frontiers
in Microbiology

ORIGINAL RESEARCH
published: 18 December 2017
doi: 10.3389/fmicb.2017.02505

Transcriptome Landscape of *Mycobacterium smegmatis*

Xinfeng Li[1†], Han Mei[1†], Fang Chen[1], Qing Tang[1], Zhaoqing Yu[1], Xiaojian Cao[1], Binda T. Andongma[1], Shan-Ho Chou[2] and Jin He[1]*

[1] *State Key Laboratory of Agricultural Microbiology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan, China,* [2] *Institute of Biochemistry and NCHU Agricultural Biotechnology Center, National Chung Hsing University, Taichung, Taiwan*

The non-pathogenic bacterium *Mycobacterium smegmatis* mc^{2}155 has been widely used as a model organism in mycobacterial research, yet a detailed study about its transcription landscape remains to be established. Here we report the transcriptome, expression profiles and transcriptional structures through growth-phase-dependent RNA sequencing (RNA-seq) as well as other related experiments. We found: (1) 2,139 transcriptional start sites (TSSs) in the genome-wide scale, of which eight samples were randomly selected and further verified by 5′-RACE; (2) 2,233 independent monocistronic or polycistronic mRNAs in the transcriptome within the operon/sub-operon structures which are classified into five groups; (3) 47.50% (1016/2139) genes were transcribed into leaderless mRNAs, with the TSSs of 41.3% (883/2139) mRNAs overlapping with the first base of the annotated start codon. Initial amino acids of MSMEG_4921 and MSMEG_6422 proteins were identified by Edman degradation, indicating the presence of distinctive widespread leaderless features in *M. smegmatis* mc^{2}155. (4) 150 genes with potentially wrong structural annotation, of which 124 proposed genes have been corrected; (5) eight highly active promoters, with their activities further determined by β-galactosidase assays. These data integrated the transcriptional landscape to genome information of model organism mc^{2}155 and lay a solid foundation for further works in *Mycobacterium*.

Keywords: *Mycobacterium smegmatis*, transcriptome, transcriptional start site, leaderless mRNA, gene structural re-annotation, operon, sub-operon, highly active promoter

OPEN ACCESS

Edited by:
Angel Angelov,
Technische Universität München,
Germany

Reviewed by:
Jörg Soppa,
Goethe Business School, Germany
Philippe H. Noirot,
Argonne National Laboratory (DOE),
United States

****Correspondence:***
Jin He
hejin@mail.hzau.edu.cn

† *These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *29 August 2017*
Accepted: *01 December 2017*
Published: *18 December 2017*

Citation:
Li X, Mei H, Chen F, Tang Q, Yu Z, Cao X, Andongma BT, Chou S-H and He J (2017) Transcriptome Landscape of Mycobacterium smegmatis. Front. Microbiol. 8:2505. doi: 10.3389/fmicb.2017.02505

INTRODUCTION

Mycobacterium tuberculosis is the pathogen responsible for tuberculosis, which is second only to HIV/AIDS as the greatest cause of death worldwide. In fact, in 2015, 10.4 million people suffered from tuberculosis with 1.8 million deaths from the disease (World Health Organization WHO, 2016). The slow-growing and pathogenic features make it difficult to study on *M. tuberculosis.* Being a "fast grower" and non-pathogen, *M. smegmatis* has been used as a research model for *M. tuberculosis* to investigate on a wide variety of mycobacterial physiological processes, including drug resistance, dormancy, fatty acid metabolism, gene regulatory networks (Morbidoni et al., 2006; Cordone et al., 2011; Yang et al., 2012; Deng et al., 2015; Liu et al., 2016).

RNA-seq has recently emerged as a method allowing people to study RNA-based regulation in a genome-wide manner. Compared to microarray, RNA-seq exhibits unique advantages including: higher sensitivity, higher throughput, larger dynamic range, better genome sequence independence, and in single-nucleotide resolution (Wang et al., 2009), which make it more ideal for comprehensive systematic and accurate transcriptome analyses. There are two types of RNA-seq experiments, namely typical RNA-seq and differential RNA-seq (dRNA-seq) experiments. In a typical RNA-seq experiment, RNAs are converted to a cDNA library and amplified by PCR. Following deep sequencing, the sequences obtained are mapped to a reference genome for further data analysis. In the dRNA-seq library method, the original RNA sample is first treated with 5′-monophosphate-RNAs (5′-p-RNAs) terminator exonuclease (TEX) to specifically degrades 5′-p-RNAs but retain 5′-triphosphate-RNAs (5′-ppp-RNAs), thus relatively enriching the primary transcripts (5′-ppp-RNAs) (Sharma and Vogel, 2014). This method has been widely used to identify the primary 5′-ends in various bacteria (Sharma et al., 2010; Wurtzel et al., 2010; Cortes et al., 2013; Thomason et al., 2015; Babski et al., 2016). Compared to dRNA-seq, typical RNA-seq can also be used to identify the 5′-ends of transcripts, but it cannot distinguish the 5′-ends of primary transcripts from the 5′-ends of processed transcripts. However, the typical RNA-seq has unique advantages in quantifying gene expression levels and identifying operon structures.

In prokaryotic genomes, the typical RNA-seq could be applied to: (1) uncover the global transcriptional profile in genome scale, identify differentially expressed genes (DEGs) and determine the differences in transcriptional levels between wild type and its mutants at distinct conditions (Wang et al., 2013b; Li et al., 2016; Hendrickson et al., 2017); (2) identify 5′-ends of transcript (Wang et al., 2013a; Liao et al., 2015); single-nucleotide resolution enables RNA-seq to globally identify the first nucleotide in 5′-ends. Previously, 5′-ends of transcript were generally identified by 5′-Rapid Amplification of cDNA Ends (5′-RACE), which is troublesome and costly; (3) correct gene structural annotation (Perkins et al., 2009); currently, the majority of sequenced bacterial genomes are annotated by an algorithm-based program which is automated but error-prone. The combination of transcriptome data with algorithmic predictions could greatly facilitate gene structural annotation. (4) discover new genes (Arnvig et al., 2011; Ignatov et al., 2013). Mapping RNA-seq data to genome makes it possible to identify new genes (especially small non-coding RNAs) in intergenic region, within a gene, as well as those in the negative strand; (5) define operon and sub-operon structures (Wang et al., 2013; Fortino et al., 2014). Some genes in prokaryotes are organized into operons, which are co-transcribed to generate polycistronic transcripts. Through transcriptome mapping, we can clearly identify whether genes are transcribed together. Moreover, in operon, some internal genes possess their own TSSs (a typical symbol of sub-operon), which can be also elucidated by RNA-seq.

For *Mycobacterium*, RNA-seq has also been applied. Wang et al. (2013) applied RNA-seq to depict the operon structures in *M. marinum* and Cortes et al. (2013) used dRNA-seq to globally identify TSSs in *M. tuberculosis*. In mc^{2}155, Shell et al. (2015) revealed a high abundance of leaderless mRNAs and small proteins by the combination of RNA-seq and ribosome profiling data. Other studies mainly focused on the identification of DEGs in different conditions (Petridis et al., 2015; Li et al., 2016; Wu et al., 2016; Hillion et al., 2017). However, no comprehensive insights have been achieved regarding *M. smegmatis* yet.

In the present study, we explored temporal gene expression profiles, identified 2139 TSSs throughout the genome, depicted 2233 operon and/or sub-operon structures, revealed a high proportion of leaderless mRNAs that were further verified them by Edman degradation, corrected structural annotation of 124 genes, and screened eight highly active promoters by combining growth-phase-dependent RNA-seq with related experiments in mc^{2}155. Together, we have comprehensively examined the expression profiles and transcriptional structures of mc^{2}155, which shall greatly facilitate further mycobacteria researches.

MATERIALS AND METHODS

Bacterial Strains and Growth Conditions

Strains used in this study are listed in **Table 1**. *Escherichia coli* strain DH5a was grown in lysogeny broth (LB) medium. mc^{2}155 wild type strain and its derivatives were grown at 37°C in Middlebrook 7H9 medium (Difco Becton Dickinson, USA) supplemented with 0.5% (v/v) glycerol and 0.05% (v/v) Tween 80, or on Middlebrook 7H10 agar (Difco Becton Dickinson, USA) supplemented with 0.5% (v/v) glycerol (Tang et al., 2014). When necessary, antibiotics were added to the culture containing either 50 μg/mL of kanamycin or 100 μg/mL of ampicillin.

RNA Isolation and RNA-Seq

mc^{2}155 cells equivalent to 30 OD_{600} (e.g., 30 mL of 1 OD_{600} of one culture) were harvested from each sample (three growth phases each in two biological replicates) and then quickly-frozen in liquid nitrogen. The bacteria pellet was transferred into a mortar containing liquid nitrogen, and was lysed by constant grinding (in the presence of liquid nitrogen), into a fine powder. The powder was transferred to a 2 mL tube containing 1 mL TRIzol Reagent (Invitrogen, CA, USA), and shaken vigorously, until the powder dissolved uniformly. Then, 200 μL of chloroform was added to the mixture and mixed vigorously for 15 s. The mixture was incubated on ice for 5 min, followed by centrifugation at 10,000 × g for 15 min at 4°C. Equal volumes (400 μL) of upper clear supernatant and isopropanol were mixed in a fresh 1.5 mL tube to precipitate RNA. It was mixed immediately by gently inverting 8∼10 times, and stored on ice for 15 min, followed by centrifugation at 10,000 × g for 30 min at 4°C. The upper clear phase was discarded, and 1 mL of 75% ethanol was added to wash the RNA pellet. The sample was gently washed by inverting 4–6 times, followed by centrifugation at 5,000 × g for 2 min at 4°C. Ethanol washing was repeated twice. Following the wash steps, ethanol was discarded. RNA was eventually dried by exposing to air for 5 min (air-dry) at room temperature. Finally, the RNA pellet was dissolved in 50 μL of RNase-free water, and stored at −80°C. Note that all reagents and materials used in this experiment were RNase-free.

For RNA-seq, 10 μg of total RNA from each sample was first treated with RNase-free DNase I (Takara, Japan) to prevent the potential contamination of genomic DNA. Ribosomal RNAs were removed using the RiboZero rRNA removal kit (Epicentre, USA) for gram-positive organisms prior to sequencing analysis.

TABLE 1 | Strains used in this study.

Strains	Characteristics or purposes	References
mc^2155	Wild-type *M. smegmatis* mc^2155	Li and He, 2012
mc^2155/pMV261-P_{MSMEG_0538}-*lacZ*	P_{MSMEG_0538} activity determination	This study
mc^2155/pMV261-P_{MSMEG_0559}-*lacZ*	P_{MSMEG_0559} activity determination	This study
mc^2155/pMV261-P_{MSMEG_0965}-*lacZ*	P_{MSMEG_0965} activity determination	This study
mc^2155/pMV261-P_{MSMEG_1060}-*lacZ*	P_{MSMEG_1060} activity determination	This study
mc^2155/pMV261-P_{MSMEG_1076}-*lacZ*	P_{MSMEG_1076} activity determination	This study
mc^2155/pMV261-P_{MSMEG_1771}-*lacZ*	P_{MSMEG_1771} activity determination	This study
mc^2155/pMV261-P_{MSMEG_2389}-*lacZ*	P_{MSMEG_2389} activity determination	This study
mc^2155/pMV261-P_{MSMEG_2756}-*lacZ*	P_{MSMEG_2756} activity determination	This study
mc^2155/pMV261-P_{MSMEG_3022}-*lacZ*	P_{MSMEG_3022} activity determination	This study
mc^2155/pMV261-P_{MSMEG_3050}-*lacZ*	P_{MSMEG_3050} activity determination	This study
mc^2155/pMV261-P_{MSMEG_3084}-*lacZ*	P_{MSMEG_3084} activity determination	This study
mc^2155/pMV261-P_{MSMEG_3255}-*lacZ*	P_{MSMEG_3255} activity determination	This study
mc^2155/pMV261-P_{MSMEG_3896}-*lacZ*	P_{MSMEG_3896} activity determination	This study
mc^2155/pMV261-P_{MSMEG_4326}-*lacZ*	P_{MSMEG_4326} activity determination	This study
mc^2155/pMV261-P_{MSMEG_4625}-*lacZ*	P_{MSMEG_4625} activity determination	This study
mc^2155/pMV261-P_{MSMEG_4891}-*lacZ*	P_{MSMEG_4891} activity determination	This study
mc^2155/pMV261-P_{MSMEG_4993}-*lacZ*	P_{MSMEG_4993} activity determination	This study
mc^2155/pMV261-P_{MSMEG_5081}-*lacZ*	P_{MSMEG_5081} activity determination	This study
mc^2155/pMV261-P_{MSMEG_5543}-*lacZ*	P_{MSMEG_5543} activity determination	This study
mc^2155/pMV261-P_{MSMEG_5789}-*lacZ*	P_{MSMEG_5789} activity determination	This study
mc^2155/pMV261-P_{MSMEG_6427}-*lacZ*	P_{MSMEG_6427} activity determination	This study
mc^2155/pMV261-P_{MSMEG_6467}-*lacZ*	P_{MSMEG_6467} activity determination	This study
mc^2155/pMV261-*MSMEG_4921*	Edman sequencing	This study
mc^2155/pMV261-*MSMEG_6422*	Edman sequencing	This study
Escherichia coli DH5α	Cloning host	Stored by our laboratory
DH5α/pMD19-*MSMEG_2196*	Identification the TSS of *MSMEG_2196*	This study
DH5α/pMD19-*MSMEG_3756*	Identification the TSS of *MSMEG_3756*	This study
DH5α/pMD19-*MSMEG_3757*	Identification the TSS of *MSMEG_3757*	This study
DH5α/pMD19-*MSMEG_4985*	Identification the TSS of *MSMEG_4985*	This study
DH5α/pMD19-*MSMEG_5223*	Identification the TSS of *MSMEG_5223*	This study
DH5α/pMD19-*MSMEG_6235*	Identification the TSS of *MSMEG_6235*	This study
DH5α/pMD19-*MSMEG_6917*	Identification the TSS of *MSMEG_6917*	This study
DH5α/pMD19-*MSMEG_6920*	Identification the TSS of *MSMEG_6920*	This study

Hundered nanograms of rRNA-depleted RNA from each sample was fragmented into 200–300 nts and used as a template for randomly primed PCR. Strand-specific cDNA libraries were prepared by standard techniques for subsequent Illumina sequencing using the mRNA-seq Sample Prep kit (Illumina, USA). The resulting cDNAs were sequenced on an Illumina HiSeq 2500 (Shao et al., 2015).

RNA-Seq Data Analysis

The quality of the raw sequence data was first assessed by FastQC, the sequences further filtered and trimmed by Trimmomatic (Bolger et al., 2014). Clean reads were then used to map to the reference genome of mc^2155 (NC_008596.1) by Bowtie2 (Langmead and Salzberg, 2012). BEDTools (Quinlan and Hall, 2010) analysis was performed on the alignments generated by Bowtie2 in order to quantify the mapped reads. To facilitate the comparison of expression levels of different samples, reads for CDSs were normalized with the total read count per sample and presented in the form of Reads Per Kilobase Per Million Reads (RPKM). The quantity of differential expressions of all transcripts was obtained using DEGseq (Wang et al., 2010) to include fold change values. In context, DEGs are those with two-fold changes or more, with FDR (false discovery rate) <0.001. As for single-nucleotide resolution transcriptome map, cleaned reads were mapped to mc^2155 genome by BLASTN to get their genome location. Then Perl script was used to count the expression level of each nucleotide. Data were visualized using Artemis software.

TSSs Validation by 5′-RACE

To verify the TSSs identified by RNA-seq, 5′-RACE was performed using a tobacco acid pyrophosphatase (TAP)-based 5′-RACE kit according to manufacturer's instructions as previously described (Ali et al., 2017). The selected regions were amplified from 5′-RACE cDNA library using adapter primers (outer and inner primers) and gene specific primers (Table S1). After multiplex amplification, PCR products were purified and ligated into pMD19-T vector, and then transformed into *E. coli*

DH5α (**Table 1**). Clones were sequenced and the first nucleotides next to 5′-RACE adaptor were considered to be the TSSs.

Edman Degradation to Identify N-Terminal Amino Acids

N-terminal sequencing was performed on proteins translated from leaderless mRNAs to identify their translation initiation amino acid residues. To get the test proteins, leaderless genes fused with 6 × His tag were over-expressed under the control of their own promoters based on the expression vector pMV261, and then, respectively, transformed into mc^2155 strain to obtain a series of derivative strains. The native promoter could hardly be induced, so two highly expressed leaderless genes *MSMEG_4921* and *MSMEG_6422* were selected. For protein purification, the two mc^2155 derivatives were harvested at mid-exponential phase, followed by cell lysis using French press. The clear lysate was loaded onto a Ni^{2+}-NTA affinity column for protein purification. Purified proteins were applied on a 15% SDS-PAGE and then transferred to PVDF blotting membrane for Edman sequencing using PPSQ-31A protein sequencer (Shimadzu, Japan). The strains and primers used are listed in **Table 1** and **Table S1**, respectively.

Operon Prediction

Operons prediction was based on the transcriptome map, visualized in Artemis. The prediction was carried out according to previously described method (Wang et al., 2013b). Briefly, genes in an operon share the same orientation, and transcripts should cover the intergenic region of the two adjacent genes. In addition, the ratio of the two gene expression levels should be <2. According to these two principles, we manually inspected the transcription of all 6,947 genes across the whole genome.

β-Galactosidase Assay

β-galactosidase activity was performed in mc^2155 by translational/transcriptional fusion of promoter to *lacZ* gene based on the expression vector pMV261. For translational fusion, the promoter regions, 5′-UTRs and partial N-terminal amino acids were fused with *lacZ* gene, while promoter and 5′-UTRs only were included in transcriptional fusion. P*hsp60* promoter was used as control when assessing the relative β-galactosidase activity of selected promoters (Yang et al., 2012). Reporter plasmids were transformed into wild type strain to obtain corresponding recombinant strains. All strains were grown in 7H9 medium at 37°C. β-galactosidase activity was determined as previously described (Miller, 1972). The strains and primers used are listed in **Table 1** and **Table S1**, respectively.

RESULTS

Global View of RNA-Seq Data at Different Growth Phases

The growth curve of mc^2155 in 7H9 broth is a prerequisite for its transcriptome profile which comprises a series of growth-phase-dependent RNA-seq experiments. Results indicate that mc^2155 reached mid-exponential phase at 16 h; early-stationary phase at 26 h and mid-stationary phase at 39 h (**Figure S1**). Two biological replicates were harvested from each of these three time points, and their total RNAs were extracted for subsequent experiments. All of the six RNA samples obtained from three different growth phases (two biological replicates) were found to exhibit high integrity with RNA Integrity Numbers (RINs) more than 9.5. Strand-specific cDNA libraries were then constructed and sequenced on an Illumina HiSeq 2000 platform. After removing low quality reads, 10 million clean reads were obtained for each library. The complete genome and trancriptome information of mc^2155 are shown in **Figure 1**.

Gene Expression Profiles during Different Growth Phases

In order to temporally investigate distinct biological processes, we analyzed gene expression profiles of the growth-phase-dependent RNA-seq data (**Table S2**). Comparison of RPKM values for each growth phase revealed significantly lower transcriptional levels for mid-stationary phase genes than those in the other two growth phases (Wilcoxon test, $p < 0.001$; **Figure 2A**). The differences in gene expression levels among the three growth phases were also analyzed with respect to different Cluster of Orthologous Groups of proteins (COG). DEGs were analyzed using mid-exponential phase RNA-seq data as control. In the mid-stationary phase, DEGs (especially down-regulated genes) are obviously increased in some clusters when compared to early-stationary phase (**Figures 2B,C**). We propose that in a cluster when the number of down regulated genes is far more than up-regulated genes, the cluster would be considered as down-regulated and *vice versa*. Thus, in the early-stationary phase (26 h), "carbohydrate transport and metabolism [G]" cluster was up-regulated, whereas "coenzyme transport and metabolism [H]", "lipid transport and metabolism [I]" and "translation, ribosomal structure, and biogenesis [J]" clusters were down-regulated (**Figure 2B**). In the mid-stationary phase, "nucleotide transport and metabolism [F]", "coenzyme transport and metabolism [H]", "translation, ribosomal structure and biogenesis [J]", and "replication, recombination and repair [L]" clusters were down-regulated; however, no cluster was up-regulated (**Figure 2C**).

Furthermore, DEGs were subjected to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways analysis to find out temporally associated biological functions. As shown in **Figure 3**, pathways with member genes significantly regulated (Q-value < 0.05) were marked in bold and underlined. As mentioned above, transcriptional levels of most genes were substantially reduced in mid-stationary phase; however, most of the genes in "inositol phosphate metabolism", "degradation of aromatic compounds" and "oxidative phosphorylation" pathways were up-regulated (**Figure 3A** and **Figure S2**). In contrast, the majority of genes were significantly down-regulated, especially genes in "translation" (including "ribosome" and "aminoacyl-tRNA biosynthesis" pathways) process (**Figure 3B** and **Figure S2**).

Genome-Wide Identification of TSSs

Transcriptional start sites (TSSs) typically refer to the 5′-ends of primary transcripts, but not 5′-ends of the processed transcripts. In our transcriptome data, most of the 5′-ends

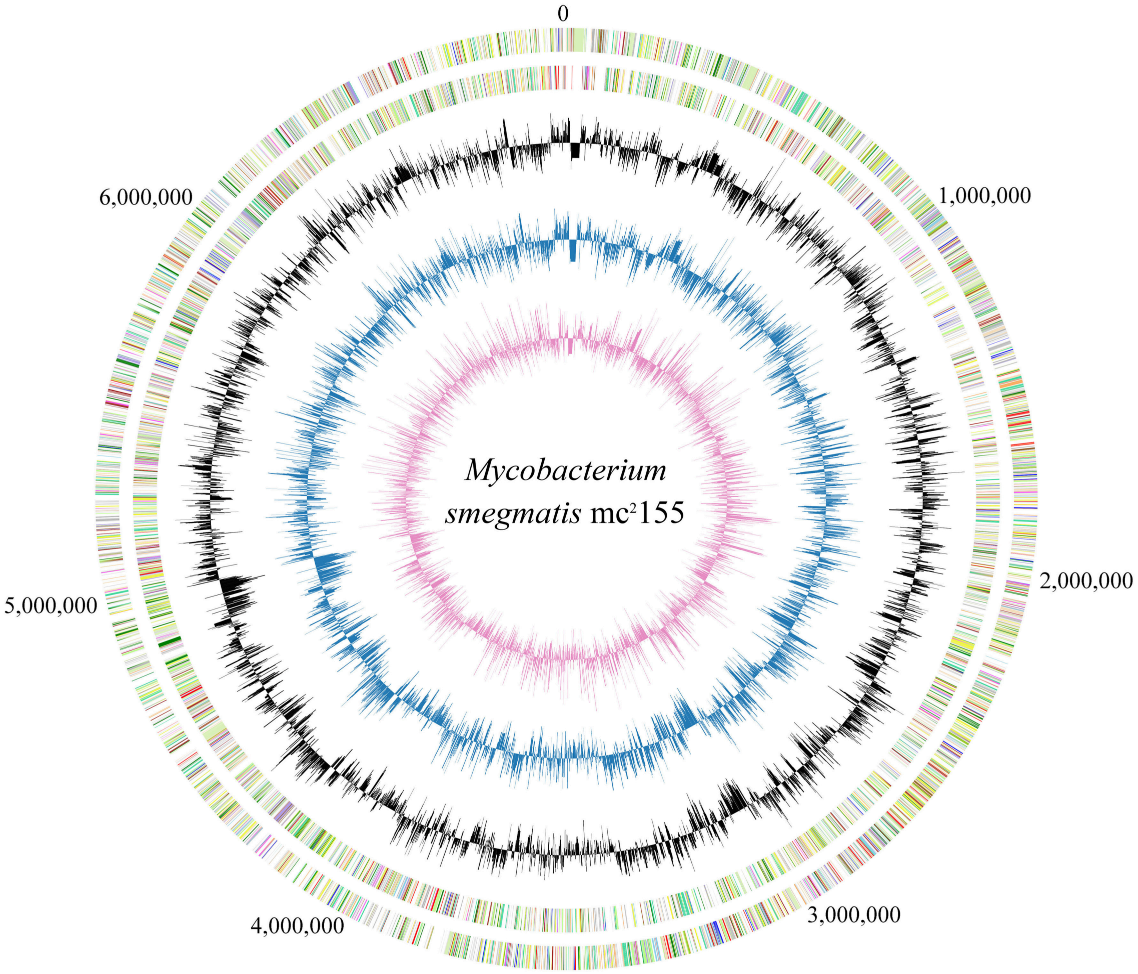

FIGURE 1 | Circular map of *M. smegmatis* mc^{2}155 genome and the corresponding transcriptome. Numbers outside circle show the genome coordinate. Moving inward, the subsequent two rings show CDSs in forward and reverse strands, respectively, with colors representing different COG categories. The inner three rings were colored in black, blue and purple representing the transcriptome maps at 16, 26, and 39 h, respectively; the uneven lines above and below the middle circle lines represent the expression level greater or lower than average.

of transcripts (except for those of processed transcripts) are indeed their TSSs, thus we used the term "TSSs" in the following text. To locate TSSs in genome-wide scale, clean reads were mapped to mc^{2}155 genome to generate single-nucleotide resolution transcriptome maps. However, the uneven transcriptome map caused by sequencing makes it hard to identify all the TSSs, especially internal TSSs which could typically imply the existence of small RNAs. Therefore, to improve the accuracy and reliability of our study, we only focused on the TSSs found at the beginning of mRNA transcripts. We manually inspected the transcriptome maps separately and successively. As shown in **Figure 4A**, the whole view of strand-specific transcriptional maps of *MSMEG_3068-3078* gene cluster was uneven, which is common in RNA-seq caused by sequencing bias. The first nucleotide significantly enriched at the beginning of transcript is considered to be the TSS. Single-nucleotide resolution of transcriptional maps showed that the transcriptions of *MSMEG_3071* and *MSMEG_3070* genes were clearly enriched at TSSs (**Figures 4B,C**). In this way, we totally identified 2139 TSSs throughout the mc^{2}155 genome (**Table S3**). To further confirm TSSs identified by RNA-seq, eight transcripts were randomly selected and verified by the canonical 5′-RACE method (**Table S4**). The results showed that TSSs identified by RNA-seq and 5′-RACE experiments overlapped at an identical nucleotides, which validates the accuracy of RNA-seq based TSSs identification. In general, promoters are located immediately upstream of TSSs. Thus, sequences 50 nts upstream of identified TSSs were scanned for potential conserved promoter motif by MEME. As a result, a conserved −10 motif (TANNNT) was found ranging 7–12 bp upstream of the identified TSSs in 69.5% of the genes, but no conserved −35 motif could be identified. These findings are similar to a previous study in *M. tuberculosis* (Cortes et al., 2013).

MSMEG_2196 is a c-di-GMP synthetase encoding gene, the TSS of which was first determined in a previous study (Bharati et al., 2013). In our study, its TSS was detected by RNA-seq and

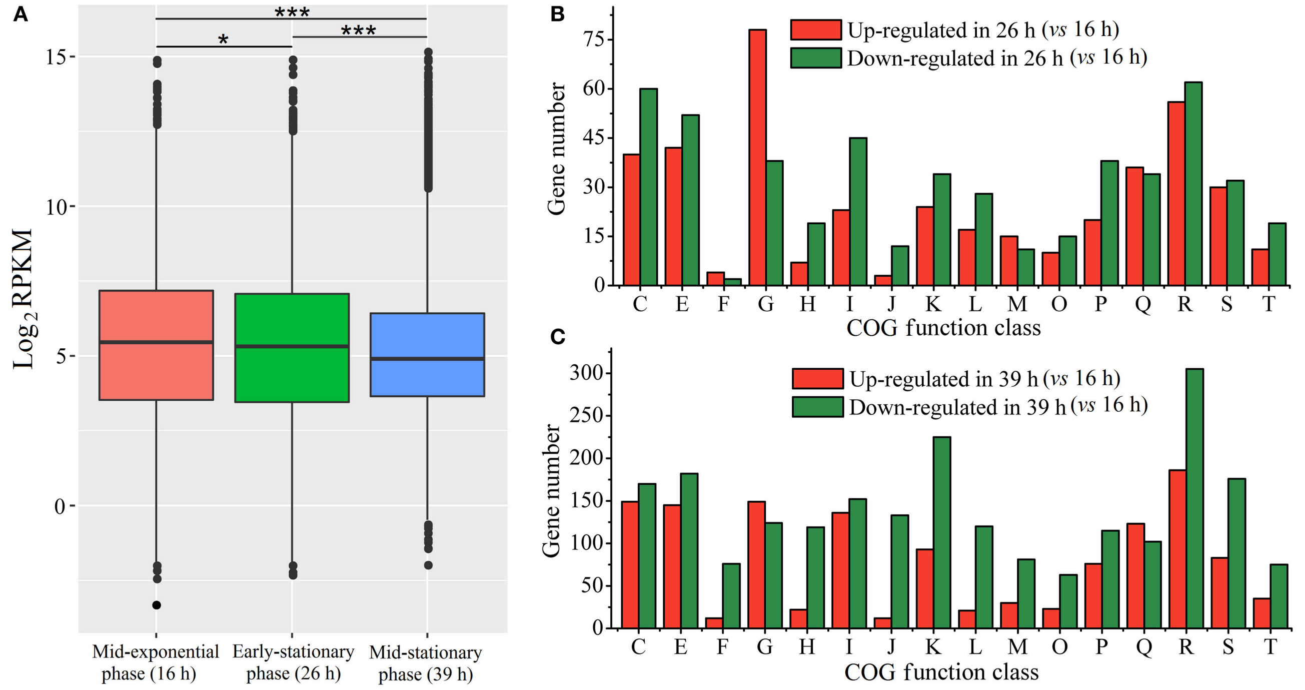

FIGURE 2 | Gene expression levels in RNA-seq data. **(A)** The median expression levels of different growth phases measured by RPKM for whole identified genes. RPKM values of each sample were analyzed using Wilcoxon test, **p* < 0.05, ***p* < 0.01, ****p* < 0.001. **(B,C)** respectively indicate the number of DEGs in distinct COG functional categories at the early-stationary phase (26 h) and mid-stationary phase (39 h). The COG functional categories are as follows: C, energy production and conversion; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination and repair; M, cell wall/membrane/envelope biogenesis; O, posttranslational modification, protein turnover, chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms.

was validated by 5′-RACE (**Figure 5**). Interestingly, the reported TSS is 88 nts downstream of our identified location. Moreover, a conserved −10 motif could be found upstream of our identified TSS but not in the previous one, indicating that our study could provide more credible data for gene organization. Taken together, our systematic identification of TSSs in genome scale could facilitates the research of gene expression regulation in mc^{2}155.

High Proportion of Leaderless mRNA in mc^{2}155

Based on TSS identification, we systematically explored the 5′-UTR length throughout the transcriptome (**Figure 6A** and **Table S3**). As shown in **Figure 6B**, 40.5% of the 5′-UTRs were <10 nts in length, and were considered to be leaderless mRNA; 28.6% of transcripts have 5′-UTR length from 11 to 50 nts, and 14.8% from 51 to 100 nts. In addition, 9.1% of transcripts exhibited long 5′-UTRs (with length more than 100 nts). Previous study indicated that long 5′-UTRs might hide regulatory elements such as riboswitch and transcription-attenuation region, which would regulate the transcription of downstream genes in a more sophisticated way (Breaker, 2011; Wang et al., 2013a; Rosinski-Chupin et al., 2014; Dersch et al., 2017). It is worth mentioning that about 7% TSSs were identified downstream of the annotated start codon, which were rather unlikely and were therefore considered to be mis-annotated.

We have observed that ∼40.5% (866/2139) of the transcripts with annotated TSSs had a 5′-UTR <10 nts in length, likely generating leaderless mRNAs. Moreover, the TSSs of 35.48% (759/2139) mRNAs overlapped with the first nucleotide of annotated start codon (AUG/GUG). In bacteria, the Shine-Dalgarno (SD) sequence helps to recruit the ribosome to the mRNA to initiate translation. One would ponder how ribosome binds to mRNA to start the translation procedure without an SD sequence. Can these mRNA be efficiently translated? Do these widespread leaderless mRNAs use other specific translation mechanisms or are they simply mis-annotated? Therefore, it is important to confirm the translation of leaderless mRNA and identify the initial amino acid of the translated protein. Shell et al. (2015) tried to reveal the widespread translation of leaderless mRNA in mc^{2}155 by ribosome profiling; however, the result could only indicate that ribosome did bind to the leaderless mRNAs but could not directly verify the translation of leaderless mRNAs. Moreover, they had also performed a proteomics study (Orbitrap LC-MS) on *M. tuberculosis* H37Rv to confirm this feature. As a result, a large number of proteins translated from leaderless mRNAs were identified. However, the key point in the validation of leaderless feature is to determine whether the first N-terminal amino acid of the translated protein is in accordance with the initial nucleotides of leaderless mRNA. LC-MS cannot be used to efficiently identify the initial amino

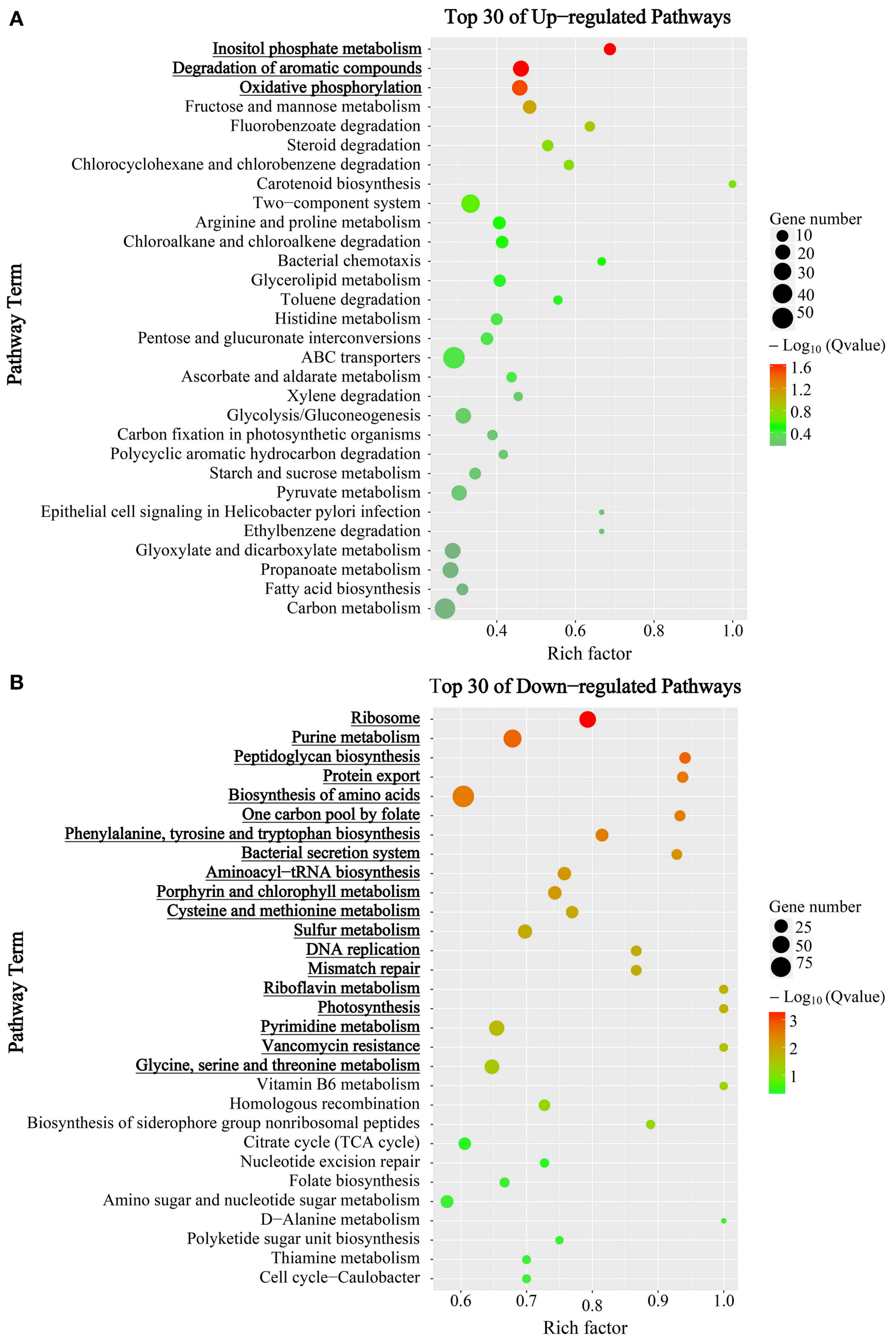

FIGURE 3 | KEGG enrichment analysis of DEGs between mid-exponential and mid-stationary phases. Y-axis label represents the distinct KEGG pathways, and X-axis label represents rich factor (rich factor = amount of DEGs in the pathway/amount of all genes in background gene set). The colors of the dots represent the Q-values of enrichment. Red color indicates high enrichment, while blue color indicates low enrichment. Pathway terms were sorted by Q-value in ascending order; and were marked in bold and underlined when Q-value < 0.05. The sizes of the dots represent the gene number of enrichment. **(A)** Top 30 up-regulated KEGG pathways. **(B)** Top 30 down-regulated KEGG pathways.

acid though some peptides detected in that study were identical to the annotated start codons. Thus, we performed Edman degradation, which is the most credible method in initial amino acid identification, on the MSMEG_4921 and MSMEG_6422 proteins both translated from leaderless mRNAs (**Figures 7A,B**). To get the test protein, leaderless genes fused with 6 × His tag were over-expressed under the control of their own promoters based on the expression vector pMV261, and were

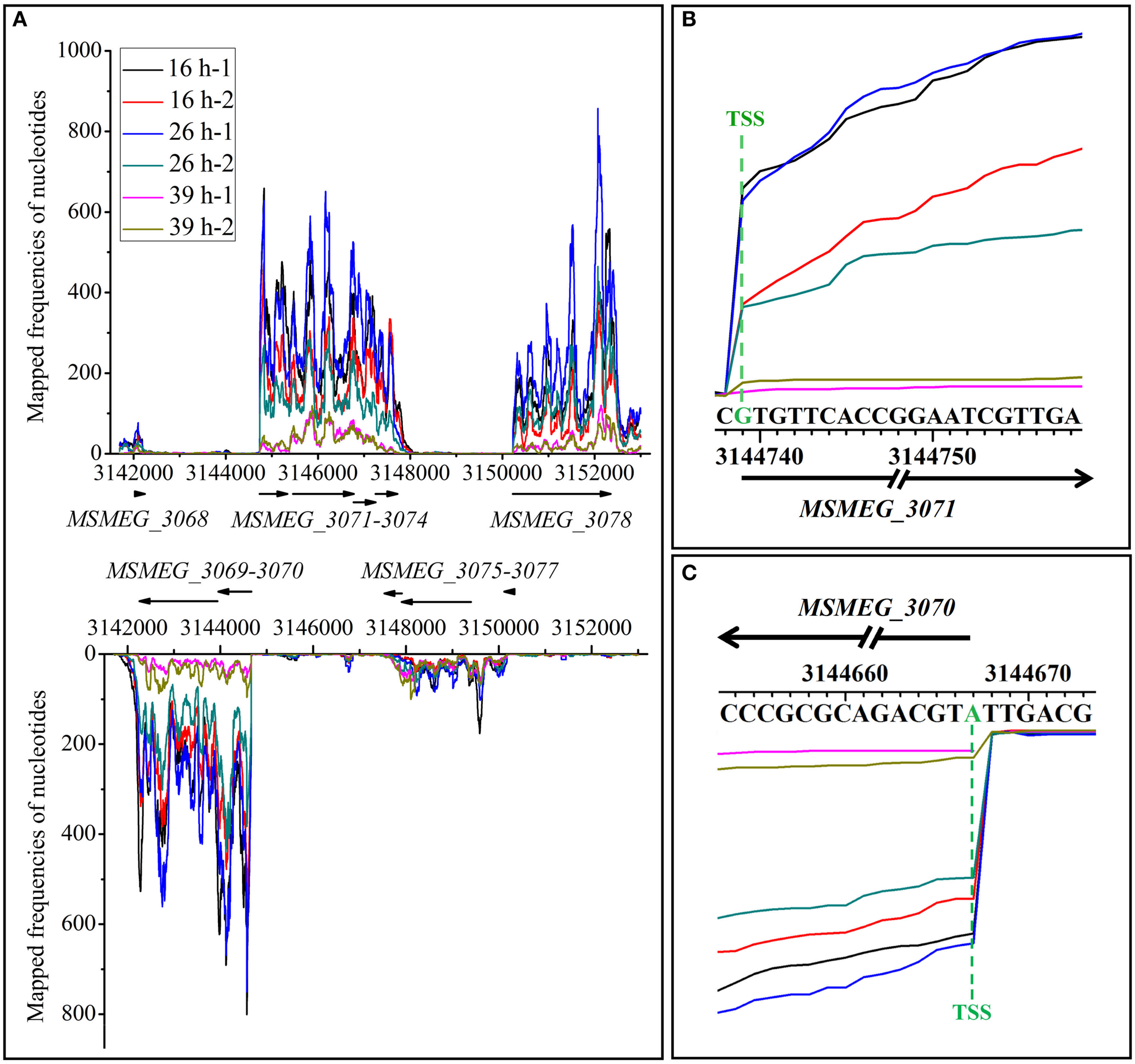

FIGURE 4 | Transcriptional maps of *MSMEG_3068-3078* loci. **(A)** Whole view of strand-specific transcriptional maps of *MSMEG_3068-3078* loci. *MSMEG_3068, MSMEG_3071-3074,* and *MSMEG_3078* genes residing in the positive strand while genes of *MSMEG_3069-3070* and *MSMEG_3075-3077* are located in the negative strand. **(B)** Single-nucleotide resolution transcriptional maps of *MSMEG_3071* and its TSS. **(C)** Single-nucleotide resolution transcriptional maps of *MSMEG_3070* and its TSS. Lines of different colors represent samples in different growth phases, −1 and −2 represent the two biological replicates. The same as below.

respectively transformed into mc^{2}155 strain to obtain a series of derivative strains. Purified proteins were applied on a 15% SDS-PAGE and then transferred to PVDF blotting membrane for Edman sequencing. The results indicated that their five N-terminal amino acids were identical to those annotated in the genome (**Figures 7C,D**). This strongly confirmed the leaderless feature of these leaderless mRNAs. Noteworthy, the N-terminal translation initiator Met was removed by methionine amino peptidase (MetAP), which is often crucial for the function and stability of proteins. Collectively, leaderless mRNA seems to be the common feature of mycobacteria. Further studies are required to understand how these leaderless mRNAs could be successfully translated into proteins. Several studies have shown that the unique translation can be initiated with high efficiency at leaderless transcripts via the undissociated 70S/80S ribosome and the initiator tRNA (Moll et al., 2004; Udagawa et al., 2004; Andreev et al., 2006; Giliberti et al., 2012). However, it is unclear whether mycobacterial strains employ this mechanism to initiate the translation of the high-proportion leaderless mRNA.

Correction of Mis-Annotated Genes in mc^{2}155

The majority of sequenced bacterial genomes were annotated by algorithm-based software, which can be error-prone. As shown in **Figure 6**, the lengths of 5′-UTRs of 150 transcripts were

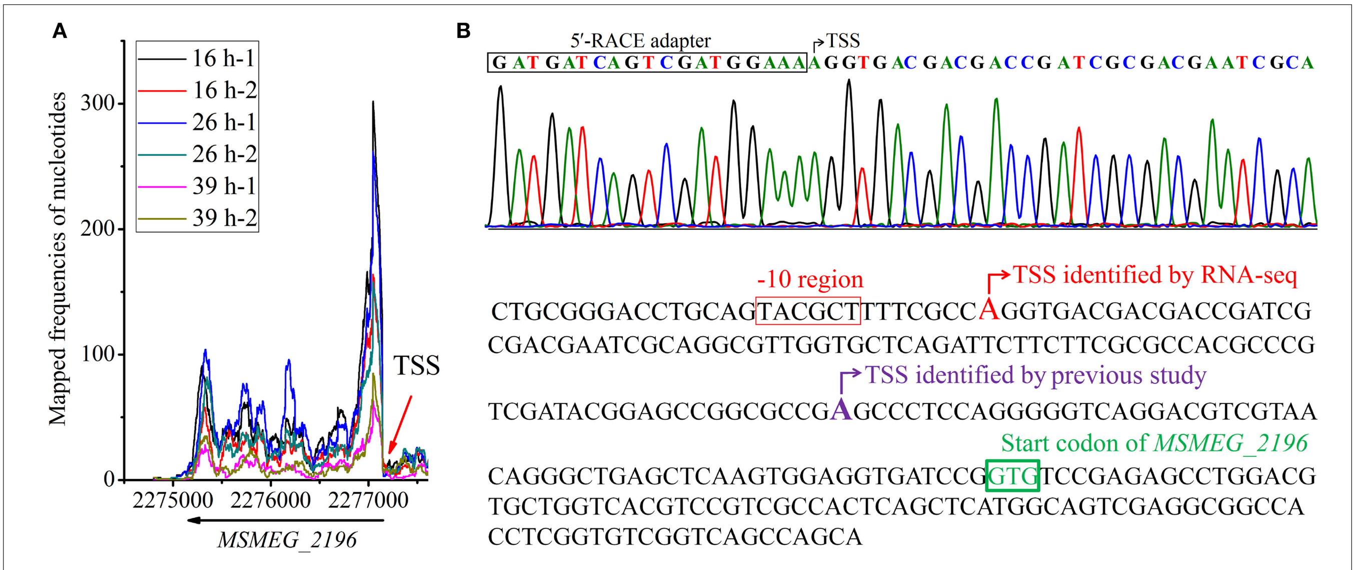

FIGURE 5 | Identification of *MSMEG_2196* TSS. **(A)** Whole view of transcriptional maps of *MSMEG_2196-2199* loci. **(B)** TSS of *MSMEG_2196* identified by 5′-RACE in this study. **(C)** Comparison of TSS of *MSMEG_2196* identified by our RNA-seq data with that previously reported (Bharati et al., 2013). The new identified TSS is colored in red, and the putative −10 motif is marked in rectangle. The previously identified TSS is colored in purple, however, no conserved −10 motif can be found.

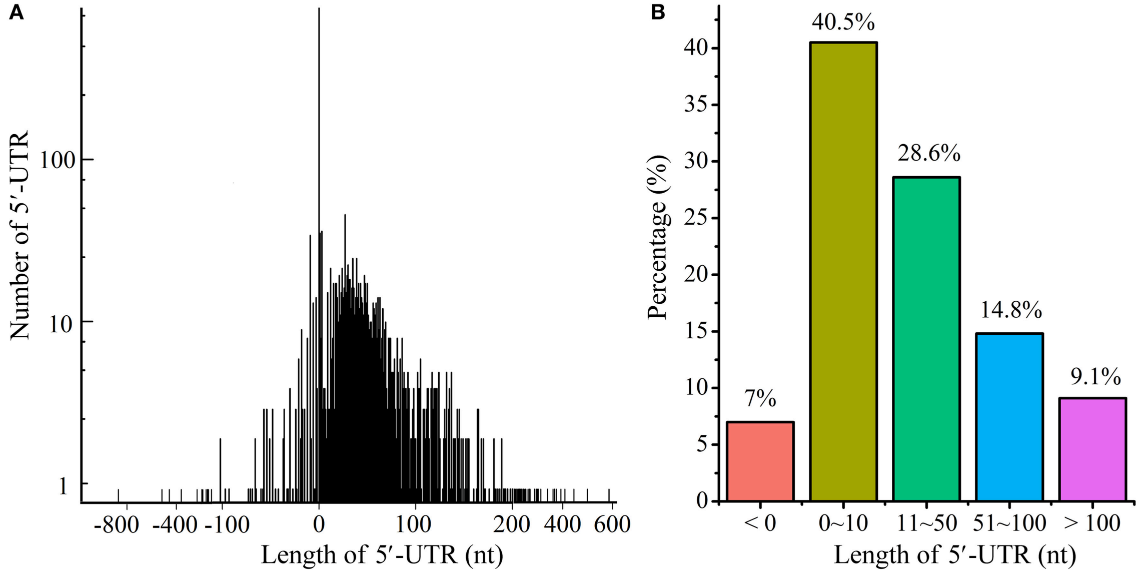

FIGURE 6 | Distribution of 5′-UTRs length. **(A)** The distribution of 5′-UTRs length. **(B)** The percentage of 5′-UTRs length. The length of 5′-UTRs <0 indicated potentially mis-annotated genes.

negative, indicating that these corresponding genes were very likely mis-annotated in these assay conditions (three different growth phases in 7H9 medium). Further explorations found that 124 out of these 150 transcripts used AUG/GUG as the first three nucleotides at TSS (**Table S5**), which were similar to the leaderless mRNAs. Given the fact that leaderless mRNAs are widespread in mc^{2}155, and an AUG/GUG at TSS is sufficient for the initial translation of leaderless mRNA in mc^{2}155 (Shell et al., 2015), we therefore propose that in these assay conditions, the 124 mis-annotated genes were transcribed into leaderless mRNA. Furthermore, distances between the identified TSSs and the first nucleotide of annotated start codons were in multiples of three in length, indicating that the corrected gene structural annotation does not shift the original open reading frame. **Figure 8** schematically shows the transcriptional maps of two mis-annotated proteins *MSMEG_1874* and *MSMEG_6901*, as well as the corrected versions. Together, we propose that RNA-seq could be highly useful to correct gene mis-annotation.

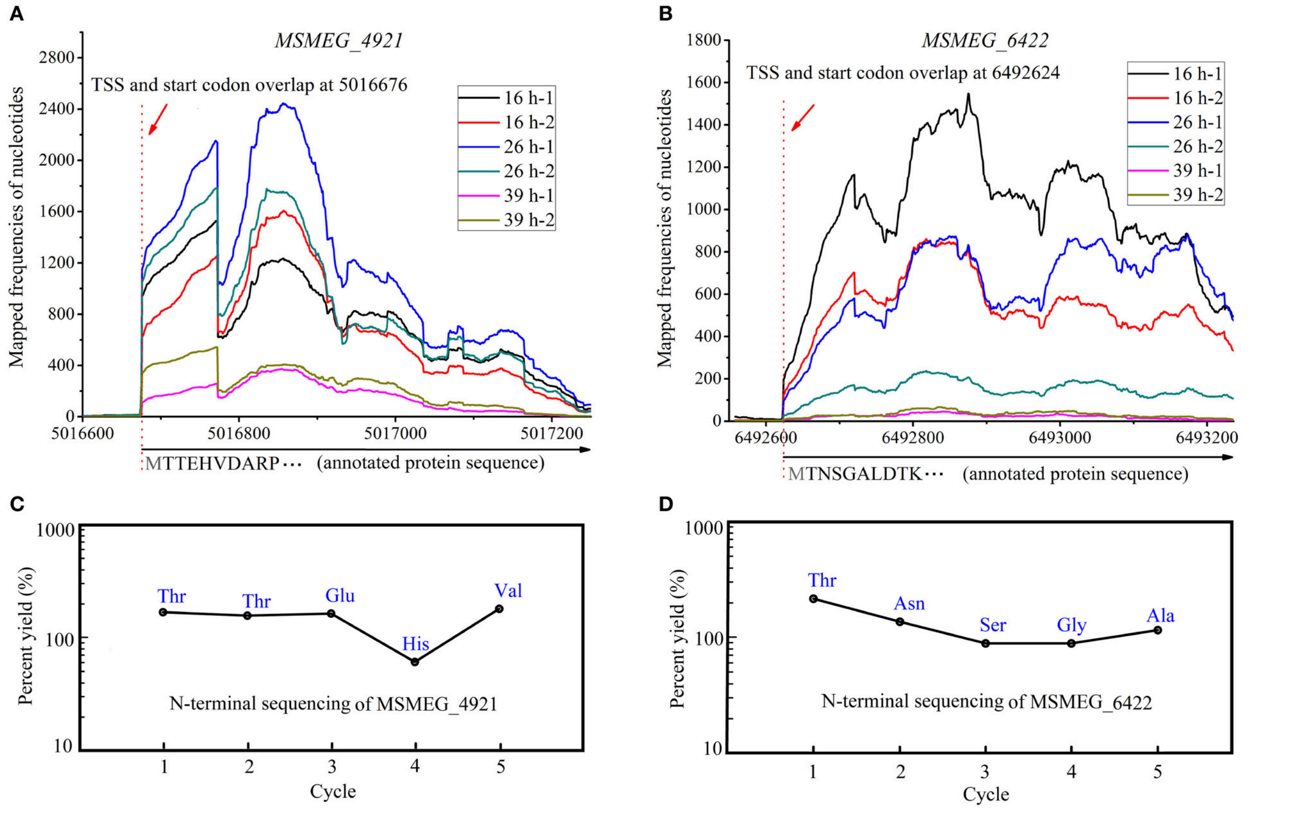

FIGURE 7 | N-terminal amino acids identification of MSMEG_4921 and MSMEG_6422. **(A)** and **(B)** respectively represent the transcriptional maps of *MSMEG_4921* and *MSMEG_6422*, while **(C)** and **(D)** respectively show the N-terminal five amino acids of MSMEG_4921 and MSMEG_6422 proteins. Both amino acid sequences are identical to the annotation files. Noteworthy, the N-terminal translation initiator Met was removed by methionine amino peptidase (MetAP), which is often crucial for the function and stability of proteins.

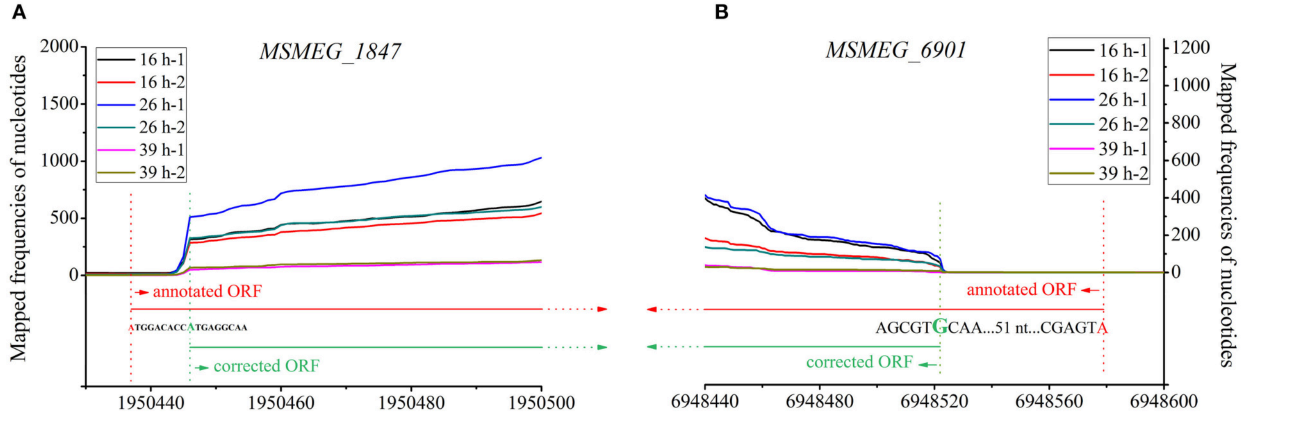

FIGURE 8 | Transcriptional maps and gene structural re-annotation of *MSMEG_1874* **(A)** and *MSMEG_6901* **(B)**. The nucleotides in red and red dot lines indicate the start codons annotated by algorithms-based software. The nucleotides in green and green dot lines indicate the TSS determined by RNA-seq, and these TSSs are also considered to be the accurate start codons of relevant genes.

Genome-Wide Identification of Operons and Sub-operons

Some genes in prokaryotes are organized into operons, which are then transcribed into polycistronic transcripts. It is typically considered that genes in an operon are functionally related. Zhang and He (2013) reported that overexpression of *MSMEG_6080* in mc^{2}155 led to cell expansion and aggregation; however, co-overexpression of *MSMEG_6080* and *MSMEG_6078*, located in a similar operon, alleviated the phenotypes described above instead. Furthermore, a gene

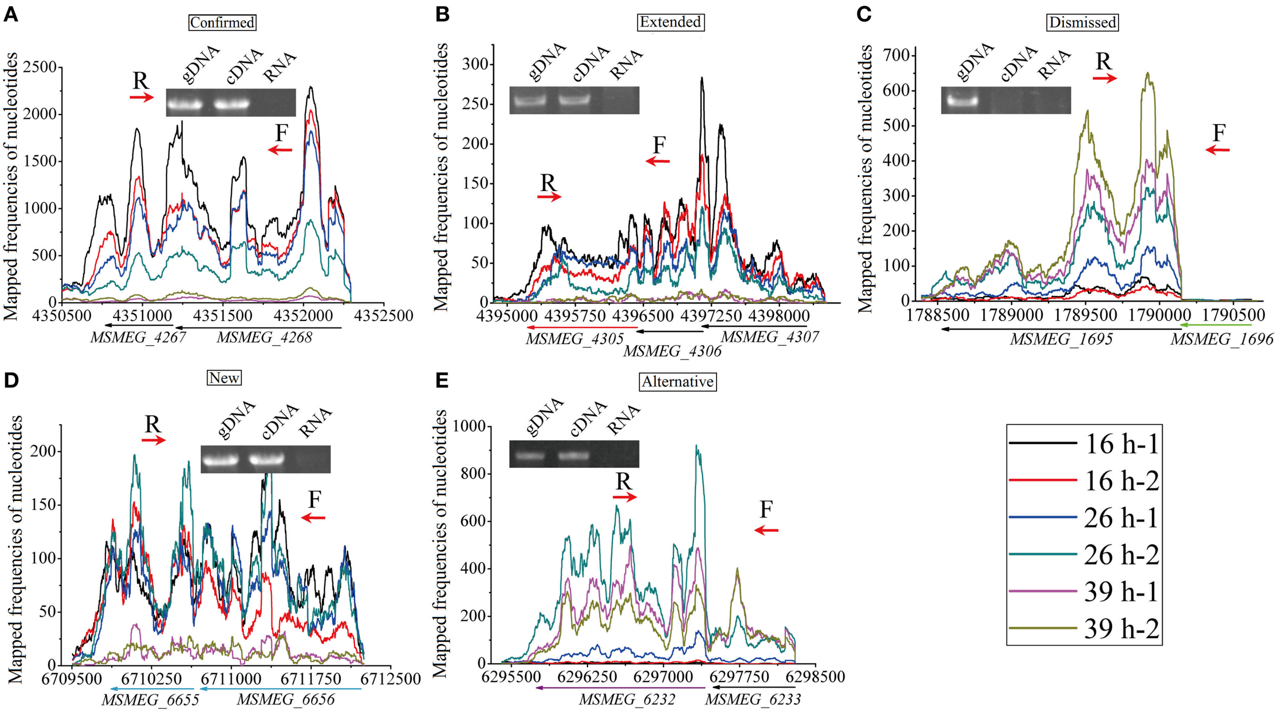

FIGURE 9 | Examples of the five groups of identified operon. The RT-PCR forward (F) and reverse (R) primers are indicated in red arrows. **(A)** Confirmed, DOOR annotated *MSMEG_4268* and *MSMEG_4267* as an operon, our RNA-seq data and RT-PCR experiment also indicate this. **(B)** Extended, DOOR annotated *MSMEG_4307* and *MSMEG_4306* as an operon, without *MSMEG_4305* (in red arrow); however, our RNA-seq data and RT-PCR experiment found that the transcription was extended to *MSMEG_4305*. **(C)** Dismissed, DOOR annotated *MSMEG_1696* to *MSMEG_1695* (in green arrow) as an operon, however, our RNA-seq data and RT-PCR experiment indicated that *MSMEG_1696* was dismissed from transcription. **(D)** New, our RNA-seq data identified new operons not found by DOOR, which was also indicated by RT-PCR. **(E)** Alternative, *MSMEG_6233* and *MSMEG_6232* (in purple arrow) were found to be co-transcribed; and *MSMEG_6232* seems to be alternatively transcribed from its own TSS.

in the middle of an operon, which can be independently transcribed in response to certain stimuli, is called sub-operon. Therefore, identification of genes that are grouped together into operons/sub-operons is important to studies of gene function and complex regulatory networks.

To investigate the structures of operons and sub-operons in genome scale, we manually inspected the transcription of all 6,947 genes across the genome, and finally identified 2,233 operons/sub-operons (**Table S6**). The 2,233 operons/sub-operons were further classified into five groups, according to the prediction of operons by DOOR (Mao et al., 2014). These five groups are: (1) confirmed, which means that our results were consistent with those predicted by DOOR (**Figure 9A**); (2) extended, in which we found that more genes could be transcribed into the polycistronic mRNA (**Figure 9B**); (3) dismissed, where less genes were found to be transcribed together (**Figure 9C**); (4) new, where some genes were transcribed together but were not annotated by DOOR (**Figure 9D**); and (5) alternative, indicating the existence of sub-operon (**Figure 9E**). To further illustrate the reliability of our results, several operons of each group were randomly selected for reverse transcription PCR (RT-PCR), and the results were consistent with RNA-seq data. The selected operons/sub-operons in each group is shown in **Table 2**.

Screening and Identification of Highly Active Promoter

As mentioned above, mc^2155 is widely accepted as a model organism in mycobacterial researches and has usually been used to overexpress homologous proteins from *M. tuberculosis* and *M. bovis* due to its non-pathogenic and fast-growing features (Soares et al., 2014; Deng et al., 2016). The temperature-sensitive P*hsp60* promoter could induce the expression of downstream genes on heat shock conditions, typically via a temperature shift from 37 to 45°C (Batoni et al., 1998). The inducibility makes P*hsp60* the most widely used promoter in mycobacteria. However, there are also some problems existing in the inducible expression system. Therefore, more promoters need to be discovered for driving gene expression in mycobacteria.

It is widely known that many non-coding RNAs such as rRNA, tRNA, 6S RNA, possess a highly active promoter. Here, we intended to screen alternate highly active promoters for gene expression. According to RNA-seq data, 22 gene candidates with high RPKM values were selected (**Table S7**). By fusing these promoter sequences to *lacZ* gene, we successfully certified that 8 out of 22 candidates exhibiting high β-galactosidase activities that are 3–5 times higher than the control promoter P*hsp60* (**Figure 10**). Additionally, some promoters showed temporal characteristics. As shown in **Figures 10C,D**, the promoter

TABLE 2 | Operon prediction in *M. smegmatis* mc^2155.

Operon groups	Total number	Number of genes selected for RT-PCR	Genes selected for RT-PCR[f]
Confirmed[a]	1635	7	*MSMEG_1556-1557,* *MSMEG_3255-3254,* *MSMEG_3540-3542,* *MSMEG_3569-3567,* *MSMEG_3655-3656,* *MSMEG_3689-3688,* *MSMEG_3935-3933*
Extended[b]	61	7	*MSMEG_1812-1811,* *MSMEG_1810-1809,* *MSMEG_3024-3025,* *MSMEG_3503-3504,* *MSMEG_3794-3793,* *MSMEG_3887-3996,* *MSMEG_4087-4085*
Dismissed[c]	167	6	*MSMEG_1525-1527,* *MSMEG_1873-1874,* *MSMEG_1951-1952,* *MSMEG_2279-2280,* *MSMEG_2356-2357,* *MSMEG_3104-3105*
New[d]	65	5	*MSMEG_1402-1403,* *MSMEG_3499-3450,* *MSMEG_3616-3615,* *MSMEG_3685-3686,* *MSMEG_3942-3941*
Alternative[e]	273	2	*MSMEG_3946-3945,* *MSMEG_6234-6535*

[a]*Confirmed: in an operon, RNA-seq annotated equal number of gene to DOOR;* [b]*Extended: in a operon, RNA-seq annotated more genes than DOOR;* [c]*Dismissed: in a operon, RNA-seq annotated less genes than DOOR;* [d]*New: this operon was not annotated by DOOR;* [e]*Alternative: sub-operon, a gene inside of an operon was transcribed independently;* [f]*of each groups, several genes were selected to perform RT-PCR.*

activities of P_{MSMEG_3050} and P_{MSMEG_3084} increased with time. In contrast, P_{MSMEG_4891} only exhibited high promoter activities in the early growth phase (**Figure 10E**). However, P_{MSMEG_5081} showed a relatively constant expression level during the whole detection period (**Figure 10F**). It is noteworthy that the β-galactosidase activity assays do not correlate well with RNA-seq data. This may be due to the fact that RNA-seq data only denote the transcriptional level. However, β-galactosidase activity assays reflect the expression levels of both transcription and translation. The disagreement between mRNA and protein levels usually happens, since many modifications may occur either at post-transcriptional or translational levels (McCarthy and Gualerzi, 1990; Alifano et al., 1994; Kozak, 2005).

DISCUSSION

This study combined RNA-seq with experimental evidences to generate a global overview of the transcriptional landscape of mc^2155, including genome-wide annotations of the transcriptional structures, as well as a bunch of non-canonical TSSs, 5′-UTRs and operons/sub-operons. For the first time, we have also performed Edman degradation to validate the translation of leaderless mRNA. In addition, eight highly active promoters were discovered for gene overexpression in mycobacteria.

In order to reveal the comprehensive transcriptome of mc^2155, we carried out RNA-seq experiments at three different growth phases. As expected, most of genes showed a lower expression level in mid-stationary phase, especially genes related to replication, translation, and DNA repair biological processes. However, the majority of genes in “inositol phosphate metabolism” and “degradation of aromatic compounds” pathways were found to be significantly up-regulated in the mid-stationary phase. Being noteworthy is that as a soil bacterium, mc^2155 is similar to other soil mycobacterial isolates such as *Mycobacterium* KMS, MCS, and JLS that share organic compounds degradation capability. Almost all genes in the “degradation of aromatic compounds” pathway exhibit increasing expression level in a temporal order (**Figure 3A** and **Figure S2**). This might indicate increased utilization of aromatic compounds as carbon source in mc^2155 during the stationary phase, demonstrating the potential of mc^2155 as an efficient pollutant eliminator.

Compared to other bacteria, mc^2155 seems to contain a strikingly high proportion (46.06%) of leaderless mRNA. Furthermore, for 40.99% of the TSSs identified in our study, the TSS overlapped with the first nucleotide of the start codon. Previously, Cortes et al. (2013) reported that 26% of TSSs overlapped with annotated start codon in *M. tuberculosis* H37Rv. Transcriptomics research in *M. avium* also revealed a high proportion (33%) of leaderless mRNAs (Ignatov et al., 2013). Furthermore, the translations of leaderless mRNAs were confirmed by Edman degradation in this study. Therefore, the high proportion of leaderless mRNAs may be a common feature of mycobacteria. In fact, Zheng et al. (2011) provided some insights into genes encoding leaderless mRNA in prokaryotes by a bioinformatics method, revealing that leaderless mRNA are widely spread in a variety of bacteria, especially in phyla *Actinobacteria* and *Deinococcus-Thermus*. Moreover, previous experimental studies have also revealed a high proportion of leaderless mRNA in some prokaryotes especially in Archaea (Torarinsson et al., 2005; Babski et al., 2016; Bauer et al., 2017).

It has long been known that the 5′-UTR of bacterial mRNAs include specific sequence elements for guiding ribosome binding, i.e., the SD motifs that interact directly with complementary motifs (anti-SD motifs) in the 16S rRNA. However, the initiation of leaderless mRNA translation is poorly understood. Currently, there are two possible pathways for initiation of leaderless mRNA translation in bacteria (verified in *E. coli*). One is mediated by the formation of the 30S–fMet-$tRNA_f^{Met}$-IF2 complex (Grill et al., 2000). With the help of initiation factor IF2, Met-$tRNA^{Met}$ can bind to the ribosome 30S subunit, and the ternary complex (30S–fMet-$tRNA_f^{Met}$-IF2) then recognizes the start codon of leaderless mRNA to initiate translation. The other mechanism is mediated by 70S ribosomes (Moll et al., 2004; Udagawa et al., 2004). The non-dissociated 70S ribosome first combines with fMet-tRNA to recognize and bind the start codon (AUG or

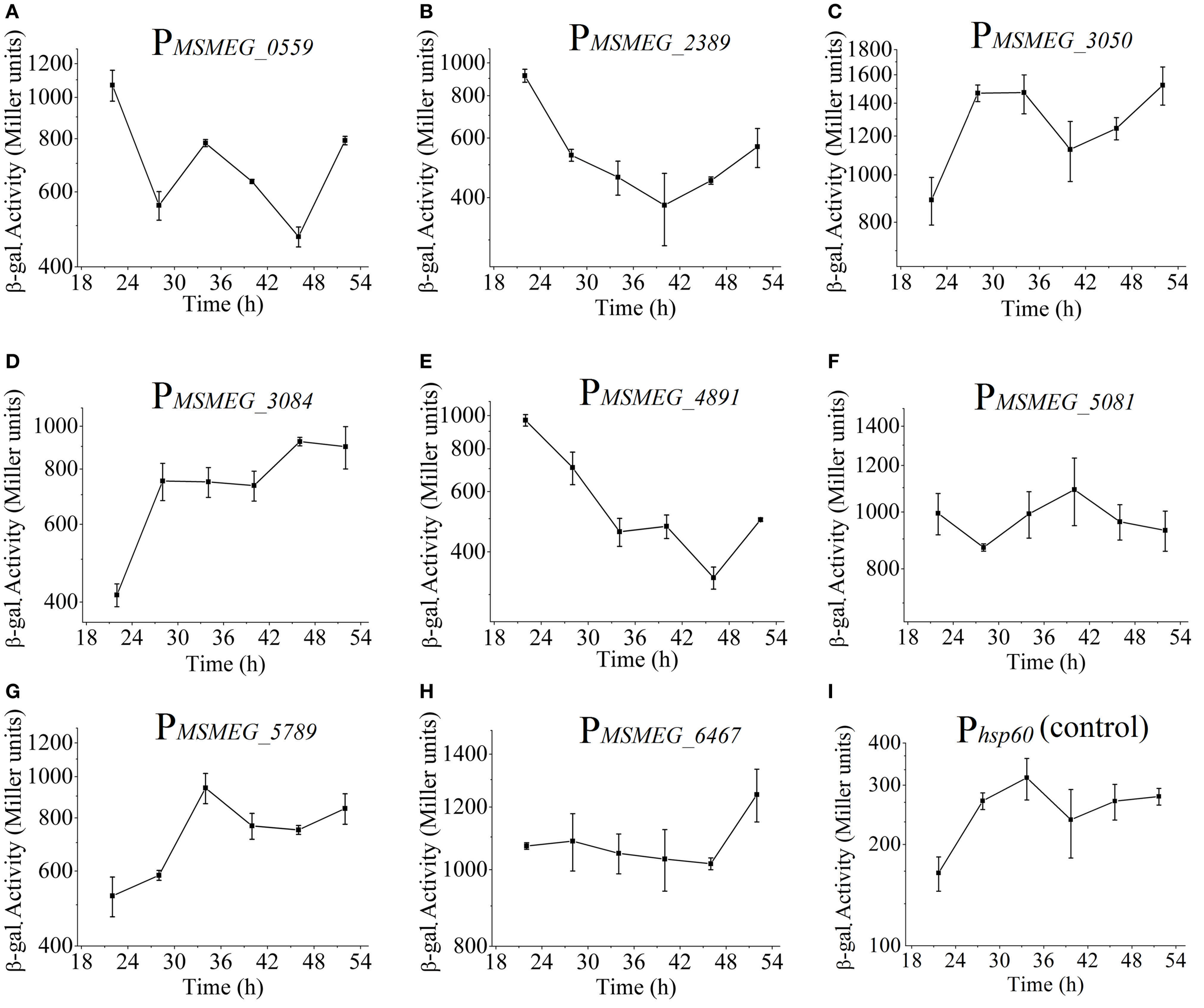

FIGURE 10 | Screening and identification of highly active promoter. P_{MSMEG_0559} represents mc^{2}155/pMV261-P_{MSMEG_0559}-*lacZ* strain, and the same as below. β-galactosidase activities for the nine strains (*Phsp60* as a control) were shown in **(A–I)**. Data represent the averages of biological triplicates. Error bars indicate standard deviation.

GUG) of leaderless mRNA to initiate translation. Importantly, both mechanisms have been found in *E. coli* that possesses a low proportion of leaderless mRNA with weak translation capability. However, robust translation of leaderless mRNA was validated by β-galactosidase reporter system in mycobacteria (Shell et al., 2015). Moreover, by analyzing the proteome data of mc^{2}155 (Chopra et al., 2014), we found that the levels of protein translated from mRNAs with or without leader sequence were almost equal. Thus, leaderless mRNAs seem to be rather abundant and are robustly translated in mycobacteria when compared to *E. coli*. Yet, the two leaderless mRNA translation pathways currently present in *E. coli* (possessing low proportion of leaderless mRNA and weak translation capability) may not be sufficient to explain the existence of high proportion of leaderless mRNA and the strong translation capability in mycobacteria. There may exist a plethora of uncharacterized mechanisms for ribosomal recognition and translational initiation of leaderless mRNAs in mycobacteria.

Bharati et al. (2013) reported that *MSMEG_2196*, a diguanylate cyclase encoded gene, could either be co-transcribed into a polycistronic mRNA (*MSMEG_2199-MSMEG_2196*) or independently as a monocistronic mRNA. Similarly, the independent transcription of CH1330 as a malate dehydrogenase gene in an operon in *Bacillus thurgiensis* CT-43 has been demonstrated (Wang et al., 2013b). It is now commonly believed that sub-operons are regulated by specific transcriptional factors and sigma factors in response to different stresses. Meanwhile, this could also be an efficient strategy for prokaryotes since they are usually composed of relatively smaller genomes. As mentioned above, we classified the identified operons into five groups, which comprise 273 sub-operons. We suppose that some of the sub-operons found here could exhibit alternate functions in mc^2155, which require further efforts to investigate.

Being widely used as a model organism in mycobacterial researches, mc^2155 should be capable of expressing and over-expressing genes from many other homologous mycobacteria. Currently, the P*hsp60* promoter is the only one that is widely used for expression of mycobacterial genes because of its heat-induced characteristic. However, there are some problems existing in this inducible expression system. Firstly, proteins produced at high temperature usually fail to fold into native structures and result in a mis-folded or inactive state. Moreover, high temperature induction may exhibit a disadvantage in introducing DNA mutations, which may disable or alter target gene expression (Al-Zarouni and Dale, 2002; Buddle et al., 2002). Additionally, there are also serious doubts about the stability of plasmids carrying the P*hsp60* promoter (Haeseleer, 1994; Al-Zarouni and Dale, 2002). Therefore, our discovery of eight highly active and temporarily-expressed promoters could contribute significantly to the gene expression in mc^2155.

Noteworthy, we found that some promoters exhibited great differences in β-galactosidase activities when fused translationally to *lacZ* gene compared to those fused transcriptionally. Translational fusion of some promoters exhibited β-galactosidase activity about 5–10 times higher than that of transcriptional fusion. For translational fusion, the promoter regions, 5′-UTRs and part of the N-terminal amino acids encoding regions were fused together to the *lacZ* gene. However, only promoter and 5′-UTRs were included in the transcriptional fusion. The CDS near the start codon may be important for translation of some genes.

DATA ACCESSION NUMBER

RNA-seq data have been submitted to GEO under the accession number GSE103158.

AUTHOR CONTRIBUTIONS

XL performed most of the experiments and made most of the data evaluation. HM completed most of the bioinformatics analysis. JH and HM gave the main idea of this research. FC, QT, ZY, and XC participated in partial experiments and interpretation of the data. XL, HM, and JH conceived the study and drafted the manuscript. JH, S-HC, and BA revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (grants 31770087 and 31270105), the Fundamental Research Funds for the Central Universities (grants 2662015PY175 and 2662017PY112), and the China Postdoctoral Science Foundation (grants 2015M582234 and 2017M622472).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02505/full#supplementary-material

Figure S1 | Growth curve of mc^2155. mc^2155 was cultured in 250 mL Erlenmeyer containing 100 mL 7H9 medium at 37°C, under a rotary shaker at 200 r/min. Overnight cultured mc^2155 was used to inoculate fresh media at a starting OD_{600} of 0.02, and OD_{600} was determined by sampling cultures in every 3 h. Samples for RNA-seq at 16, 26, and 39 h, corresponding to mid-exponential, early-stationary and mid-stationary phases respectively, were collected, with two biological replicates.

Figure S2 | Heatmap of all genes in the three pathways. Heatmap in each picture were measured by Log_2 (fold_change) of selected genes. **(A)** Degradation of aromatic compounds pathway. **(B)**Translation process. **(C)** TInositol phosphate metabolism.

Table S1 | Primers used in this study.

Table S2 | Gene expression profiles of *M. smegmatis* mc^2155.

Table S3 | Genome-wide identification of TSSs in *M. smegmatis* mc^2155.

Table S4 | Eight genes randomly selected to perform 5′-RACE experiments.

Table S5 | Structural re-annotation of genes in *M. smegamtis* mc^2155.

Table S6 | Genome-wide operon (sub-operon) identificaiton.

Table S7 | Expression profiles of genes selected to screen high activity promoters.

REFERENCES

Ali, M. K., Li, X., Tang, Q., Liu, X., Chen, F., Xiao, J., et al. (2017). Regulation of inducible potassium transporter KdpFABC by the KdpD/KdpE two-component system in *Mycobacterium smegmatis*. *Front. Microbiol.* 8:570. doi: 10.3389/fmicb.2017.00570

Alifano, P., Bruni, C. B., and Carlomagno, M. S. (1994). Control of mRNA processing and decay in prokaryotes. *Genetica* 94, 157–172. doi: 10.1007/BF01443430

Al-Zarouni, M., and Dale, J. W. (2002). Expression of foreign genes in *Mycobacterium bovis* BCG strains using different promoters reveals instability of the hsp60 promoter for expression of foreign genes in *Mycobacterium bovis* BCG strains. *Tuberculosis* 82, 283–291. doi: 10.1054/tube.2002.0374

Andreev, D. E., Terenin, I. M., Dunaevsky, Y. E., Dmitriev, S. E., and Shatsky, I. N. (2006). A leaderless mRNA can bind to mammalian 80S ribosomes and direct polypeptide synthesis in the absence of translation initiation factors. *Mol. Cell Biol.* 26, 3164–3169. doi: 10.1128/MCB.26.8.3164-3169.2006

Arnvig, K. B., Comas, I., Thomson, N. R., Houghton, J., Boshoff, H. I., Croucher, N. J., et al. (2011). Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. *PLoS Pathog.* 7:e1002342. doi: 10.1371/journal.ppat.1002342

Babski, J., Haas, K. A., Näther-Schindler, D., Pfeiffer, F., Förstner, K. U., Hammelmann, M., et al. (2016). Genome-wide identification of transcriptional start sites in the haloarchaeon *Haloferax volcanii* based on differential RNA-Seq (dRNA-Seq). *BMC Genomics* 17:629. doi: 10.1186/s12864-016-2920-y

Batoni, G., Maisetta, G., Florio, W., Freer, G., Campa, M., and Senesi, S. (1998). Analysis of the *Mycobacterium bovis* hsp60, promoter activity in recombinant *Mycobacterium avium*. *FEMS. Microbiol. Lett.* 169, 117–124. doi: 10.1111/j.1574-6968.1998.tb13307.x

Bauer, J. S., Fillinger, S., Förstner, K., Herbig, A., Jones, A. C., Flinspach, K., et al. (2017). dRNA-seq transcriptional profiling of the FK506 biosynthetic gene cluster in *Streptomyces tsukubaensis* NRRL18488 and general analysis of the transcriptome. *RNA Biol.* 30, 1–10. doi: 10.1080/15476286.2017.1341020

Bharati, B. K., Swetha, R. K., and Chatterji, D. (2013). Identification and characterization of starvation induced msdgc-1 promoter involved in the c-di-GMP turnover. *Gene* 528, 99–108. doi: 10.1016/j.gene.2013.07.043

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Breaker, R. R. (2011). Prospects for riboswitch discovery and analysis. *Mol. Cell.* 43, 867–879. doi: 10.1016/j.molcel.2011.08.024

Buddle, B. M., Wards, B. J., Aldwell, F. E., Collins, D. M., and Lisle, G. W. (2002). Influence of sensitisation to environmental mycobacteria on subsequent vaccination against bovine tuberculosis. *Vaccine* 20, 1126–1133. doi: 10.1016/S0264-410X(01)00436-4

Chopra, T., Hamelin, R., Armand, F., Chiappe, D., Moniatte, M., and McKinney, J. D. (2014). Quantitative mass spectrometry reveals plasticity of metabolic networks in *Mycobacterium smegmatis*. *Mol. Cell Proteomics* 13, 3014–3028. doi: 10.1074/mcp.M113.034082

Cordone, A., Audrain, B., Calabrese, I., Euphrasie, D., and Reyrat, J. M. (2011). Characterization of a *Mycobacterium smegmatis* uvrA mutant impaired in dormancy induced by hypoxia and low carbon concentration. *BMC Microbiol.* 11, 1–9. doi: 10.1186/1471-2180-11-231

Cortes, T., Schubert, O. T., Rose, G., Arnvig, K. B., Comas, I., Aebersold, R., et al. (2013). Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Rep*. 5, 1121–1131. doi: 10.1016/j.celrep.2013.10.031

Deng, G., Zhang, F., Yang, S., Kang, J., Sha, S., and Ma, Y. (2016). *Mycobacterium tuberculosis* Rv0431 expressed in *Mycobacterium smegmatis*, a potentially mannosylated protein, mediated the immune evasion of RAW 264.7 macrophages. *Microb. Pathog.* 100, 285–292. doi: 10.1016/j.micpath.2016.10.013

Deng, W., Zeng, J., Xiang, X., Li, P., and Xie, J. (2015). PE11 (Rv1169c) selectively alters fatty acid components of *Mycobacterium smegmatis* and host cell interleukin-6 level accompanied with cell death. *Front. Microbiol.* 6:613. doi: 10.3389/fmicb.2015.00613

Dersch, P., Khan, M. A., Mühlen, S., and Görke, B. (2017). Roles of regulatory RNAs for antibiotic resistance in bacteria and their potential value as novel drug targets. *Front. Microbiol.* 8:803. doi: 10.3389/fmicb.2017.00803

Fortino, V., Smolander, O. P., Auvinen, P., Tagliaferri, R., and Greco, D. (2014). Transcriptome dynamics-based operon prediction in prokaryotes. *BMC. Bioinformatics* 15:145. doi: 10.1186/1471-2105-15-145

Giliberti, J., O'Donnell, S., Etten, W. J., and Janssen, G. R. (2012). A 5′-terminal phosphate is required for stable ternary complex formation and translation of leaderless mRNA in *Escherichia coli. RNA* 18, 508–518. doi: 10.1261/rna.027698.111

Grill, S., Gualerzi, C. O., Londei, P., and Bläsi, U. (2000). Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation. *EMBO J.* 19, 4101–4110. doi: 10.1093/emboj/19.15.4101

Haeseleer, F. (1994). Structural instability of recombinant plasmids in mycobacteria. *Res. Microbiol.* 145, 683–687. doi: 10.1016/0923-2508(94) 90040-X

Hendrickson, E. L., Beck, D. A., Miller, D. P., Wang, Q., Whiteley, M., Lamont, R. J., et al. (2017). Insights into dynamic polymicrobial synergy revealed by time-coursed RNA-seq. *Front. Microbiol.* 8:261. doi: 10.3389/fmicb.2017.00261

Hillion, M., Bernhardt, J., Busche, T., Rossius, M., Maaß, S., Becher, D., et al. (2017). Monitoring global protein thiol-oxidation and protein S-mycothiolation in *Mycobacterium smegmatis* under hypochlorite stress. *Sci. Rep*. 7:1195. doi: 10.1038/s41598-017-01179-4

Ignatov, D., Malakho, S., Majorov, K., Skvortsov, T., Apt, A., and Azhikina, T. (2013). RNA-seq analysis of *Mycobacterium avium* non-coding transcriptome. *PLoS ONE* 8:e74209. doi: 10.1371/journal.pone.0074209

Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* 361, 13–37. doi: 10.1016/j.gene.2005.06.037

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, Q., Ge, F., Tan, Y., Zhang, G., and Li, W. (2016). Genome-wide transcriptome profiling of *Mycobacterium smegmatis* mc2155 cultivated in minimal media supplemented with cholesterol, androstenedione or glycerol. *Int. J. Mol. Sci.* 17:689. doi: 10.3390/ijms17050689

Li, W., and He, Z. G. (2012). LtmA, a novel cyclic di-GMP-responsive activator, broadly regulates the expression of lipid transport and metabolism genes in *Mycobacterium smegmatis*. *Nucleic Acids Res.* 40, 11292–11307. doi: 10.1093/nar/gks923

Liao, Y., Huang, L., Wang, B., Zhou, F., and Pan, L. (2015). The global transcriptional landscape of *Bacillus amyloliquefaciens* XH7 and high-throughput screening of strong promoters based on RNA-seq data. *Gene* 571, 252–262. doi: 10.1016/j.gene.2015.06.066

Liu, Y., Wang, H., Cui, T., Zhou, X., Jia, Y., Zhang, H., et al. (2016). NapM, a new nucleoid-associated protein, broadly regulates gene expression and affects mycobacterial resistance to anti-tuberculosis drugs. *Mol. Microbiol.* 101, 167–181. doi: 10.1111/mmi.13383

Mao, X., Ma, Q., Zhou, C., Chen, X., Zhang, H., Yang, J., et al. (2014). DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res.* 42, D654–D659. doi: 10.1093/nar/gkt1048

McCarthy, J. E., and Gualerzi, C. (1990). Translational control of prokaryotic gene expression. *Trends Genet.* 6, 78–85. doi: 10.1016/0168-9525(90) 90098-Q

Miller, J. H. (1972). *Experiments in Molecular Genetics,* Vol. 58. New York, NY: Cold Spring Harbor Laboratory Press, 893–924.

Moll, I., Hirokawa, G., Kiel, M. C., Kaji, A., and Bläsi, U. (2004). Translation initiation with 70S ribosomes: an alternative pathway for leaderless mRNAs. *Nucleic Acids Res.* 32, 3354–3363. doi: 10.1093/nar/gkh663

Morbidoni, H. R., Vilchèze, C., and Kremer, L. (2006). Dual inhibition of mycobacterial fatty acid biosynthesis and degradation by 2-alkynoic acids. *Chem. Biol.* 13, 297–307. doi: 10.1016/j.chembiol.2006.01.005

Perkins, T. T., Kingsley, R. A., Fookes, M. C., Gardner, P. P., James, K. D., Yu, L., et al. (2009). A strand-specific RNA-seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet*. 5:e1000569. doi: 10.1371/journal.pgen.1000569

Petridis, M., Benjak, A., and Cook, G. M. (2015). Defining the nitrogen regulated transcriptome of *Mycobacterium smegmatis* using continuous culture. *BMC Genomics* 16:821. doi: 10.1186/s12864-015-2051-x

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rosinski-Chupin, I., Soutourina, O., and Martin-Verstraete, I. (2014). Riboswitch discovery by combining RNA-seq and genome-wide identification of transcriptional start sites. *Methods. Enzymol*. 549, 3–27. doi: 10.1016/B978-0-12-801122-5.00001-5

Shao, Z. H., Ren, S. X., Liu, X. Q., Xu, J., Yan, H., Zhao, G. P., et al. (2015). A preliminary study of the mechanism of nitrate-stimulated remarkable increase of rifamycin production in Amycolatopsis mediterranei U32 by RNA-seq. *Microb. Cell Fact*. 14:75. doi: 10.1186/s12934-015-0264-y

Sharma, C. M., and Vogel, J. (2014). Differential RNA-seq: the approach behind and the biological insight gained. *Curr. Opin. Microbiol*. 19, 97–105. doi: 10.1016/j.mib.2014.06.010

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., et al. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* 464, 250–255. doi: 10.1038/nature08756

Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., et al. (2015). Leaderless transcripts and small proteins are common features of the mycobacterial translational landscape. *PLoS Genet*. 11:e1005641. doi: 10.1371/journal.pgen.1005641

Soares, A., Rizzi, C., Peiter, A. C., Labonde, J., and Dellagostin, O. A. (2014). Expression of recombinant *Mycobacterium bovis*, antigen 85B by *Mycobacterium smegmatis* mc2155. *BMC. Proc*. 8:P169. doi: 10.1186/1753-6561-8-S4-P169

Tang, Q., Li, X., Zou, T., Zhang, H., Wang, Y., Gao, R., et al. (2014). *Mycobacterium smegmatis* BioQ defines a new regulatory network for biotin metabolism. *Mol. Microbiol.* 94, 1006–1023. doi: 10.1111/mmi.12817

Thomason, M. K., Bischler, T., Eisenbart, S. K., Förstner, K. U., Zhang, A., Herbig, A., et al. (2015). Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. *J. Bacteriol.* 197, 18–28. doi: 10.1128/JB.02096-14

Torarinsson, E., Klenk, H. P., and Garrett, R. A. (2005). Divergent transcriptional and translational signals in Archaea. *Environ. Microbiol.* 7, 47–54. doi: 10.1111/j.1462-2920.2004.00674.x

Udagawa, T., Shimizu, Y., and Ueda, T. (2004). Evidence for the translation initiation of leaderless mRNAs by the intact 70S ribosome without its dissociation into subunits in eubacteria. *J. Biol. Chem.* 279, 8539–8546. doi: 10.1074/jbc.M308784200

Wang, J., Ai, X., Mei, H., Fu, Y., Chen, B., Yu, Z., et al. (2013a). High-throughput identification of promoters and screening of highly active promoter-5′-UTR DNA region with different characteristics from *Bacillus thuringiensis*. *PLoS ONE* 8:e62960. doi: 10.1371/journal.pone.0062960

Wang, J., Han, M., Zheng, C., Qian, H., Cui, C., and Yang, F. (2013b). The metabolic regulation of sporulation and parasporal crystal formation in *Bacillus thuringiensis* revealed by transcriptomics and proteomics. *Mol. Cell Proteomics* 12, 1363–1376. doi: 10.1074/mcp.M112.023986

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26, 136–138. doi: 10.1093/bioinformatics/btp612

Wang, S., Dong, X., Zhu, Y., Wang, C., Sun, G., Luo, T., et al. (2013). Revealing of *Mycobacterium marinum* transcriptome by RNA-seq. *PLoS ONE* 8:e75828. doi: 10.1371/journal.pone.0075828

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

World Health Organization (WHO) (2016). Global Tuberculosis Report. Available online at: www.who.int/tb/publications/global_report/en/ (Accessed October 13, 2016).

Wu, M. L., Gengenbacher, M., Chung, J. C., Chen, S. L., Mollenkopf, H. J., Kaufmann, S. H., et al. (2016). Developmental transcriptome of resting cell formation in *Mycobacterium smegmatis*. *BMC Genomics* 17:837. doi: 10.1186/s12864-016-3190-4

Wurtzel, O., Sapra, R., Chen, F., Zhu, Y., Simmons, B. A., and Sorek, R. (2010). A single-base resolution map of an archaeal transcriptome. *Genome Res.* 20, 133–141. doi: 10.1101/gr.100396.109

Yang, M., Gao, C., Cui, T., An, J., and He, Z. G. (2012). A TetR-like regulator broadly affects the expressions of diverse genes in *Mycobacterium smegmatis*. *Nucleic Acids Res.* 40, 1009–1020. doi: 10.1093/nar/gkr830

Zhang, L., and He, Z. G. (2013). Radiation-sensitive gene A (RadA) targets DisA, DNA integrity scanning protein A, to negatively affect cyclic di-AMP synthesis activity in *Mycobacterium smegmatis*. *J. Biol. Chem.* 288, 22426–22436. doi: 10.1074/jbc.M113.464883

Zheng, X., Hu, G. Q., She, Z. S., and Zhu, H., (2011). Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics* 12:361. doi: 10.1186/1471-2164-12-361

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.