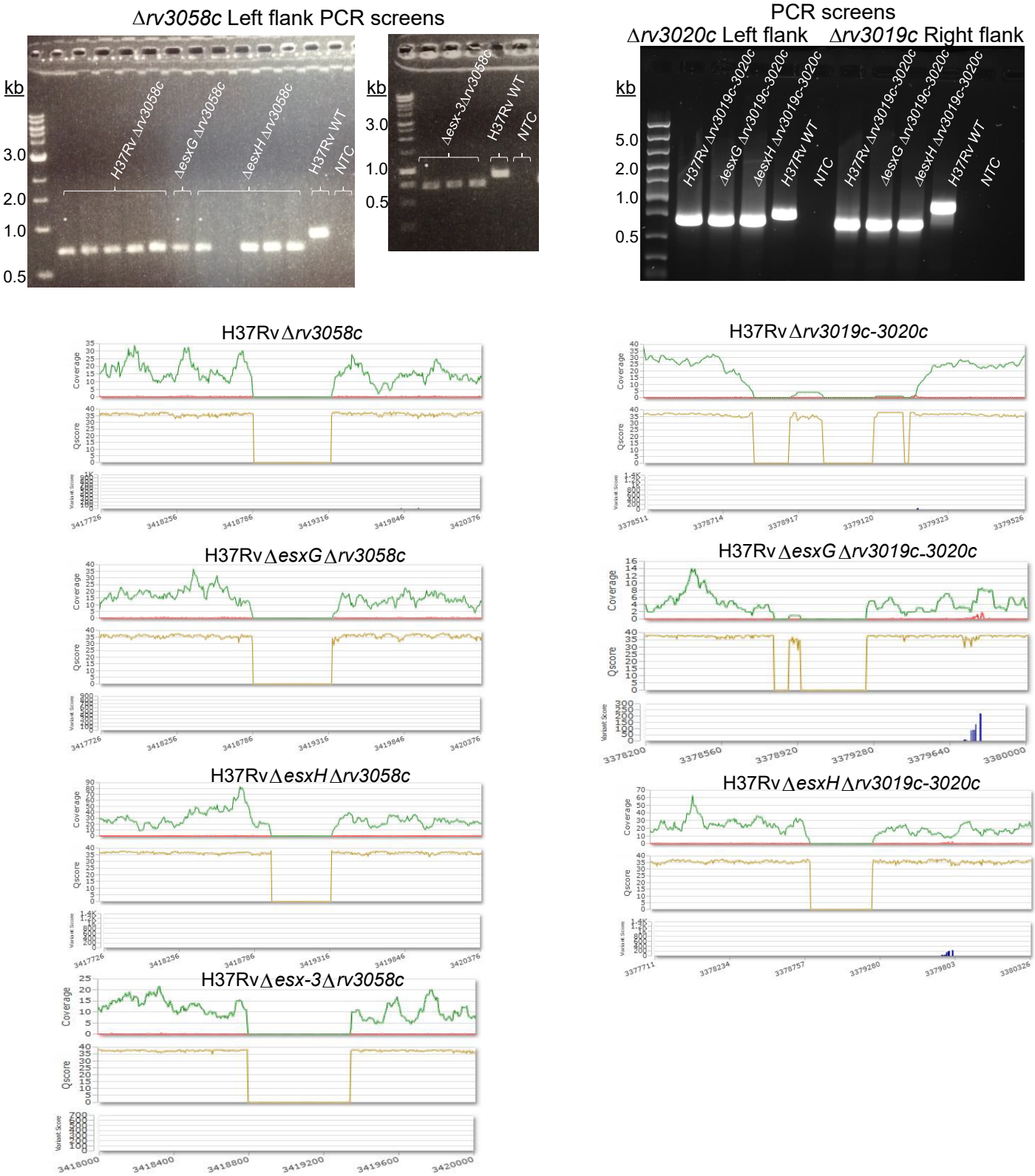
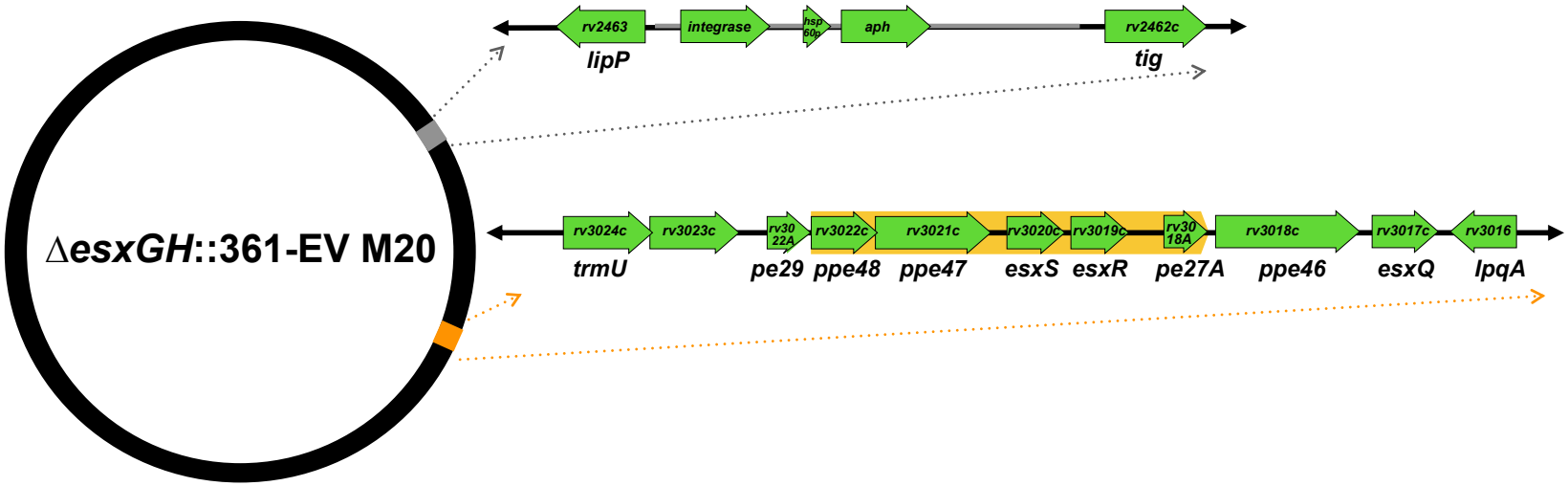


Figure S3. Confirmation of gene deletions by PCR and WGS. Upper panels show the results of PCR analyses at the *rv3058c* locus (two gel images on left) and the *esxS-esxR* locus (gel image on right) for the indicated clones. Note that for the *rv3058c* locus left flanks were screened, while for the *esxS-esxR* locus both the left and right flanks were screened. The sequences of primers used for the "3 primer" PCR screens are provided in Dataset S2. In each case, for the mutant clones, a smaller product is amplified as compared with the parental H37Rv WT strain. Where more than one transductant clone was screened for an individual mutant, an asterisk indicates the clone selected for use. Whole genome sequencing of gDNA from each strain was performed on a Miseq instrument. Lower panels show coverage maps obtained from the Miseq Reporter software, with genome position based on the standard NBCI annotation along the x-axis and fold coverage along the y-axis. The absence of reads in the targeted regions is evident for the deletion mutants. For uncertain reasons, the *rv3058c* deletion in the Δ*esxH* mutant background appears to be less extensive than in the other strains, although the majority of the open reading frame remains without mapped reads in this isolate confirming *rv3058c* deletion.

A. Nanopore sequencing: Δ*esxGH*::361-EV M20 stitched contigs



B. Illumina sequence reads mapped on Δ*esxGH*::361-EV M20

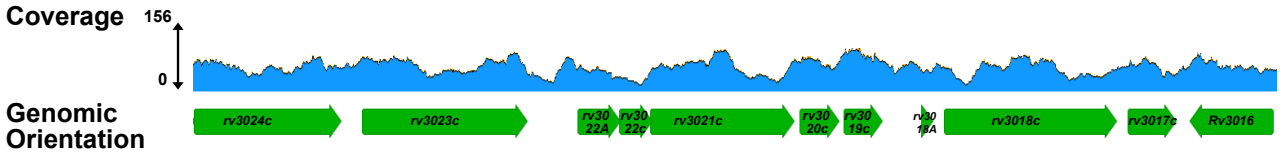


Figure S4. Nanopore sequencing of Δ*esxGH*::361-EV M20 reveals no evidence of *esxRS* locus amplification. A. Nanopore sequencing and Canu assembly of Δ*esxGH*::361-EV M20 resulted in a single circular contig of about 4.4 Mb which represents the complete genome, with no evidence of tandem repeats upon annotation. The single copy of *ppe48* (*rv3022c*)-*ppe47* (*rv3021c*)-*esxS* (*rv3020c*)-*esxR* (*rv3019c*)-*pe27A* (*rv3018A*) is highlighted in orange. A single copy of *integrase*-*hsp*$_{60}$ promoter-*aph* from the pMV361 vector inserted between *lipP* (*rv2463*) and *tig* (*rv2462c*) as revealed by Nanopore sequencing and annotation is shown in grey on the assembly. **B.** The Illumina reads are uniformly distributed across the genome when mapped to the Δ*esxGH*::361-EV M20 assembled chromosome.

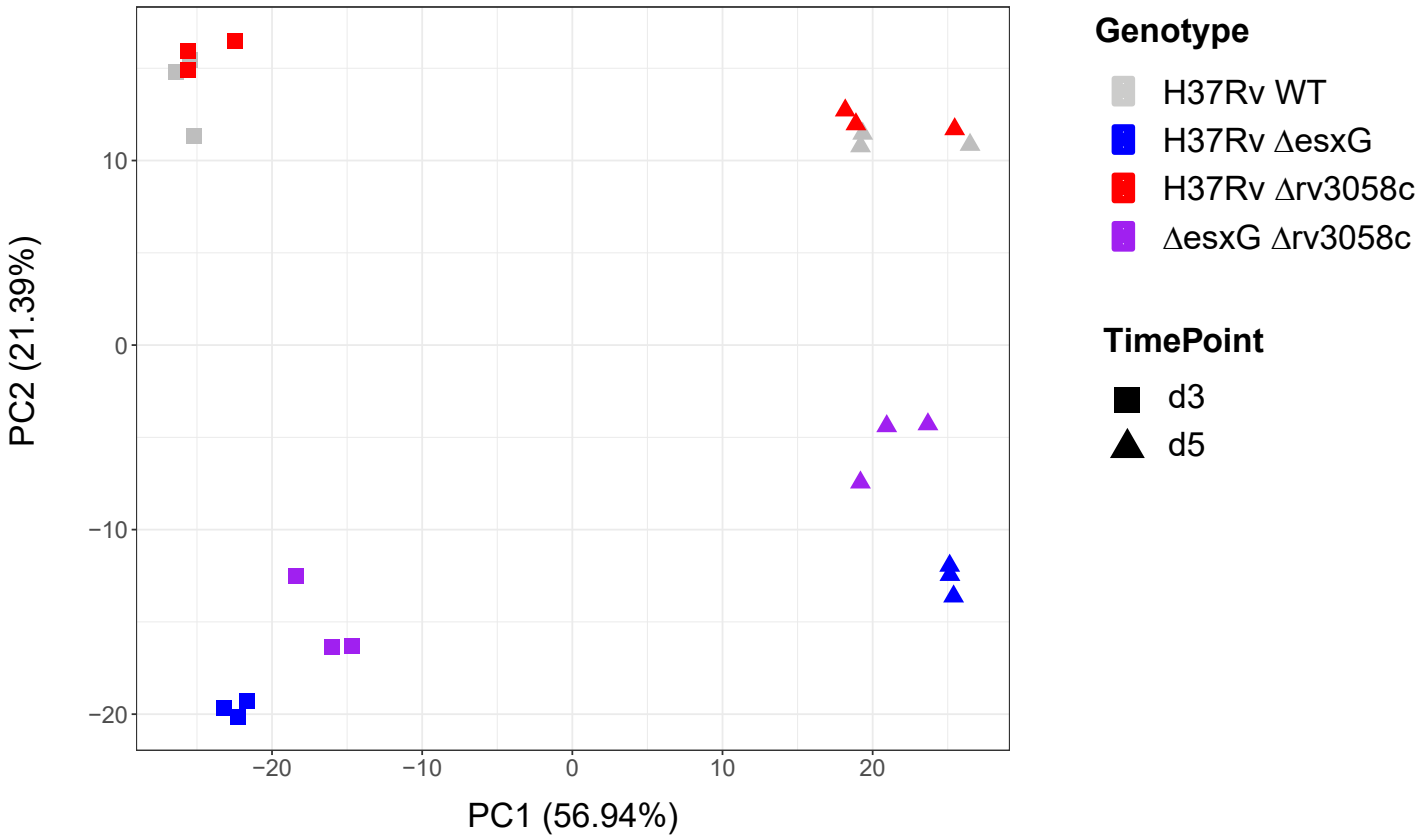


Figure S5. Principal component analysis of RNA-seq data obtained from strains growing in 7H9 medium. The two-dimensional principal component analysis (PCA) plot of RNA-seq samples based on their gene expression profiles illustrates that the first two components (PC1 and PC2) successfully separate out the time points and genotypes. The values indicated in parentheses are the percent variance explained by each principal component. Each point represents an RNA-Seq sample. Sample groups are color-coded by genotype, while the shapes indicate the time point of sample collection as indicated in the legend on the right.

Supplementary Tables

Table S1.

WGS of apparent suppressor mutants[1]

Isolate name	Isolate	Phenotype of original isolate	Phenotype upon regrowth and plating on unsupplemented 7H10	High frequency SNPs/findings on WGS
Δ*esxG*-in vivo-1	H37Rv Δ*esxG* 8 wk mouse #4 lung colony #1	High (~50K) lung cfu C57BL/6 aerosol	Growth	Rv3058c arg31his
Δ*esxG*-in vivo-2	H37Rv Δ*esxG* 16 wk mouse #3 lung colony #1	Very high (~150K) lung cfu C57BL/6 aerosol	Not tested	Rv3058c arg31his
Δ*esxG*-in vivo-3	H37Rv Δ*esxG* 16 wk mouse #3 lung colony #2	Very high (~150K) lung cfu C57BL/6 aerosol	Not tested	Rv3058c arg31his
Δ*esxG*-in vitro-1	H37Rv Δ*esxG* 7H11 + Zn large #1	Large colony isolated 7H11 supplemented with Zn but no additional iron complexes	Growth	Rv3058c arg31his
Δ*esxG*-in vitro-2	H37Rv Δ*esxG* 7H11 + Zn large #2	Large colony isolated 7H11 supplemented with Zn but no additional iron complexes	Not tested	Rv3058c arg31his
Δ*esxGH*::361-EV M18	H37Rv Δ*esxGH*::pMV361 empty vector mouse #18 colony #1	High (~30K) lung cfu C57BL/6 aerosol	Growth	Markedly increased coverage across *esxRS* locus

Mutations identified by PCR and targeted sequencing

Isolate name	Isolate	Organ burden after aerosol infection C57BL/6	Phenotype upon regrowth and plating on unsupplemented 7H10	SNPs on PCR and Sanger sequencing of Rv3058c
Δ*esxG*-in vivo-4	H37Rv Δ*esxG* 3 wk mouse #2 spleen colony #1	Lung ~500K Spleen ~0.5K	Growth	Rv3058c his22asp
Δ*esxG*-in vivo-5	H37Rv Δ*esxG* 3 wk mouse #2 spleen colony #2	Lung ~500K Spleen ~0.5K	Not tested	Rv3058c his22asp

Δ*esxG*-in vivo-6	H37Rv Δ*esxG* 8 wk mouse #4 spleen colony #1	Lung ~50K Spleen ~5K	Not tested	Rv3058c arg31his
Δ*esxG*-in vivo-7	H37Rv Δ*esxG* 8 wk mouse #4 spleen colony #2	Lung ~50K Spleen ~5K	Not tested	Rv3058c arg31his
Δ*esxG*-in vivo-8	H37Rv Δ*esxG* 8 wk mouse #4 lung colony #2	Lung ~50K Spleen ~5K	Not tested	Rv3058c arg31his
Δ*esxG*-in vivo-9	H37Rv Δ*esxG* 8 wk mouse #4 lung colony #3	Lung ~50K Spleen ~5K	Not tested	Rv3058c arg31his
Δ*esxG*-in vivo-10	H37Rv Δ*esxG* 16 wk mouse #3 lung colony #3	Lung ~150K Spleen ~100K	Not tested	Rv3058c Arg31his
Δ*esxG*-in vivo-11	H37Rv Δ*esxG* 16 wk mouse #3 lung colony #4	Lung ~150K Spleen ~100K	Not tested	Rv3058c arg31his
Δ*esxH*-in vivo-1	H37Rv Δ*esxH* early morbidity IV-infected CBA mouse lung colony #1	N/A	Growth	Rv3058c leu33pro
Δ*esxH*-in vivo-2	H37Rv Δ*esxH* early morbidity IV-infected CBA mouse lung colony #2	N/A	Not tested	Rv3058c leu33pro
Δ*esxH*-in vivo-3	H37Rv Δ*esxH* early morbidity IV-infected CBA mouse spleen colony #1	N/A	Not tested	Rv3058c leu33pro
Δ*esxH*-in vivo-4	H37Rv Δ*esxH* early morbidity IV-infected CBA mouse spleen colony #2	N/A	Not tested	Rv3058c leu33pro
Δ*esxH*-in vivo-5	H37Rv Δ*esxH* early morbidity IV-infected CBA mouse spleen colony #3	N/A	Not tested	Rv3058c leu33pro
Δ*esxH*-in vivo-6	H37Rv Δ*esxH* early morbidity IV-infected CBA mouse liver colony #1	N/A	Not tested	Rv3058c leu33pro
Δ*esxH*-in vivo-7	H37Rv Δ*esxH* early morbidity IV-infected CBA mouse liver colony #2	N/A	Not tested	Rv3058c leu33pro
Δ*esxH*-in vivo-8	H37Rv Δ*esxH* early morbidity IV-infected CBA mouse liver colony #3	N/A	Not tested	Rv3058c leu33pro

[1]Our analysis does not exclude the possibility that individual colonies isolated from the same mouse are siblings.

Table S2. Summary of the high frequency variant analysis from whole genome sequencing and targeted PCR and Sanger sequencing of *rv3058c*.

	SNP_1	SNP_2	SNP_3
Isolate containing SNP	Δ*esxG*-in vivo-1*, -2*, -3*, -6, -7, -8, -9, -10, -11 Δ*esxG*-in vitro-1*, -2*	Δ*esxG*-in vivo-4, -5	Δ*esxH*-in vivo-1, -2, -3, -4, -5, -6, -7, -8
How identified	WGS for isolates indicated with * PCR/Sanger sequencing of *rv3058c* for remainder	PCR/Sanger sequencing of *rv3058c*	PCR/Sanger sequencing of *rv3058c*
Gene/locus	*rv3058c*	*rv3058c*	*rv3058c*
Base change in SNP	92 G > A	64 C > G	98 T > C
Amino acid change	R31H	H22D	L33P

Table S3. Plamids and phages used in this study.

Plasmid	Relevant genotype/Description	Source
pYUB1471	Allelic exchange substrate (AES) cloning vector for specialized transduction	(2)
pYUB2981	pYUB1471 into which the flanking regions of *M. tuberculosis* H37Rv *rv3058c* were cloned.	This work
pYUB2982	pYUB1471 into which the flanking regions of *M. tuberculosis* H37Rv *esxS-esxR* were cloned.	This work
phAE159	Shuttle phasmid mycobacterial vector for specialized transduction	(2)
phAE1151	phAE159 with pYUB2981 cloned into the PacI site	This work
phAE1152	phAE159 with pYUB2982 cloned into the PacI site	This work
phAE280	phage to deliver $\gamma\delta$ resolvase for unmarking	(2)
pUC57K-GH	pUC57-kan::*esxG-esxH*; *M. tuberculosis* H37Rv *esxG* (*rv0287*)-*esxH* (*rv0288*) cloned into EcoRV site of pUC57-kan	This work (obtained from Genewiz)
pUC57K-RS	pUC57-kan::*esxS-esxR*; *M. tuberculosis* H37Rv *esxS* (r*v3020c*)-*esxR* (*rv3019c*) cloned into EcoRV site of pUC57-kan	This work (obtained from Genewiz)
pMV361.kan	integrative *E. coli* mycobacterial shuttle vector containing *hsp60* promoter	(13)
pMV361K-esxGH	pMV361.kan::*esxG-esxH* ; XmnI/HindIII fragment of pUC57K-GH cloned downstream of the *hsp60* promoter of pMV361	This work
pMV361K-esxRS	pMV361.kan::*esxS-esxR* ; XmnI/HindIII fragment of pUC57K-RS cloned downstream of the *hsp60* promoter of pMV361	This work
pUC-3058c-VP16-HA	pUC57-kan with codon-optimized *M. tuberculosis* H37Rv Rv3058c-Herpes simplex VP16-HA tag coding sequences cloned into EcoRV site	This work (obtained from GenScript)
pUC-3058c-R31HVP16-HA	pUC57-kan with codon-optimized *M. tuberculosis* H37Rv Rv3058c R31H-Herpes simplex VP16-HA tag coding sequences cloned into EcoRV site	This work (obtained from GenScript)
pSEAP2-control	Vector expressing secreted alkaline phosphatase (SEAP) under control of the SV40 early promoter and SV40 enhancer	Clontech
p3058c-VP16	Expression vector for SEAP assay; SEAP liberated by EcoRI/XbaI digestion	This work

	was replaced with codon-optimized Rv3058c-VP16-HA coding sequences, followed by the SV40 late polyadenylation signal	
p3058c-R31H-VP16	Expression vector for SEAP assay; SEAP liberated by EcoRI/XbaI digestion was replaced with codon-optimized Rv3058c R31H-VP16-HA coding sequences, followed by the SV40 late polyadenylation signal	This work
pUC-up-pe29	pUC57-kan with ~200 bp promoter region of *pe29* coding sequences cloned into EcoRV site	This work (obtained from GenScript)
pUC-up-0834c	pUC57-kan with ~200 bp promoter region of *rv0834c* coding sequences cloned into EcoRV site	This work (obtained from GenScript)
pSEAP2-basic	Vector lacking eukaryotic promoter and enhancer sequences; has MCS that allows putative promoter DNA fragments to be inserted upstream of the SEAP gene	Clontech
pUP-pe29-pSEAP2	pSEAP2-basic with ~200 bp promoter region of *pe29* cloned into MCS upstream of SEAP gene	This work
pUP-0834c-pSEAP2	pSEAP2-basic with ~200 bp promoter region of *rv0834c* cloned into MCS upstream of SEAP gene	This work

Table S4. *Mtb* strains used in this study.

Strain Name	Description	Construction	Source
H37Rv	wildtype	N/A	Trudeau Institute
mc^{2}7844	H37Rv Δ*esx-3 (rv0282 – rv0292*)	SpTr of H37Rv followed by unmarking with phAE180	(14)
mc^{2}7845	H37Rv Δ*esxG (rv0287)*	SpTr of H37Rv followed by unmarking with phAE180	(14)
mc^{2}7846	H37Rv Δ*esxH (rv0288)*	SpTr of H37Rv followed by unmarking with phAE180	(14)
mc^{2}7847	H37Rv Δ*esxG-esxH (rv0287 – rv0288)* (abbreviated Δ*esxGH*)	SpTr of H37Rv followed by unmarking with phAE180	(14)
mc^{2}7866	H37Rv Δ*esxG* (*rv0287*), attBL5::pYUB1944 (abbreviated Δ*esxG* compl.)	Transformation of mc^{2}7845 with pYUB1944 (integrating plasmid encoding *esxG-esxH* expressed from hsp60 promoter)	(14)
JTTB1	H37Rv Δ*rv3058c* $\gamma\delta$ *(sacB-hyg)* $\gamma\delta$	SpTr of H37Rv with phAE1151	This work
JTTB2	H37Rv Δ*esx-3* Δ*rv3058c* $\gamma\delta$ *(sacB-hyg)* $\gamma\delta$	SpTr of mc^{2}7844 with phAE1151	This work
JTTB3	H37Rv Δ*esxG* Δ*rv3058c* $\gamma\delta$ *(sacB-hyg)* $\gamma\delta$	SpTr of mc^{2}7845 with phAE1151	This work
JTTB4	H37Rv Δ*esxH* Δ*rv3058c* $\gamma\delta$ *(sacB-hyg)* $\gamma\delta$	SpTr of mc^{2}7846 with phAE1151	This work
JTTB5	H37Rv Δ*rv3019c-rv3020c* $\gamma\delta$ *(sacB-hyg)* $\gamma\delta$ (abbreviated Δ*esxRS*)	SpTr of H37Rv with phAE1152	This work
JTTB6	H37Rv Δ*esxG* Δ*rv3019c-rv3020c* $\gamma\delta$ *(sacB-hyg)* $\gamma\delta$ (abbreviated Δ*esxG* Δ*esxRS*)	SpTr of mc^{2}7845 with phAE1152	This work
JTTB7	H37Rv Δ*esxH* Δ*rv3019c-rv3020c* $\gamma\delta$ *(sacB-hyg)* $\gamma\delta$ (abbreviated Δ*esxH* Δ*esxRS*)	SpTr of mc^{2}7846 with phAE1152	This work
JTTB8	H37Rv Δ*rv3058c*	Unmarking of JTTB1 with phAE280	This work

JTTB9	H37Rv Δ*esxG* Δ*rv3058c*	Unmarking of JTTB3 with phAE280	This work
JTTB10	H37Rv Δ*esxH* Δ*rv3058c*	Unmarking of JTTB4 with phAE280	This work
JTTB11	H37Rv Δ*esxH*::pMV361-*esxG-esxH* (abbreviated Δ*esxH*::361-*esxGH*)	Transformation of mc^{2}7846 with pMV361K-esxGH	This work
JTTB12	H37Rv Δ*esxH*::pMV361-*esxS-esxR* (abbreviated Δ*esxH*::361-*esxRS*)	Transformation of mc^{2}7846 with pMV361K-esxRS	This work
JTTB13	H37Rv::pMV361EV (abbreviated H37Rv::361-EV)	Transformation of H37Rv with pMV361.kan empty vector	This work
JTTB14	H37Rv::pMV361-*esxG-esxH* (abbreviated H37Rv::361-*esxGH*)	Transformation of H37Rv with pMV361K-esxGH	This work
JTTB15	H37Rv::pMV361-*esxS-esxR* (abbreviated H37Rv::361-*esxRS*)	Transformation of H37Rv with pMV361K-esxRS	This work
JTTB16	H37Rv Δ*esxG-esxH*::pMV361 (abbreviated Δ*esxGH*::361-EV)	Transformation of mc^{2}7847 with pMV361.kan empty vector	This work
JTTB17	H37Rv Δ*esxG-esxH*::pMV361-*esxG-esxH* (abbreviated Δ*esxGH*::361-*esxGH*)	Transformation of mc^{2}7847 with pMV361K-esxGH	This work
JTTB18	H37Rv Δ*esxG-esxH*::pMV361-*esxS-esxR* (abbreviated Δ*esxGH*::361-*esxRS*)	Transformation of mc^{2}7847 with pMV361K-esxRS	This work
JTTB19	H37Rv Δ*esx-3*::pMV361 (abbreviated Δ*esx-3*::361-EV)	Transformation of mc^{2}7844 with pMV361.kan empty vector	This work
JTTB20	H37Rv Δ*esx-3*::pMV361-*esxG-esxH* (abbreviated Δ*esx-3*::361-*esxGH*)	Transformation of mc^{2}7844 with pMV361K-esxGH	This work
JTTB21	H37Rv Δ*esx-3*::pMV361-*esxS-esxR* (abbreviated Δ*esx-3*::361-*esxRS*)	Transformation of mc^{2}7844 with pMV361K-esxRS	This work

[a] SpTr, specialized transduction

Table S5. Genes differentially expressed in H37Rv Δ*rv3058c* vs. H37Rv WT at day 3 of growth in 7H9 medium, and expression ratios of these genes in Δ*esxG* Δ*rv3058c* vs. Δ*esxG* also at day 3.

Gene Identifier	Gene name	Expression ratio (Δ*rv3058c*/WT)	FDR	Expression ratio (Δ*esxG*Δ*rv3058c*/Δ*esxG*)	FDR	description
rv0834c	PE_PGRS14	3.21	1.88E-75	5.45	4.91E-135	PE-PGRS family protein PE_PGRS14
rv0989c	grcC2	4.04	4.37E-65	5.40	5.76E-109	Probable polyprenyl-diphosphate synthase GrcC2 (polyprenyl pyrophosphate synthetase)
rv3017c	esxQ	2.75	2.88E-20	2.04	3.09E-12	ESAT-6 like protein EsxQ (TB12.9)
rv3018c	PPE46	3.50	9.73E-24	2.60	8.15E-17	PPE family protein PPE46
rv3019c	esxR	3.03	4.27E-21	3.20	2.35E-29	Secreted ESAT-6 like protein EsxR (TB10.3)
rv3020c	esxS	2.63	7.17E-10	2.63	4.91E-12	ESAT-6 like protein EsxS
rv3021c	PPE47	2.28	8.99E-07	2.08	1.72E-07	PPE family protein PPE47

Legends for Datasets

Dataset S1, Whole genome sequencing variant analysis of Δ*esxG* and Δ*esxGH* isolates.

Results of variant analysis of Illumina Miseq data derived from the indicated strains using CLC Genomics Workbench as described in SI Materials and Methods.

Dataset S2, Oligonucleotides used in the study

The oligonucleotides used in this study include primers for AES construction and to verify deletions and unmarking (sheets 1-3); primers and oligonucleotides used in the SEAP assay and in the EMSA (sheet 4); qPCR and qRT-PCR primers (sheet 5); and primers to amplify Southern blot probe (sheet 6).

Dataset S3, RNA-seq data analysis.

Results of transcriptome analysis using RNA-seq as described in SI Materials and Methods. The index sheet provides a list of each comparison, including the strains and sample preparation time points. For each comparison, the unfiltered (unfilt) lists include all genes whose expression was detected, while the filtered (filt) lists include only those that met the criteria for differential expression as detailed in SI Materials and Methods.

SI References

1. Bardarov S*, et al.* (2002) Specialized transduction: an efficient method for generating marked and unmarked targeted gene disruptions in Mycobacterium tuberculosis, M. bovis BCG and M. smegmatis. *Microbiology* 148(Pt 10):3007-3017.
2. Jain P*, et al.* (2014) Specialized transduction designed for precise high-throughput unmarked deletions in Mycobacterium tuberculosis. *MBio* 5(3):e01245-01214.
3. Larsen MH, Biermann K, Tandberg S, Hsu T, & Jacobs WRJ (2007) Genetic Manipulation of Mycobacterium tuberculosis. *Current Protocols in Microbiology*, (John Wiley & Sons, Inc.), Vol 6, pp 10A.2.1-10A.2.21.
4. Brown T (1993) Southern blotting. *Current Protocols in Molecular Biology* 21(1):2.9.1-2.9.20.
5. Livak KJ & Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25(4):402-408.
6. Shetty AC*, et al.* (2019) CAVERN: Computational and visualization environment for RNA-seq analyses. *American Society of Human Genetics*.
7. Andrews S (2010) FastQC A Quality Control tool for High Throughput Sequence Data.
8. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357-359.
9. Anders S, Pyl PT, & Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166-169.
10. Anders S & Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
11. Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27-30.

12. Chen B*, et al.* (2011) Einstein Contained Aerosol Pulmonizer (ECAP): Improved Biosafety for Multi-Drug Resistant (MDR) and Extensively Drug Resistant (XDR) Mycobacterium tuberculosis Aerosol Infection Studies. *Appl Biosaf* 16(3):134-138.
13. Stover CK*, et al.* (1991) New use of BCG for recombinant vaccines. *Nature* 351(6326):456-460.
14. Tufariello JM*, et al.* (2016) Separable roles for Mycobacterium tuberculosis ESX-3 effectors in iron acquisition and virulence. *Proc Natl Acad Sci U S A* 113(3):E348-357.