Proteogenomic Analysis of *Mycobacterium tuberculosis* By High Resolution Mass Spectrometry*🅂

Dhanashree S. Kelkar‡§, Dhirendra Kumar¶, Praveen Kumar‡, Lavanya Balakrishnan‡‖, Babylakshmi Muthusamy‡, Amit Kumar Yadav¶, Priyanka Shrivastava¶, Arivusudar Marimuthu‡‡‡, Sridhar Anand§§, Hema Sundaram§§, Reena Kingsbury§§, H. C. Harsha‡, Bipin Nair§, T. S. Keshava Prasad‡**‡‡, Devendra Singh Chauhan¶¶, Kiran Katoch¶¶, Vishwa Mohan Katoch‖‖, Prahlad Kumar§§, Raghothama Chaerkady‡[a], Srinivasan Ramachandran¶, Debasis Dash¶, and Akhilesh Pandey‡[ab]**

The genome sequencing of H37Rv strain of *Mycobacterium tuberculosis* was completed in 1998 followed by the whole genome sequencing of a clinical isolate, CDC1551 in 2002. Since then, the genomic sequences of a number of other strains have become available making it one of the better studied pathogenic bacterial species at the genomic level. However, annotation of its genome remains challenging because of high GC content and dissimilarity to other model prokaryotes. To this end, we carried out an in-depth proteogenomic analysis of the *M. tuberculosis* H37Rv strain using Fourier transform mass spectrometry with high resolution at both MS and tandem MS levels. In all, we identified 3176 proteins from *Mycobacterium tuberculosis* representing ~80% of its total predicted gene count. In addition to protein database search, we carried out a genome database search, which led to identification of ~250 novel peptides. Based on these novel genome search-specific peptides, we discovered 41 novel protein coding genes in the H37Rv genome. Using peptide evidence and alternative gene prediction tools, we also corrected 79 gene models. Finally, mass spectrometric data from N terminus-derived peptides confirmed 727 existing annotations for translational start sites while correcting those for 33 proteins. We report creation of a high confidence set of protein coding regions in *Mycobacterium tuberculosis* genome obtained by high resolution tandem mass-spectrometry at both precursor and fragment detection steps for the first time. This proteogenomic approach should be generally applicable to other organisms whose genomes have already been sequenced for obtaining a more accurate catalogue of protein-coding genes. ***Molecular & Cellular Proteomics 10: 10.1074/mcp.M111.011627, 1–12, 2011.***

Mycobacterium *tuberculosis* continues to be a significant health burden, especially in the developing countries. Emergence of drug-resistant strains and a higher incidence of tuberculosis in people with HIV/AIDS have further worsened the situation. In the past, researchers have used proteomics for investigating the biology of this pathogen (1–7). There are a number of published studies pertaining to annotation of the *Mycobacterium tuberculosis* genome. The whole genome sequence of *M. tuberculosis* first became available for H37Rv strain in 1998, which was followed by that of CDC1551 and several other strains (8, 9). Accurate annotation of protein coding genes from any genome is a continuously evolving process. This is highly evident in the case of *M. tuberculosis*. Cole and colleagues reported the presence of 3924 open reading frames (ORFs) in H37Rv genome (8). In a re-annotation effort by the same authors, the gene number was revised to 3995 (10). As of March 2011, the TubercuList database contains 4012 annotated protein coding genes in the *M. tuberculosis* genome (11). de Souza *et al.* have carried out a comparison of two different gene annotation sets for *M. tuberculosis* H37Rv strain (Sanger and TIGR annotations) and reported that ~50% of the genes have different translation start sites (12). In the same study, using proteomic data for 449 culture filtrate proteins, the authors were able to correct

From the ‡Institute of Bioinformatics, International Technology Park, Bangalore, 560 066, India, §Amrita School of Biotechnology, Amrita University, Kollam 690 525, India, ¶Institute of Genomics and Integrative Biology (CSIR), Mall Road, Delhi 110 007, India, ‖Department of Biotechnology, Kuvempu University, Shimoga, 577 451, India, **Centre of Excellence in Bioinformatics, School of Life Sciences, Pondicherry University, Pondicherry 605014, India, ‡‡Manipal University, Madhav Nagar, Manipal, Karnataka 576104, India, §§National Tuberculosis Institute, Bellary Road, Bangalore 560 003, India, ¶¶National JALMA Institute for Leprosy and Other Mycobacterial Diseases, Tajganj, Agra, 282001, India, ‖‖Department of Health Research, Government of India and Director-General, Indian Council of Medical Research, V. Ramalingaswami Bhawan, Ansari Nagar, New Delhi 110029, India, [a]McKusick-Nathans Institute of Genetic Medicine and Departments of Biological Chemistry, Pathology and Oncology, Johns Hopkins University School of Medicine, Baltimore, MD 21205

Received May 24, 2011, and in revised form, September 28, 2011
Published, MCP Papers in Press, October 3, 2011, DOI 10.1074/mcp.M111.011627

annotations of 24 genes. Finally, the possibility of existence of many CDSs, which are not yet annotated in H37Rv genome, has also been suggested (13).

A direct evidence of translational potential of a genomic region can be obtained from peptide data from mass spectrometry-based proteomics (14–16). Other information such as N-terminal acetylation of peptides can be used for translational start site assignment. That the annotations in H37Rv genome are still not final is indicated by a recent analysis by de Souza *et al.*, where they used clustered database of annotated CDSs and flanking regions from five *M. tuberculosis* strains and three *M. bovis* strains to search mass spectrometric data to identify missing proteins from H37Rv genome. These investigators found peptide evidence for 24 genes incorrectly annotated in H37Rv genome (17). In the present study, we have carried out an in-depth proteomic analysis of *M. tuberculosis* using high resolution Fourier transform mass spectrometry. Cell lysates and culture filtrates were fractionated using various methodologies followed by tandem MS (MS/MS)[1] analysis on an LTQ-Orbitrap Velos ETD mass spectrometer. The mass spectrometry-derived data were analyzed using a six-frame translation of genome sequences in addition to searches of a protein database of *M. tuberculosis* H37Rv. We used two gene prediction programs (*FgeneSB* and *GeneMark*) to obtain alternative gene models (18, 19). In addition to gene predictions, we also used a comparative proteomic approach to validate alternative gene structures. From this analysis, 3176 proteins were identified representing ~80% of the total proteome of *M. tuberculosis*. A total of ~250 peptides that did not match existing annotations were identified. On the basis of these high confidence peptide identifications, we were able to delineate 41 novel genes in the H37Rv genome and correct 79 gene models. We were also able to identify alternative translational start sites for 33 proteins in addition to confirming translational start sites of 727 proteins using N terminus-derived peptides.

EXPERIMENTAL PROCEDURES

Culturing and Protein Extraction from Mycobacterium tuberculosis—M. tuberculosis H37Rv strain was grown in Middlebrook 7H9 media with OADC supplement. Colonies from Lowenstein-Jensen media slants were used to inoculate 1 liter of Middlebrook 7H9 media. Cultures were grown at 37 °C in a stationary condition for 5 weeks. At the end of 5 weeks, the cells were pelleted by spinning after washing 3 times using chilled phosphate buffer saline. Cell lysis was carried out by bead beating (0.1 mm zirconia beads) in presence of lysis buffer. Three percent SDS was used as lysis buffer for SDS-PAGE fractionation and 9 M urea was used as lysis buffer for in-solution digestion. For preparation of culture filtrate proteins, the *M. tuberculosis* H37Rv strain was grown in Proskauer-Beck media at 37 °C for 6 weeks. Proskauer-Beck media was chosen as it does not have added protein supplements. The cells were removed by filtering through 0.22-μm membrane filter. Filtrate was concentrated using 3 kDa cutoff filters (Millipore Corporation, Billerica, MA).

Trypsin Digestion and Fractionation—To obtain maximum proteome coverage, cell lysates were fractionated using three different methods. Protein level fractionation was carried out by SDS-PAGE. Two hundred micrograms of protein was loaded onto a 10% SDS-PAGE gel and stained using colloidal Coomassie blue stain. After removing excess stain, the lane was cut into 35 bands and subjected to in-gel tryptic digestion as described previously (20). In-gel reduction was carried out using 10 mM dithiothreitol followed by alkylation using 20 mM iodoacetamide. In-gel digestion was carried out using trypsin (Promega, Madison, WI) at 37 °C for 12 h. Peptides were extracted from the gel and dried using vacuum drying process as explained earlier (21).

Peptide level fractionation was carried out using strong cation exchange chromatography (SCX) and isoelectric focusing (offgel). In-solution trypsin digestion was carried out for ~1 mg of cell lysate protein following reduction and alkylation. Digestion was carried out at 37 °C for 12 h using trypsin (Promega) at the concentration of 1:20. Peptides were cleaned using C_{18} Sep-Pak columns. Half the amount of the peptides was used for SCX fractionation on polysulfoethyl A column (PolyLC, 200 × 2.1 mm; 5 μm, 200A) and the other half was used for fractionation by IEF using Agilent's 3100 OFFGEL fractionator. Twenty-four fractions were collected by offgel fractionation method over the pH range of 3 to 12. Two hundred micrograms of culture filtrate protein was fractionated by 10% SDS-PAGE and 36 bands were cut and in-gel digestion was done as described above for cell lysates.

LC-MS/MS Analysis—We carried out total of 123 (LC)-MS/MS runs (Cell lysate in-gel, 32; SCX fractions, 31; offgel fractions, 24; and culture filtrate ingel fractions 36). All of the mass spectrometry analyses were carried out on an LTQ-Orbitrap Velos ETD mass spectrometer (Thermo Scientific) interfaced with an Agilent 1200 series high-performance liquid chromatography system. The peptides from each fraction were analyzed using reversed phase nano scale liquid chromatography coupled to tandem mass spectrometry. The reversed phase nano scale liquid chromatography system consisted of a desalting column (75 μm × 2 cm, C_{18} material 5–10 μm, 120Å) and an analytical column (75 μm × 10 cm, C_{18} material 5 μm, 120Å) with an electrospray emitter tip 8 μm (New Objective Woburn, MA, USA) maintained at 2.0 kV ion spray voltage. The mass spectrometry analysis on the LTQ-Orbitrap Velos was carried out in a data-dependent manner with survey scan resolution $r = 60{,}000$ at m/z 400, scan range of m/z 350 to 1800. Following every survey scan, up to 15 most abundant precursor ions were picked for MS/MS fragmentation by collision induced dissociation (higher energy collision induced dissociation mode was used for analysis of SCX and offgel, collision induced dissociation mode was used for in-gel fractions). Fragment ions were detected in Orbitrap with resolution $r = 15{,}000$ at m/z 400. In the case of culture filtrate samples, MS/MS scans were acquired in the LTQ mass analyzer. Lock mass option was enabled, which helps maintaining high mass accuracy by real time calibration using polysiloxane ions from air. Further, ions picked for MS/MS were dynamically excluded for next 30 s. Normalized collision energy for MS/MS was set to 35%, and the transfer tube temperature was maintained at 220 °C.

Database Searches for Peptide and Protein Identification—Raw data files were processed to generate peak list files using Proteome Discoverer software version 1.2 (Thermo scientific). Filtering parameters used were: (1) Allowed precursor mass range was 500 Da to 5000 Da, (2) Precursor charge state was allowed from 1 to 5, (3) Minimum number of peaks in a spectra was chosen to be 5, (4) Signal to noise ratio was set as 3, (5) For precursors with unrecognized

[1] The abbreviations used are: MS/MS, tandem MS; PSM, Peptide spectrum match; FDR, False discovery rate; TSS, translational start site; GSSPs, Genome search specific peptides; ORF, open reading frame; SCX, strong cation exchange; CDS, Coding sequence.

charge state, default charge states of 2 and 3 were allowed. The protein database used for MS/MS searches was downloaded from NCBI for *M. tuberculosis* H37Rv strain (updated April 12, 2009). The genome sequence for H37Rv strain was downloaded from NCBI ftp site (NCBI Reference Sequence: NC_000962.2). Using in-house Python scripts, a six-frame translated database was created containing translated sequences from stop codon to stop codon. As *Mycobacterium* is known to use GTG and TTG as initiator methionine codons, wherever GTG and TTG codon was encountered it was translated as initiator methionine in addition to valine and leucine, respectively (8). The variant peptides thus obtained were appended to the genome translation databases as separate entries. Commonly encountered contaminants like BSA, trypsin, and keratins were added to the databases that were used for the MS/MS ion search. Total number of sequences in protein database, including contaminants, was 4015 whereas in genome translation database 320,958 sequences were present. Three different search algorithms, Mascot (version 2.2), Sequest (SCM build 59), and MassWiz (version 1.6.4.3-A) were used to analyze the data. Search parameters used were as follows: (a) Trypsin as a proteolytic enzyme allowing up to one missed cleavage, for MS/MS ion search against genome database, semitryptic cleavage was allowed; (b) Peptide mass error tolerance of 20 ppm; (c) Fragment mass error tolerance of 0.1 Da (for culture filtrate data 0.8 Da mass error was allowed as it was acquired in LTQ mass analyzer); (d) Fixed post-translational modifications was carbamidomethylation of cysteine residues. Variable modifications allowed were oxidation of methionine, acetylation of peptide N terminus and formylation of methionine. PSMs were filtered based on score threshold for 1% false discovery rate (FDR) established using decoy database search. A reverse sequence database was searched separately in addition to forward (target) database and FDR was calculated at every Peptide spectrum match (PSM) score value as

$$\% \text{ FDR} = (\text{Number of hits in reverse database at or above the score/Total number of hits in target and reverse database at or above the score}) \times 100$$

A 1% FDR threshold was applied to search results from individual data files. Because six different types of searches were performed with the same data set (genome and protein database searches using Mascot, Sequest, and MassWiz), peptide sequences assigned to a single spectrum in each search were compared. Spectra that were assigned different sequences in different searches were omitted from further analysis. The protein identification list was generated by grouping proteins based on shared peptides. Proteins that have at least one unique peptide were selected from the group. From a group in which no protein could be distinctly selected above others, that is, all proteins had equal evidence, it was represented by only one in the final list marked with an asterisk and remaining equivalent protein IDs were reported in parentheses. Proteins with no unique peptide evidence (subset proteins) were reported in a separate list. Peptides identified with PTMs which were used for protein N terminus analysis were manually evaluated for spectral assignments.

Analysis Using the MassWiz MS/MS Search Algorithm—MassWiz is a recently developed algorithm for peptide and protein identification from mass spectrometry data (22). MassWiz uses empirically determined abundance weights given to different ion series, their continuity, ion intensities, and supporting ions according to different instrument types. These weights help discriminate the good peptide spectrum matches from poor ones. MassWiz incorporates an integrated filtering module that picks high intensity peaks from dynamic mass bins determined from precursor neutral masses. This removes poor spectra prior to database search. This scoring scheme plugged with a simple spectral filtering approach enables MassWiz to maximize correct peptide identifications while lowering false identifications. MassWiz is open-source and is well suited for high resolution mass spectrometry data because the scoring function takes the mass accuracies of the matching fragment peaks into account.

Workflow for Genome Annotation—Peptide identifications from mass spectrometric data obtained in high resolution mode at both MS and MS/MS levels were used for proteogenomic analysis. This was done because we wanted to base our novel findings on high-confident peptide identification. Peptides obtained after applying the 1% FDR cutoff were selected for genome annotation analysis. Genome coordinates of all the peptides were found out using the tblastn program. Peptides mapping to multiple places in the H37Rv genome were not considered for further proteogenomic analysis. Genome search specific peptides were identified by excluding those peptides which mapped to known proteins from translated genome database search results. Genome search specific peptides were categorized as (1) mapping to intergenic region, (2) partially overlapping annotated genes, and (3) completely mapping to annotated genes. Alternative gene models were searched using two different gene prediction programs—*FgeneSB, GeneMark 2.5* for prokaryotes and *orfind* tool, for peptides that did not agree with the gene model that they were mapped to (categories 2 and 3) and those that mapped to intergenic region (category 1). Novel genes and gene model modifications thus obtained using peptide evidence and gene prediction tools were checked for their conservation across *Mycobacterium spp.* (in some cases *Corynebacterium* strains) using protein blast. MS/MS spectra used for proposing novel ORFs or changes in gene structure were manually validated from raw files for the sequence assignments.

RESULTS

LC-MS/MS Analysis of Intracellular and Culture Filtrate Proteins of M. tuberculosis H37Rv—Protein identification was carried out from both cell lysates and culture filtrates of *M. tuberculosis* H37Rv. Cell lysates were fractionated by three different fractionation methods: (1) SDS-PAGE followed by in-gel trypsin digestion, (2) Strong cation exchange chromatography, and (3) offgel electrophoresis (IEF), the latter two being at the peptide level. Culture filtrate proteins were fractionated by SDS-PAGE alone. In all, 123 LC-MS/MS runs were carried out. Approximately 1,800,000 MS/MS spectra were obtained, out of which 394,952 were assigned to peptide sequences using three different MS/MS search algorithms. PSMs were filtered for first rank assignments that passed a 1% FDR threshold. The total number of unique peptide sequences obtained was 35,562. The number of peptides identified exclusively from different fractionation methods was, 5788 from in-gel digestion, 5149 from strong cation exchange chromatography and 5291 from offgel fractionation. Figs. 1*A* and 1*B* show distribution of peptides and proteins identified from the various fractionation methods applied for analysis of cell lysate proteins. The complete list of

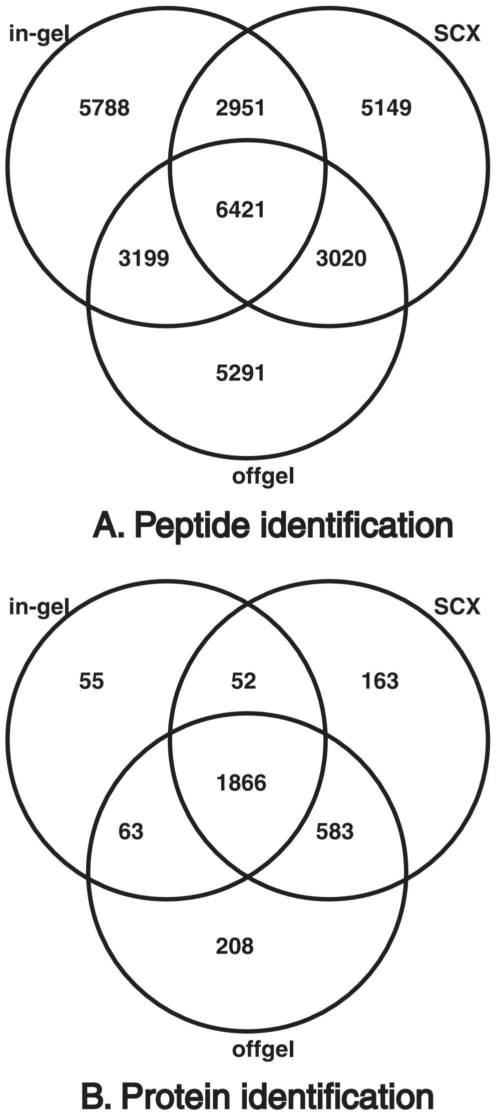


FIG. 1. **Identification of peptides and proteins by different fractionation methods.** *A*, Shows the distribution of peptides identified from *M. tuberculosis* cell lysates using three different fractionation methods—SDS-PAGE, Strong cation exchange chromatography (SCX), and offgel, *B*, Shows the distribution of proteins identified from cell lysates using the three fractionation methods.

peptides identified in our study along with PSM scores, charge, *m/z* value and post-translational modifications is provided in supplementary Data file 1. The complete set of raw mass spectrometry data (.raw files and peak list files) generated from this study has been made available through the Tranche server (http://proteomecommons.org/tranche, hash values for downloading the data are given at the end of the article). The raw data has also been deposited to PeptideAtlas (www.peptideatlas.org) where the raw data along with peptide identifications can be accessed using the accession number PAe001767 (23). The peptide and protein identification data have also been submitted to Open Source Drug Discovery portal (http://sysborg2.osdd.net/).

Confirmation of Annotated Protein Coding Genes in M. tuberculosis Genome—NCBI RefSeq database has 3988 protein sequences for H37Rv strain, whereas TubercuList database lists 4012 protein coding genes in the H37Rv genome (release R21, March 2010) (11). We have identified a total of 3176 mycobacterial proteins with at least one unique peptide identified, comprising ~80% of the total proteome of the *M. tuberculosis.* Proteins identified based on shared peptide evidence are listed separately (supplementary Data file 2). As an illustration of the high coverage of many of the proteins, Fig. 2 depicts the identification of peptides mapping to 93% of the chaperonin GroEL (Rv0440) gene product where we identified 65 unique peptides on the basis of 14,818 PSMs. Other predominant proteins identified in this analysis are molecular chaperone *DnaK* (Rv0350) with 70 peptides (7503 PSMs), *MetE* (Rv1133c) with 56 peptides (6260 PSMs), fatty acid synthase (Rv2524c) with 175 unique peptides (2609 PSMs), polyketide synthase (Rv3800c) with 112 unique peptides (3581 PSMs), and glutamine synthetase (Rv2220) with 35 unique peptides (3411 PSMs). Interestingly, we identified a peptide YLTWGLR mapping to gene Rv0157A, which is annotated as a probable pseudogene (11).

Identification of Hypothetical Proteins—Mycobacterium represents an evolutionarily distinct group of microorganisms. Many of the genes in *Mycobacteria* are functionally uncharacterized because of lack of sequence similarity to any of the known genes from model prokaryotes (8). In the TubercuList database, 1081 such proteins are present, which are categorized as conserved hypothetical proteins as these are conserved across related organisms (MTB complex, *M. smegmatis, M. marinum*) (13). We identified 829 proteins that are annotated as conserved hypothetical proteins of which 233 have not been shown to be translated by any earlier proteomic study.

Identification of Novel Protein Coding Genes in H37Rv Genome Using Genome Search Specific Peptides (GSSPs)—After excluding peptides which map to currently annotated proteins (from RefSeq database), the results from the genome database search provided a list of GSSPs. On the basis of these GSSPs, novel genes and gene structure refinements were proposed by proteogenomic analysis. Fig. 3 gives a schematic of our proteogenomic annotation strategy, which is described in detail in the methods section. Out of a total 36,785 unique peptides that were identified in this study, 243 peptides either mapped to regions of genome where no gene was present or did not agree with the gene model they were mapping to. Two gene prediction programs, *FgeneSB and GeneMark* as well as the *Orfind* tool from NCBI were used to find ORFs in the region to which these GSSPs were mapped (18, 19). Based on this, we were able to predict the presence of 41 novel protein coding genes. Further, we also checked the conservation of these predicted genes across mycobacterial or related species using protein blast algorithm. Forty of the novel ORFs were already annotated in other strains of

A

```
mAKTIAYDEEARRglerGLNALADAVKVTLGPKgrNVVLEKKWGAPTITNDGVSIAKEIE
LEDPYEKIGAELVKEVAKKTDDVAGDGTTTATVLAQALVREGLRNVAAGANPLGLKRgie
kAVEKVTETLLKGAKevetkEQIAATAAISAGDQSIGDLIAEAMDKVGNEGVITVEESNT
FGLQLELTEGMRFDKGYISGYFVTDPERQEAVLEDPYILLVSSKVSTVKDLLPLLEKVIG
AGKPLLIIAEDVEGEALSTLVVNKIRgtfksvavkAPGFGDRRKAMLQDMAILTGGQVIS
EEVGLTLENADLSLLGKarkVVVTKDETTIVEGAGDTDAIAGRVAQIRQEIENSDSDYDR
EKLQERlakLAGGVAVIKAGAATEVELKERkHRIEDAVRnakAAVEEGIVAGGGVTLLQA
APTLDELKLEGDEATGANIVKVALEAPLKQIAFNSGLEPGVVAEKvrNLPAGHGLNAQTG
VYEDLLAAGVADPVKVTRSALQNAASIAGLFLTTEAVVADKPEKEKASVPGGGDMGGMDF
```

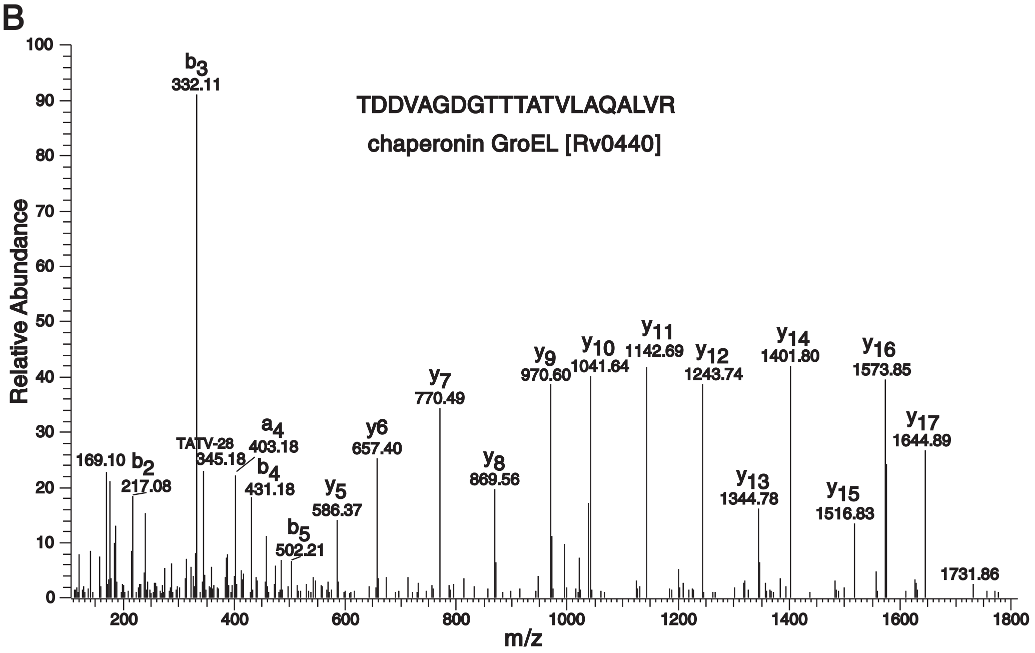


FIG. 2. **Identification of chaperonin GroEL (Rv0440) with representative MS/MS spectrum.** *A*, Chaperonin GroEL was identified on the basis of 65 unique peptides from a total 14,151 MS/MS spectra with 93% protein coverage. *B*, A representative MS/MS spectrum of peptide TDDVAGDGTTTATVLAQALVR identified from GroEL (Rv0440).

M. tuberculosis. It is interesting to note that the length of these novel proteins was short (an average length of 86.5 amino acids) and it is likely that some of these were missed from genome annotation of the reference strain owing to their small size. As mentioned earlier, TubercuList contains an additional 24 genes as compared with the RefSeq data set. We identified seven of these proteins, which are implicated in virulence, detoxification, and adaptation. Fig. 4 depicts identification of a novel 69 amino acid long ORF in the H37Rv genome along with the corresponding MS/MS spectra. Table I lists the novel genes found in this study along with details of the supporting peptide evidence and genome co-ordinates of the novel genes.

Changes in Current Gene Structures Using Peptide Data—Apart from identification of novel ORFs, we corrected the gene coordinates using genome search specific peptide data. Using this strategy, we corrected 79 gene models of which, 78 were N-terminal extensions and one was merging of two genes into one. TubercuList has annotated corrected coordinates (in agreement with our analysis) for six of these genes. Supplementary data file 3 gives the list of genes where modification in the structure is suggested in our analysis. It also contains information about old co-ordinates, modified co-ordinates, and corresponding peptide evidence. Fig. 5 depicts an example of correction of a gene model by extension of the N terminus (Rv2241). The current coordinates of the gene Rv2241 (*AceE*), which codes for pyruvate dehydrogenase E1 component are 2,512,452 to 2,515,244. We found peptides HDLAQNSNSASEPDR and HDLAQNSNSASEPDRVR mapping 66 bp upstream to the gene. *FgeneSB* and *GeneMark* predicted a longer gene model, which includes the peptide identified by us.

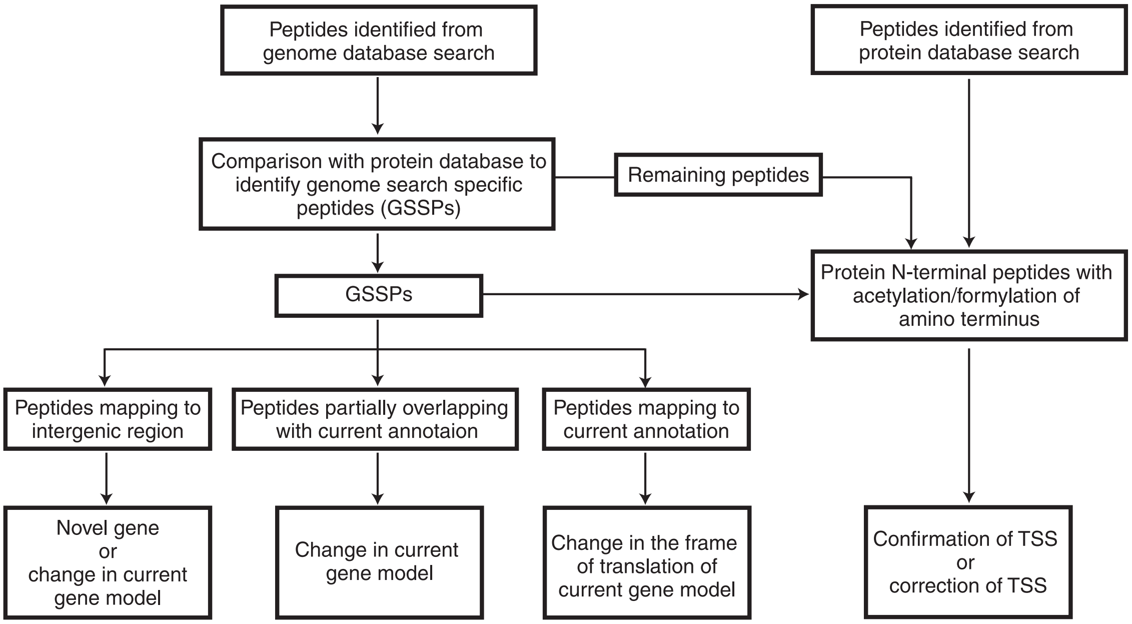

Peptides identified from genome database search
Peptides identified from protein database search
Comparison with protein database to identify genome search specific peptides (GSSPs)
Remaining peptides
GSSPs
Protein N-terminal peptides with acetylation/formylation of amino terminus
Peptides mapping to intergenic region
Peptides partially overlapping with current annotaion
Peptides mapping to current annotation
Novel gene or change in current gene model
Change in current gene model
Change in the frame of translation of current gene model
Confirmation of TSS or correction of TSS

FIG. 3. **Schematic representation of proteogenomic annotation strategy.** Peptide qualifying 1% FDR threshold from both protein and genome search were considered for proteogenomic analysis. From six frame translated genome database search results, peptides which mapped to known proteins from protein database were eliminated to obtain a set of genome search specific peptides (GSSPs). GSSPs were classified as mapping to intergenic region partially overlapping with current gene annotation and completely mapping within current gene annotations which were further used to modify the present genome annotation. N-terminal peptides identified from protein database searches were used to either confirm or propose change in the annotated translational start sites.

In case of gene Rv1181 (PKS4), a peptide TVVTASSFDELSAALR was found to partially overlap the N terminus of the gene. This finding suggested that actual start site of the protein was located further upstream of the annotated start site. Using a comparative genomic approach, a similar region in closely related strains was represented by one long protein whereas in H37Rv two separate CDSs, Rv1180 and Rv1181, were reported in the genome as a stop codon. However, a single nucleotide variation (A->C) has been reported at position 1,315,191 in the H37Rv genome, which converts the stop codon to a tyrosine codon supporting the presence of a longer protein instead of two independent coding regions (11).

In another example of N-terminal extension, peptides MIIDLHVQR and VLTIHGVTEHGR were found to map upstream of a gene Rv3203 (genome coordinates 3,580,638 to 3,580,638). However, these peptides are in +2 frame and Rv3203 is translated in +3 frame. We also identified seven unique peptides mapping to Rv3203 in +3 frame. A closer examination revealed that orthologous proteins in other mycobacterial strains including H37Ra and CDC1551 are longer by 37 amino acids, which includes both of the novel peptides that were identified. Ioerger *et al.* have reported an indel (T->-) at position 3,580,637 in the H37Rv genome, which is a probable sequencing error in the H37Rv reference genome (24). Our findings indicate that there is indeed a sequencing error and that the coordinates of the gene Rv3203 should be corrected to be 3,580,526 to 3,581,309.

Correction of gene coordinates was also carried out using protein N-terminal peptides as described in the following section.

Identification of N-terminal Peptides and Confirmation/Correction of Translational Start Sites (TSS)—Approximately 60% of the genes in H37Rv genome use ATG as start codon, ~33% use GTG as start codon and ~5% use TTG as a start codon. As described in the methods section, while creating six frame translated database from genome we added variant peptides in which GTG and TTG codons were translated into methionine residues instead of valine or leucine, respectively. Protein N-terminal peptides can be indicated either by post-translational modifications like acetylation or nontryptic nature at the N terminus of the peptide. We used these criteria to assign correct translational start sites. We permitted semitryptic peptides in MS/MS search of the genome database for the purpose of identifying such peptides. We looked for fMet containing peptides, deformylated protein N-terminal peptides, peptides with initiator methionine cleaved, and peptides with acetylated N terminus after initiator methionine was deformylated or cleaved to identify protein N-terminal peptides. N-terminal acetylation after removal of initiator methionine is known to occur in prokaryotes, however, at lower frequency than in eukaryotes. Rison *et al.* have shown that fraction of proteins that are acetylated at the N-terminal in *M. tuberculosis* is more than that found in *E. coli* (25). We found 874 protein N-terminal peptides out of which, 253 peptides were acetylated at the

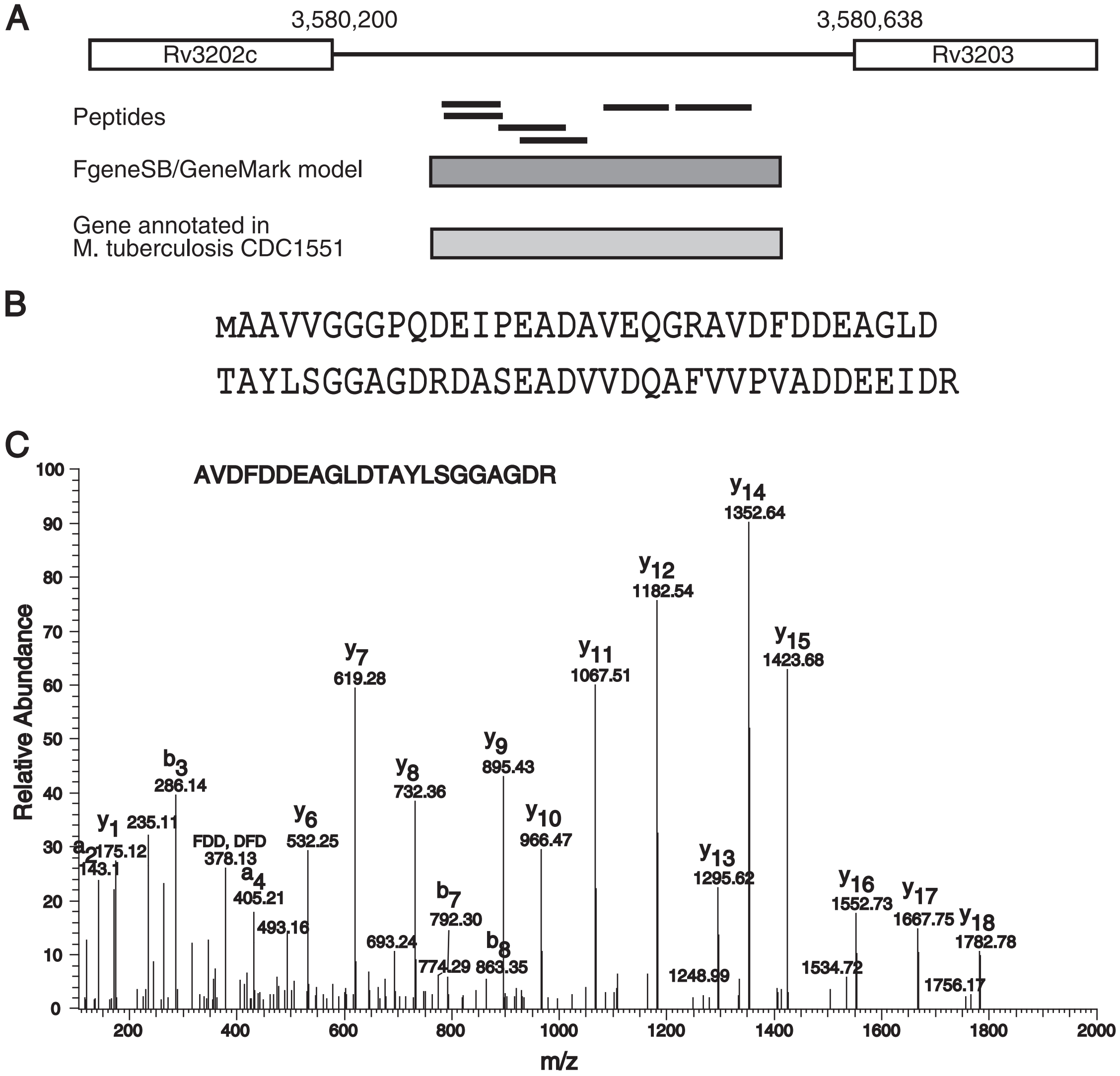


FIG. 4. **Identification of a novel gene based on peptides mapping to an intergenic region.** *A*, Six peptides mapped to an intergenic region between genes Rv3202c and Rv3203. Gene prediction algorithms *FgeneSB* and *GeneMark* supported the presence of this additional gene. In addition, a protein corresponding to this novel gene has been annotated in *M. tuberculosis* CDC1551 genome. *B*, Protein sequence of a novel gene. Identified region is indicated by taller text. *C*, A representative MS/MS spectrum for identification of genome search specific peptide AVDFDDEAGLDTAYLSGGAGDR is shown.

N termini, 237 peptides had initiator methionine deformylated, and 384 peptides had cleaved initiator methionine. Thus, we were able to confirm the TSS of 727 proteins (supplementary Data file 5) and reannotate TSS of 33 other proteins (supplementary Data file 4). We also identified N-terminal peptides which confirmed the translational start site of four Novel ORFs (Table I).

We found 21 examples of N-terminal peptides mapping upstream to the currently annotated translational start site of the gene. On the other hand, we corrected TSSs of 12 genes based on a protein N-terminal peptide mapping downstream of the annotated TSS. One such example is illustrated in Fig. 6 where a peptide with nontryptic N terminus, M.GDASLTTELGR.V, was mapped downstream of the annotated TSS of gene Rv1106c, which can be used as evidence for shortening the CDS length of the gene model. Interestingly, in the case of eight proteins (Rv2175c, Rv2847c, Rv3133c, Rv1683, Rv1612, Rv3131, Rv2986c, and Rv3001c) we found peptide evidence indicating translation initiation at two different sites.

DISCUSSION

M. tuberculosis H37Rv genome sequence has been available for more than 10 years and has been characterized previously by a number of groups. Thus, it was surprising to identify many novel regions with protein coding potential. Even more surprising was the fact that most of these novel ORFs were already annotated in other strains of *M. tuberculosis* but were missing from the primary genome se-

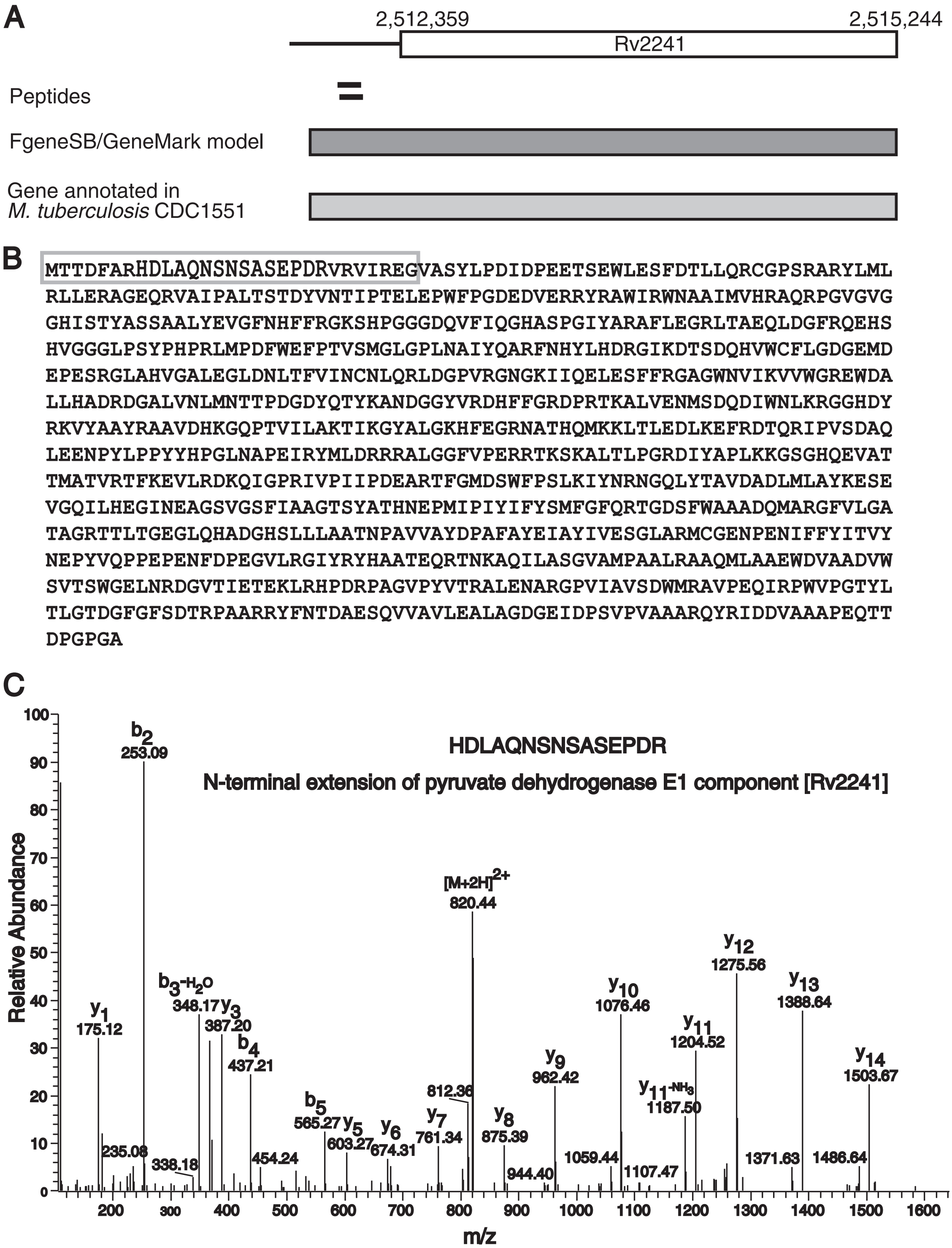
Table I

Identification of Novel ORFs using peptide evidence. Note: Score indicated here is the highest score obtained for the peptide sequence by individual search engines.

Novel ORF No.	Peptide	Mascot score	Sequest score	MassWiz score	Annotated in other M.tb strains?	Annotated in TubercuList?	Genome co-ordinates of novel ORF	Length (amino acids)
Novel ORF#1	ELFGPNPIEPPTDIAPDPDSTK	73.09	3.38	–	Yes	no	31824–31967	47
Novel ORF#2	WGFGDLAVCDGEK	75.2	–	–	Yes	No	65012–65350	112
Novel ORF#2	PHQPDMTK	55.96	–	–	Yes	No	65012–65350	112
Novel ORF#2	DPPDPHQPDMTK	61.49	–	–	Yes	No	65012–65350	112
Novel ORF#3	MTRPGAEEGDSAGPTSTR	38.35	–	972.62	Yes	No	87798–88004	68
Novel ORF#4	LVGAVRLPVEHR	53.72	–	–	Yes	No	197612–198286	224
Novel ORF#5	CTCGDELAPVK	64.89	–	–	Yes	Yes (Rv0186A)	218390–218536	48
Novel ORF#5	IEVPcHcAGAGDAYR	35.92	3.63	1673.7	Yes	Yes (Rv0186A)	218390–218536	48
Novel ORF#5	VRIEVPCHCAGAGDAYR	91.28	3.55	1901.94	Yes	Yes (Rv0186A)	218390–218536	48
Novel ORF#5	TNYEAGTLLTCSHEGCGCR	49.43	–	2254.13	Yes	Yes (Rv0186A)	218390–218536	48
Novel ORF#6	VAAALDTLAAAPPEDR	113.61	3.6	1863.07	Yes	No	274710–274862	50
Novel ORF#7	VIGPDDDPEFLRR	–	–	950.68	Yes	No	622121–622282	53
Novel ORF#7	RVIGPDDDPEFLR	53.15	–	–	Yes	No	622121–622282	53
Novel ORF#8	GQAGIVDDGAVLIHVPGECPHPGEHVPRS	–	4.14	1797.41	Yes	No	665434–665601	55
Novel ORF#8	GQAGIVDDGAVLIHVPGECPHPGEHVPR	–	5.77	1942.3	Yes	No	665434–665601	55
Novel ORF#8	RGQAGIVDDGAVLIHVPGECPHPGEHVPR	–	4.11	1302.51	Yes	No	665434–665601	55
Novel ORF#8	QTIEPGWLYITAHR	60.36	4.38	1583.66	Yes	No	665434–665601	55
Novel ORF#8	CKQTIEPGWLYITAHR	–	4.61	901.11	Yes	No	665434–665601	55
Novel ORF#9	EGAGHFNR	45.31	–	976.96	Yes	No	704478–704654	86
Novel ORF#10	SEGHSAGGR	–	–	836.42	Yes	No	1836814–1837032	72
Novel ORF#11	IIPASAGPLDSLISTGSVQPAR	–	2.95	1292.08	Yes	Yes (Rv1962A)	2205277–2205540	87
Novel ORF#12	EAQLLDILR	57.45	–	757.46	Yes	Yes (Rv1982A)	2225841–2226101	86
Novel ORF#12	HALSAQLAFLESR	37.12	–	1278.73	Yes	Yes (Rv1982A)	2225841–2226101	86
Novel ORF#13	YLHELDAQLLTGQIDR	–	4.32	928.42	Yes	Yes (Rv1991A)	2234643–2234891	82
Novel ORF#14	MGITSVSVHSGAIAATPGSVAAAER	45.96	5.81	1519.32	Yes	No	2249260–2249457	65
Novel ORF#15	ANDTDDAHWSTIDDFDKR	–	5.16	2165.79	Yes	Yes (Rv2142A)	2402507–2402722	71
Novel ORF#16	ELLEQISTLIR	57.26	–	1440.27	Yes	No	2510351–2510548	65
Novel ORF#16	ADMRELLEQISTLIR	41.99	–	1438.63	Yes	No	2510351–2510548	65
Novel ORF#16	ADLYAAVDAMR	68.51	3.82	1246.3	Yes	No	2510351–2510548	65
Novel ORF#16	QTAPPVRPMTSDQLPATK	42.97	–	1404.7	Yes	No	2510351–2510548	65
Novel ORF#16	GERPGFQSDSAAR	73.3	–	–	Yes	No	2510351–2510548	65
Novel ORF#17	LSEPWTHPPILWAATDEVVGSAHGGHGH DASEFTVGGGASGTW	52.66	–	1717.11	Yes	No	2772098–2772331	77
Novel ORF#18	GKLDGAAAGQQR	–	3.17	970.98	Yes	No	2869253–2869627	124
Novel ORF#18	FSHVEER	–	–	947.8	Yes	No	2869253–2869627	124
Novel ORF#18	ATIGSFNALREDFTALR	36.31	–	726.81	Yes	No	2869253–2869627	124
Novel ORF#18	ATIGSFNALR	64.75	–	–	Yes	No	2869253–2869627	124
Novel ORF#18	VLAGAADRDVTEFVGEFR	58.58	–	1027.44	Yes	No	2869253–2869627	124
Novel ORF#18	ASEQDAAAAR	77.12	–	1052.12	Yes	No	2869253–2869627	124
Novel ORF#19	PASLIDMR	50.09	–	–	Yes	No	3356736–3357011	91
Novel ORF#20	MAVDVDQWPTVGQILPVVYSPK	53.25	3.23	1063.28	Yes	No	3502945–3503277	110
Novel ORF#21	VWITTSRPGEEPTRIEVVLIAAYR	–	–	620.84	Yes	No	3540988–3541167	59
Novel ORF#22	DTDIGQPCSPEGAK	64.09	–	–	Yes	No	476394–476639	81
Novel ORF#22	LWGNPGPIYCER	64.25	–	1273.71	Yes	No	476394–476639	81

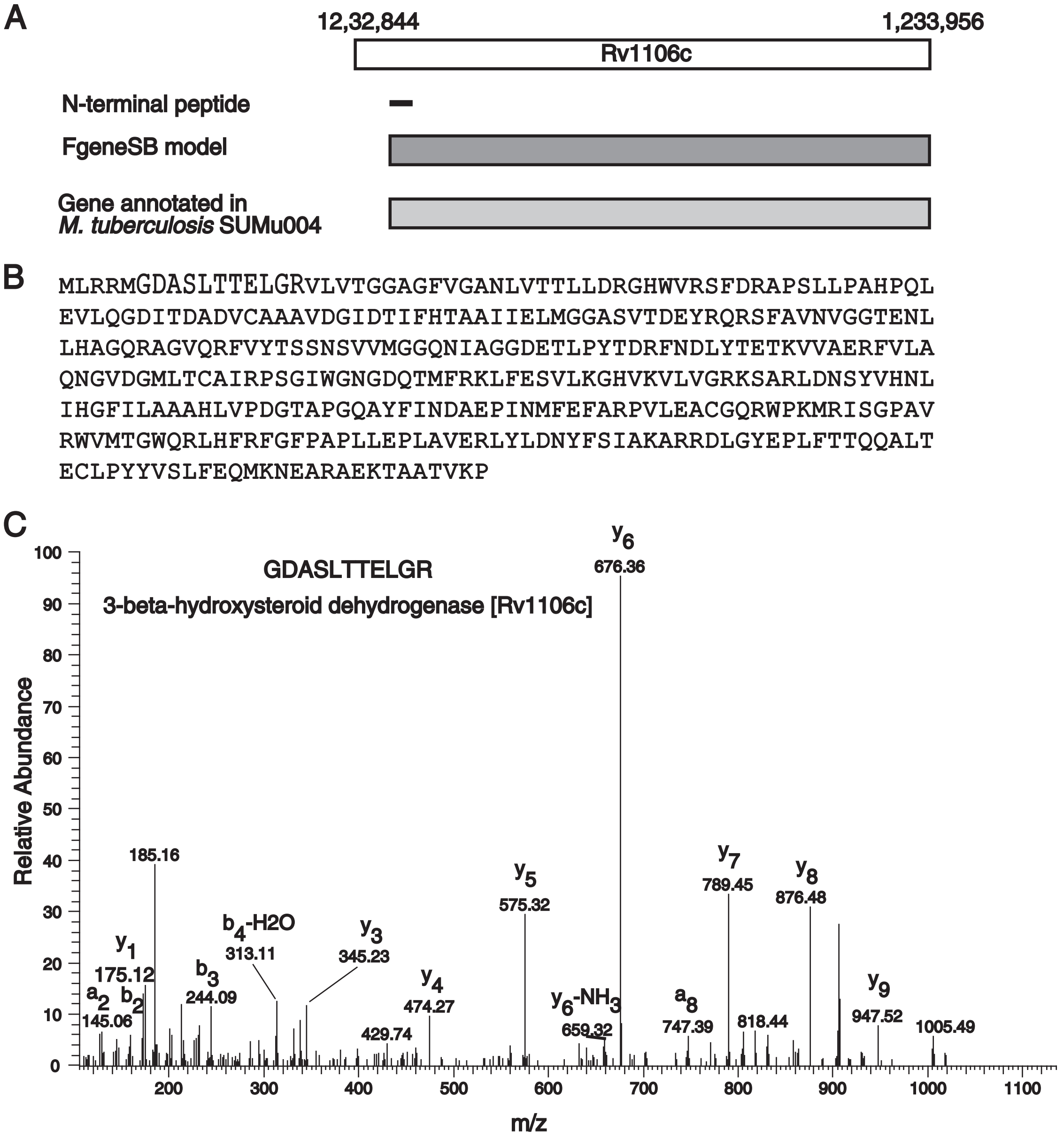
TABLE I—*continued*

Novel ORF No.	Peptide	Mascot score	Sequest score	MassWiz score	Annotated in other M.tb strains?	Annotated in TubercuList?	Genome co-ordinates of novel ORF	Length (amino acids)
Novel ORF#23	TTIDLPQDLHK	40.25	–	699.4	Yes	Yes (Rv0616A)	710782–711009	75
Novel ORF#23	TLSETVADLMR	51.28	–	841.44	Yes	Yes (Rv0616A)	710782–711009	75
Novel ORF#23	TGLPLVSVGTVVTSEDVR	73.85	–	1289.42	Yes	Yes (Rv0616A)	710782–711009	75
Novel ORF#24	WVGLAGVAGVVAGGALVAR	59.44	–	1526.96	Yes	No	1242712–1242912	66
Novel ORF#24	RAYTPDEVR	45.51	3.17	946.13	Yes	No	1242712–1242912	66
Novel ORF#25	SPQESSSEGETK	66.39	–	1192.12	Yes	No	1282030–1282218	62
Novel ORF#26	AVAGMVGGGTGLGGAGR	40.29	–	–	Yes	No	1414903–1415058	51
Novel ORF#27	RITEFATPQEAMEHR	44.92	4.16	868.83	Yes	No	1539204–1539440	78
Novel ORF#27	ITEFATPQEAMEHR	52.82	3.73	1351.87	Yes	No	1539204–1539440	78
Novel ORF#28	AETDQRDLR	53.9	–	–	Yes	No	1960837–1961037	66
Novel ORF#29	SRTPLRPPVAPSEGVAADSVAVCR	–	6	1426.07	Yes	Yes (Rv2063A)	2321057–2321467	136
Novel ORF#30	LADGEEVHGECDELTINPATGVLTVCR	55.96	4.53	1658.51	Yes	No	2680458–2680667	69
Novel ORF#30	VDGFEETTTHYSPSAWR	–	–	1023.85	Yes	No	2680458–2680667	69
Novel ORF#30	GVGVRPSLVSTAQ	–	–	800.85	Yes	No	2680458–2680667	69
Novel ORF#31	AKLPSGAELLFCQHHANEHEAK	–	6.48	1261.85	Yes	No	3020323–3020511	62
Novel ORF#32	VHNLDPELVDEHAR	44.44	6.48–	2186.32	Yes	No	3056019–3056423	134
Novel ORF#33	LAPMWLPFR	51.81	–	–	Yes	No	3392812–3393201	129
Novel ORF#33	YGPFRVEAPLSSVR	–	–	1050.95	Yes	No	3392812–3393201	129
Novel ORF#33	WWTAVGPR	41.03	–	–	Yes	No	3392812–3393201	129
Novel ORF#34	LRDEDVSTEPDAW	–	–	827.68	Yes	No	3467352–3467612	86
Novel ORF#35	TIGIVYVHGDPVDYLDRDQMAK	–	4.76	1310.96	Yes	No	3467609–3467926	105
Novel ORF#36	AGEVDTVVAGPADDR	61.4	–	976.81	No	No	3493222–3493704	160
Novel ORF#37	LPLSASVWDIAQR	100.17	–	1129.16	Yes	No	3556855–3557061	68
Novel ORF#38	VNHAASAISPSLNENSSSGSPK	54.57	4.93	1480.7	Yes	No	3571628–3571799	57
Novel ORF#39	AAVVGGGPQDEIPEADAVEQGR	151.94	–	–	Yes	No	3580286–3580495	69
Novel ORF#39	AVVGGGPQDEIPEADAVEQGR	145.58	–	–	Yes	No	3580286–3580495	69
Novel ORF#39	AVDFDDEAGLDTAYLSGGAGDR	120.04	4.89	1971.04	Yes	No	3580286–3580495	69
Novel ORF#39	DASEADVVDQAFVVPVADDEEIDR	91.67	3.21	1348.58	Yes	No	3580286–3580495	69
Novel ORF#40	HLRGVAGALGR	–	–	1120.33	Yes	No	4258204–4260120	638
Novel ORF#41	MQLRDILSLLGHR	–	2.91	–	Yes	No	4361278–4361580	100



```
MTTDFARHDLAQNSNSASEPDRVRVIREGVASYLPDIDPEETSEWLESFDTLLQRCGPSRARYLML
RLLERAGEQRVAIPALTSTDYVNTIPTELEPWFPGDEDVERRYRAWIRWNAAIMVHRAQRPGVGVG
GHISTYASSAALYEVGFNHFFRGKSHPGGGDQVFIQGHASPGIYARAFLEGRLTAEQLDGFRQEHS
HVGGGLPSYPHPRLMPDFWEFPTVSMGLGPLNAIYQARFNHYLHDRGIKDTSDQHVWCFLGDGEMD
EPESRGLAHVGALEGLDNLTFVINCNLQRLDGPVRGNGKIIQELESFFRGAGWNVIKVVWGREWDA
LLHADRDGALVNLMNTTPDGDYQTYKANDGGYVRDHFFGRDPRTKALVENMSDQDIWNLKRGGHDY
RKVYAAYRAAVDHKGQPTVILAKTIKGYALGKHFEGRNATHQMKKLTLEDLKEFRDTQRIPVSDAQ
LEENPYLPPYYHPGLNAPEIRYMLDRRRALGGFVPERRTKSKALTLPGRDIYAPLKKGSGHQEVAT
TMATVRTFKEVLRDKQIGPRIVPIIPDEARTFGMDSWFPSLKIYNRNGQLYTAVDADLMLAYKESE
VGQILHEGINEAGSVGSFIAAGTSYATHNEPMIPIYIFYSMFGFQRTGDSFWAAADQMARGFVLGA
TAGRTTLTGEGLQHADGHSLLLAATNPAVVAYDPAFAYEIAYIVESGLARMCGENPENIFFYITVY
NEPYVQPPEPENFDPEGVLRGIYRYHAATEQRTNKAQILASGVAMPAALRAAQMLAAEWDVAADVW
SVTSWGELNRDGVTIETEKLRHPDRPAGVPYVTRALENARGPVIAVSDWMRAVPEQIRPWVPGTYL
TLGTDGFGFSDTRPAARRYFNTDAESQVVAVLEALAGDGEIDPSVPVAAARQYRIDDVAAAPEQTT
DPGPGA
```

FIG. 5. **N-terminal extension of a gene using peptides mapping upstream to an annotated start site.** *A*, Two peptides were mapped to the upstream region of gene Rv2241. Gene prediction algorithms *FgeneSB* and *GeneMark* predicted presence of longer gene extending N-terminal of Rv2241 by 29 amino acids. *M. tuberculosis* CDC1551 ortholog supports the N-terminal extension. *B*, Protein sequence of modified gene model for Rv2241. Sequence in gray box indicates the extended N terminus. *C*, A representative MS/MS spectrum for the peptide HDLAQNSNSASEPDR is shown.



MLRRMGDASLTTELGRVLVTGGAGFVGANLVTTLLDRGHWVRSFDRAPSLLPAHPQL
EVLQGDITDADVCAAAVDGIDTIFHTAAIIELMGGASVTDEYRQRSFAVNVGGTENL
LHAGQRAGVQRFVYTSSNSVVMGGQNIAGGDETLPYTDRFNDLYTETKVVAERFVLA
QNGVDGMLTCAIRPSGIWGNGDQTMFRKLFESVLKGHVKVLVGRKSARLDNSYVHNL
IHGFILAAAHLVPDGTAPGQAYFINDAEPINMFEFARPVLEACGQRWPKMRISGPAV
RWVMTGWQRLHFRFGFPAPLLEPLAVERLYLDNYFSIAKARRDLGYEPLFTTQQALT
ECLPYYVSLFEQMKNEARAEKTAATVKP

FIG. 6. **Correction of translational start site by identification of N-terminal peptide.** *A*, A Peptide GDASLTTELGR with a nontryptic N terminus was mapped downstream of the annotated translational start site of the gene Rv1106c. Gene model by Gene prediction program *FgeneSB* and orthologous protein from *M. tuberculosis* Sumu004 support the re-annotation of the TSS by shortening protein length by 4 amino acids. *B*, Protein sequence of Rv1106c is shown. Text in gray indicates amino acids removed in the revised gene model for Rv1106c. *C*, A representative MS/MS spectrum for the peptide GDASLTTELGR is shown.

quence of *M. tuberculosis*. We have demonstrated the power of the proteogenomic approach to annotate a genome and refine the annotation of a well studied genome. Though relatively new, this approach of using mass spectrometry-based proteomic data to identify protein coding regions in the genome can prove to be an essential complementary method along with computational methods for annotating newly sequenced genomes in the future.

The Data Associated With This Manuscript May be Downloaded from the ProteomeCommons.org Tranche Network Using the Following Hash—Raw data files from SCX fractions of *M. tb* lysates-levqcXkyUcwFeX7HvH6I1bpG5FU3CRCC-hKdNdBw9rMbqENYrrkYGs3FEn9FCHbACDEYcHWVEHp/f-gdsNZo7JjHTTG7AAAAAAAAAPuw = =

Raw data files from offgel fractions of *M. tb* lysates-yFxQ0-dmYGnRAOEUvYelrONUysTDqJNRUO6R9vI3duPd8J0RwE-gPCM33VHRGqoA+B89hYmRK2MEGAEtGh5BnF7NNVtMc-AAAAAAAAPNw = =

Raw data files from ingel fractions of *M. tb* lysates-qiSIhC-XpIQhe0HbOPkZ56QCR5T9Q6nPUQPpIQ8fgZ42KEPS2I Nj4v8n9Pk3r9O9EYS6fiTwxNpAiBV9IVsgWOVRrqqoAAAA-AAAAMSA = =

Raw data files from *M. tb* culture filtrates-SI4cLiunsLWXUS8ksiEaxrJfbS97/2koDxJCPc134jV8+T/oV0sLQRXf+vlYyaaCkvzN4QKNW5A9eLR4JCrKfCiA3/IAAAAAAAABiA = =

Peak list files-Nep0iN9CduLbC+j7UhzE1kfF4O7vaCskEMcbWcTBoAuMEt9QN1Oy6Hj0x+eCujysogXrS45dJkrW26wuPD79lnaXVxUAAAAAAAAzkA = =

Acknowledgments—We thank the Department of Biotechnology (DBT), India for research support to the Institute of Bioinformatics. We thank Agilent Technologies for instrumentation support.

* Dhanashree S. Kelkar is the recipient of a Senior Research Fellowship from the University Grants Commission (UGC), India. Amit Kumar Yadav is the recipient of a Senior Research Fellowship from Council of Scientific and Industrial Research (CSIR), India. Dhirendra Kumar is the recipient of DST-INSPIRE Junior Research Fellowship from the Department of Science and Technology (DST), India. Dr. Debasis Dash is the recipient of an Indo-US Research Fellowship awarded by the Indo-US Science and Technology Forum. Dr. T. S. Keshava Prasad is the recipient of research grants on “Establishment of a National Database on Tuberculosis” and “Development of Infrastructure and a Computational Framework for Analysis of Proteomic Data” from DBT. Dr. Raghothama Chaerkady is supported by a research grant “Genomic and Proteomic Analysis of Chronic Meningitis” from DBT.

S This article contains supplemental Data S1 to S5.

[b] To whom correspondence should be addressed: McKusick-Nathans Institute of Genetic Medicine and Departments of Biological Chemistry, Oncology, and Pathology, Johns Hopkins University, Baltimore, MD 21205. Fax: 1-410-502-7544; E-mail: pandey@jhmi.edu.

Authors' contribution: AP, DD and SR initiated the project. DSK, PS, KK, HS, RK, DC, SA procured the samples. DSK was responsible for sample processing and fractionation. DSK and RC were responsible for mass spectrometric analysis. AY, DK and DD developed the MassWiz algorithm. DSK, DK designed and carried out data analysis. PK (IOB), AY, LB, BM, AM and HCH assisted in bioinformatics analyses. DSK, AP drafted the initial manuscript with additional input provided by TSKP, DD, SR, KK, VMK, HCH, BN, PK (NTI). This manuscript has been seen and approved by all authors.

REFERENCES

1. Jungblut, P. R., Schaible, U. E., Mollenkopf, H. J., Zimny-Arndt, U., Raupach, B., Mattow, J., Halada, P., Lamer, S., Hagens, K., and Kaufmann, S. H. (1999) Comparative proteome analysis of Mycobacterium tuberculosis and Mycobacterium bovis BCG strains: towards functional genomics of microbial pathogens. *Mol. Microbiol.* **33,** 1103–1117
2. Jungblut, P. R., Müller, E. C., Mattow, J., and Kaufmann, S. H. (2001) Proteomics reveals open reading frames in Mycobacterium tuberculosis H37Rv not predicted by genomics. *Infect. Immun.* **69,** 5905–5907
3. Gu, S., Chen, J., Dobos, K. M., Bradbury, E. M., Belisle, J. T., and Chen, X. (2003) Comprehensive proteomic profiling of the membrane constituents of Mycobacterium tuberculosis strain. *Mol. Cell. Proteomics* **2,** 1284–1296
4. Mawuenyega, K. G., Forst, C. V., Dobos, K. M., Belisle, J. T., Chen, J., Bradbury, E. M., Bradbury, A. R., and Chen, X. (2005) Mycobacterium tuberculosis functional network analysis by global subcellular protein profiling. *Mol. Biol. Cell* **16,** 396–404
5. Mattow, J., Siejak, F., Hagens, K., Becher, D., Albrecht, D., Krah, A., Schmidt, F., Jungblut, P. R., Kaufmann, S. H., and Schaible, U. E. (2006) Proteins unique to intraphagosomally grown Mycobacterium tuberculosis. *Proteomics* **6,** 2485–2494
6. Målen, H., Berven, F. S., Fladmark, K. E., and Wiker, H. G. (2007) Comprehensive analysis of exported proteins from Mycobacterium tuberculosis H37Rv. *Proteomics* **7,** 1702–1718
7. Målen, H., Pathak, S., Softeland, T., de Souza, G. A., and Wiker, H. G. (2010) Definition of novel cell envelope associated proteins in Triton X-114 extracts of Mycobacterium tuberculosis H37Rv. *BMC Microbiol.* **10,** 132
8. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., and Barrell, B. G. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393,** 537–544
9. Fleischmann, R. D., Alland, D., Eisen, J. A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E., Kolonay, J. F., Nelson, W. C., Umayam, L. A., Ermolaeva, M., Salzberg, S. L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jacobs Jr., W. R., Jr., Venter, J. C., and Fraser, C. M. (2002) Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. *J. Bacteriol.* **184,** 5479–5490
10. Camus, J. C., Pryor, M. J., Médigue, C., and Cole, S. T. (2002) Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology* **148,** 2967–2973
11. The Tuberculist Database [http://tuberculist.epfl.ch]
12. de Souza, G. A., Målen, H., Softeland, T., Saelensminde, G., Prasad, S., Jonassen, I., and Wiker, H. G. (2008) High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using Mycobacterium tuberculosis as an example. *BMC Genomics* **9,** 316
13. Lew, J. M., Kapopoulou, A., Jones, L. M., and Cole, S. T. (2011) TubercuList - 10 years after. *Tuberculosis* **91,** 1–7
14. Pandey, A., and Lewitter, F. (1999) Nucleotide sequence databases: a gold mine for biologists. *Trends Biochem. Sci.* **24,** 276–280
15. Mann, M., and Pandey, A. (2001) Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem. Sci.* **26,** 54–61
16. Castellana, N., and Bafna, V. (2010) Proteogenomics to discover the full coding content of genomes: a computational perspective. *J. Proteomics* **73,** 2124–2135
17. de Souza, G. A., Arntzen, M. Ø., Fortuin, S., Schürch, A. C., Målen, H., McEvoy, C. R., van Soolingen, D., Thiede, B., Warren, R. M., and Wiker, H. G. (2011) Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. *Mol. Cell. Proteomics* **10,** M110.002527
18. Besemer, J., Lomsadze, A., and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29,** 2607–2618
19. FGENESB: Bacterial Operon and Gene Prediction [http://linux1.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb]
20. Amanchy, R., Kalume, D. E., and Pandey, A. (2005) Stable isotope labeling with amino acids in cell culture (SILAC) for studying dynamics of protein abundance and posttranslational modifications. *Sci. STKE.* 2005, pl2
21. Harsha, H. C., Molina, H., and Pandey, A. (2008) Quantitative proteomics using stable isotope labeling with amino acids in cell culture. *Nat. Protoc.* **3,** 505–516
22. Yadav, A. K., Kumar, D., and Dash, D. (2011) MassWiz: A Novel Scoring Algorithm with Target-Decoy Based Analysis Pipeline for Tandem Mass Spectrometry. *J. Proteome. Res.* **10,** 2154–2160
23. Deutsch, E.W., Lam, H., and Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO. Rep.* **9,** 429–434
24. Ioerger, T. R., Feng, Y., Ganesula, K., Chen, X., Dobos, K. M., Fortune, S., Jacobs, W. R., Jr., Mizrahi, V., Parish, T., Rubin, E., Sassetti, C., and Sacchettini, J. C. (2010) Variation among genome sequences of H37Rv strains of Mycobacterium tuberculosis from multiple laboratories. *J. Bacteriol.* **192,** 3645–3653
25. Rison, S. C., Mattow, J., Jungblut, P. R., and Stoker, N. G. (2007) Experimental determination of translational starts using peptide mass mapping and tandem mass spectrometry within the proteome of Mycobacterium tuberculosis. *Microbiology* **153,** 521–528