RNA-Seq Analysis of *Mycobacterium avium* Non-Coding Transcriptome

Dmitriy Ignatov¹, Sofia Malakho²,³, Konstantin Majorov⁴, Timofey Skvortsov¹, Alexander Apt⁴, Tatyana Azhikina¹

¹ Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia, ² Center of Innovations and Technologies “Biologically Active Compounds and their Applications”, Moscow, Russia, ³ Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia, ⁴ Central Institute for Tuberculosis, Moscow, Russia

**Abstract**

Deep sequencing was implemented to study the transcriptional landscape of *Mycobacterium avium*. High-resolution transcriptome analysis identified the transcription start points for 652 genes. One third of these genes represented leaderless transcripts, whereas the rest of the transcripts had 5' UTRs with the mean length of 83 nt. In addition, the 5' UTRs of 6 genes contained SAM-IV and Yokk types of riboswitches. 87 antisense RNAs and 10 intergenic small RNAs were mapped. 6 intergenic small RNAs, including 4.5S RNA and mpB, were transcribed at extremely high levels. Although several intergenic sRNAs are conserved in *M. avium* and *M. tuberculosis*, both of these species have unique intergenic sRNAs. Moreover, we demonstrated that even conserved small RNAs are regulated differently in these species. Different sets of intergenic sRNAs may underlie differences in physiology between conditionally pathogenic *M. avium* and highly specialized pathogen *M. tuberculosis*.

---

**Introduction**

Infections caused by mycobacteria, other than *Mycobacterium tuberculosis* and *M. leprae*, pose a serious medical problem. At least 60 species of environmental mycobacteria are known to cause opportunistic infections in humans [1], and *M. avium* is of special interest because of its environmental diversity and ability to infect birds, animals and immune compromised humans. According to the current taxonomy, *M. avium* includes 4 subspecies, *M. avium* avium (MAA), *M. avium* hominisuis (MAH), *M. avium* silvaticum (MAS), and *M. avium* paratuberculosis (MAP) [2,3]. MAA and MAS are specific avian pathogens causing a tuberculosis-like disease in birds [4], whereas MAP is a well-known pathogen causing Johne’s disease, i.e., chronic enteritis of ruminants [5]. MAH causes disseminated infections in patients with severely compromised T cell immunity, and pulmonary infections in children and aged individuals [6,7,8,9]. After entering the host, *M. avium* is engulfed by macrophages and resides within phagosomes, demonstrating capacity to block phagosome-lysosome fusion and prevent acidification of the internal phagosomal compartment, similarly to *M. tuberculosis* [10]. Although in humans major features of acquired immunity against *M. tuberculosis* and *M. avium* are similar, *M. tuberculosis* is much more aggressive and virulent, whereas *M. avium* affects only immune compromised individuals [11].

In the past decade, studies have revealed abundant non-coding transcriptomes in bacteria. Non-coding transcripts, including 5'UTRs, antisense transcripts, and intergenic small RNAs, primarily play roles in the regulation of gene expression. Pathogenic bacteria encounter diverse environmental conditions and therefore require fast regulatory circuits to survive. Together with regulatory proteins and 2-component systems, regulatory RNAs allow pathogens to adjust their metabolism and express factors that are necessary for subverting host defense [12]. 5' UTRs in bacteria are of varying lengths reaching several hundred nucleotides. They can alter their conformation in response to changes in temperature, pH, or metabolite concentrations. Changes in conformation subsequently regulate the expression of downstream genes by transcription termination/antitermination or blocking/unblocking of ribosome-binding sites [13]. Intergenic sRNAs, which are relatively short transcripts (~50–300 nucleotides), act on distantly encoded targets. Most of them regulate mRNAs through short, imperfect base-pairing interactions, whereas others, e.g., *Escherichia coli* 6S and CsrB RNAs, modify protein activities [14]. Antisense transcription may be generated by either bona fide antisense RNAs (asRNAs), which do not encode proteins, or by overlapping parts of mRNAs [15]. It has been reported that antisense transcription may regulate the transcription and translation of coding transcripts. An attractive model for the activity of antisense transcripts is the digestion of antisense-coding RNA duplexes by Rnase III [23].

Investigations of mycobacterial non-coding transcriptomes have focused mainly on *M. tuberculosis*. The first mycobacterial small RNAs were mapped in *M. tuberculosis* by cloning and sequencing a fraction of short RNAs [16], as well as in *M. bovis* BCG by using a combination of computer prediction and cloning strategies [17]. Recently, bacterial RNA-seq has emerged as an accurate tool for studying bacterial
transcriptomes [21]. For instance, Pellin et al. performed high-throughput sequencing of sRNA fractions of \textit{M. tuberculosis} and identified \~2,000 sRNA candidates by combining information regarding the read coverage with conservation analysis of intergenic regions [18]. A subsequent microarray analysis further confirmed the expression of 258 of these sRNA candidates, including 22 intergenic sRNAs and 152 antisense RNAs [19].

Arnvig et al. utilized RNA-seq for \textit{M. tuberculosis} whole-transcriptome profiling at the exponential and stationary phases of growth and demonstrated abundant anti-sense transcription generated by overlapping 3’ UTRs and distinct asRNAs [20]. Overall, in \textit{M. tuberculosis}, more than 20 intergenic sRNAs were identified and mapped (16,17,20,21), and for a few of these sRNAs, implications in pathogenesis have been suggested. The levels of MTS194, MTS479, and MTS2822 were found to be increased in response to H2O2 mimicking oxidative stress inside host macrophages. Accumulation of MTS997, MTS1338, and MTS2823 at high levels was demonstrated during the transition from exponential to stationary phases of growth and along the course of infection [17,21]. There is little doubt that intergenic sRNAs play an important role in mycobacterial physiology; however, the mechanism of their action remains unknown. Thus, \textit{Mycobacteria} lack the Hfq protein facilitating interactions between intergenic sRNAs and their targets in gram-negative microorganisms [22].

Until recently, no data were published on non-coding transcriptome of \textit{M. avium}, prompting us to fill this gap in our knowledge. By cloning and sequencing a fraction of short RNAs, we identified several trans-encoded sRNAs in \textit{M. avium} subsp. \textit{avium} TMC724 (ATCC25291) [23]. In the present study, we utilized a more powerful RNA-seq approach to profile the transcriptome of \textit{M. avium} TMC724 in the mid-logarithmic growth phase. The single-nucleotide resolution data enabled us to identify multiple non-coding transcripts and exact transcript boundaries. In addition, 2 intergenic sRNAs with the highest in \textit{vitro} expression levels were evaluated in the mouse infection model by means of qRT-PCR, and demonstrated profound expression differences in the lung tissue of genetically susceptible and resistant mice.

**Results and Discussion**

The total amount of intergenic transcription exceeds the transcription of protein-coding sequences in logarithmically grown \textit{M. avium}

We used Illumina sequencing to study the transcriptome of \textit{M. avium} TMC724. RNA was isolated from bacteria grown to the mid-exponential growth phase. 16S and 23S rRNA were removed with MicrobExpress kit (Life Technologies). To prepare cDNA for sequencing we utilized the Illumina RNA ligation protocol, which allows not only investigation of bacterial transcriptome in an unbiased manner but also preservation of information regarding the transcription directionality [24]. This protocol have shown a high level of technical reproducibility [25]. At present, complete sequencing of \textit{M. avium} TMC724 genome is not finished yet. GenBank contains 258 contigs of the genome that were produced in a whole genome shotgun sequencing project under the accession number CP000479.1, since the genome sequence of the latter was completed and annotated. Our sequencing approach provided 42.2 million reads, 65–76 nucleotides (nt) in length. The mapping statistics is shown in Table 1. Totally, 14 million reads did not map to the \textit{M. avium} MAH104 genome, whereas 18.8 million reads mapped to a single rRNA operon. Such a large number of unmapped reads (33%) may be explained by technical artifacts of sequencing.

In addition, while 3.6 million reads mapped to coding sequences (CDS) in the sense orientation, only 0.2 million reads mapped to CDS in the antisense orientation (5.4% of that mapping in sense). In contrast, 4.4 million reads, mapping to intergenic loci, constituted 123% of that mapping to CDS in sense (Table 1). Below, we show that such an enormous level of intergenic transcription is caused by several highly expressed small RNAs. Arnvig et al. have shown that in log-phase grown \textit{M. tuberculosis}, reads mapping antisense to CDS constituted 16.1% of those mapping to CDS in the sense orientation, whereas those mapping to IGR constituted 23.4% of those mapping to CDS in the sense orientation [20]. Discrepancy in the amount of reads mapping antisense to CDS or to intergenic loci between our studies may be explained by different techniques of sample preparation for sequencing or by true differences in physiology between these microorganisms.

For rapid and convenient visualization of RNA-seq data, we produced the so-called transcriptional profile. For each nucleotide of the genome, we counted the number of reads overlapping the nucleotide. Such transcriptional profiles were made for forward and reverse strands of the genome and visualized in the Artemis genome browser [26].

**The genome of \textit{M. avium} hominisuis strain 104 possesses 25 large sequence polymorphisms, absent in \textit{M. avium avium} TMC724.** The major source of \textit{M. avium} strain diversity is known to be their genomic heterogeneity, mainly large sequence polymorphisms (LSPs), which are genomic regions present in some strains but not others [27]. To identify \textit{M. avium} 104 genomic loci that are potentially absent in \textit{M. avium} TMC724,

| Table 1. Mapping statistics. |
|-----------------------------|
| RNA-seq results, mln of reads | % of total | % of CDS | % of CDS (\textit{M. tuberculosis}) |
| total mapped reads | 28,2 | 67 |
| reads mapped to rRNA operon | 18,8 | 67 |
| reads mapped to CDS | 3,6 | 13 |
| reads antisense to CDS | 0,2 | 0,7 | 5,4 | 16,1 |
| reads mapped to IGR | 4,4 | 16 | 123,1 | 23,4 |

Number of reads mapped to rRNA operon, CDS in sense and antisense orientation, intergenic regions (IGR) are shown in the first column. Percentages of reads mapped to these loci relative to all mapped reads are shown in the second column. In the third column percentages of reads mapped to CDS in antisense orientation and IGR relative to the number of reads mapped to CDS are represented. In the fourth column the same values for \textit{M. tuberculosis} are shown [20].

doi:10.1371/journal.pone.0074209.t001
we aligned 258 contigs of the *M. avium* TMC724 genome with the complete sequence of the *M. avium* 104 genome. As a result, 25 such loci with lengths of more than 500 nt were found (Table 2). We observed that almost no reads mapped to these loci and that each locus was located in the gaps of *M. avium* TMC724 contigs. On the basis of these findings, we consider these loci to be LSPs specific to *M. avium* 104, but not *M. avium* TMC724. Totally, these 25 loci make up 750 kb of genetic material and contain 808 genes. Because we mapped the reads on genome of *M. avium* TMC724, conditionally pathogenic *M. avium* TMC724 contains 2 additional clusters of *mce* genes, MAV\_2532-2537 and MAV\_5047-5051 located in LSPs 14 and 22, respectively. The *mce* operons have been implicated in the pathogenesis of mycobacteria, and their products were suggested to be transporter systems [29,30,31,32].

76% of *M. avium* genes are transcribed at a considerable level

RPKM, a measure of relative gene expression [33], normalizes the number of reads mapping to a gene to both the feature lengths and the total number of reads in an experiment. For the total number of reads in the experiment, we chose all mapped reads, except those mapping to the rRNA operon. We excluded these reads because of rRNA depletion, which may skew RPKM values. On the other hand, we included reads representing the non-coding transcriptome, because it includes important un-annotated features. Interestingly, the distribution of RPKM values for all genes was continuous; that is, despite significant variability in features. Interestingly, the distribution of RPKM values for all genes was continuous, that is, despite significant variability in RPKM values, almost all genes were transcribed, and there were only 71 genes with a RPKM value of 0. Among these, 53 encode...
transposases and 6 (MAV_0763, MAV_2077, MAV_2417, MAV_2855, MAV_4320, and MAV_4779) contain short ORFs encoding putative proteins. However, these proteins have no homologs based on a psi-blast search; therefore, these 6 non-transcribed genes are probably the results of misannotation. The RPKM values for all annotated CDSs present in the MAA TMC724 genome are given in Table S1.

Transcription initiation is a stochastic event; even repressed genes are transcribed in a certain population of cells. Nevertheless, there is a possibility that a low level of transcription detected for a subpopulation of genes results from technical artifacts of sample preparation or the mapping of reads. To account for possible technical noise, we defined a gene to be transcribed when its RPKM value is more than 7. In the absence of additional information (e.g., information regarding expressed proteins [25]), the choice of such a threshold is somewhat arbitrary. A RPKM value of 7 was chosen because genes that have lower RPKM tend to have higher transcription from the antisense strand than that from the sense strand. Accordingly, we consider that 3,408 of 4,501 genes were transcribed and represent 76% of the genes under consideration. Similar results were obtained for *M. tuberculosis*; 78% of annotated genes were transcribed [20].

To study the relative transcription of genes representing different functional categories, we chose 10% of genes with the highest RPKM values and studied which functional classes were overrepresented among these gene groups (Figure 1).

The functional classes for *M. avium* 104 were retrieved from www.TIGR.org. Their over- or underrepresentation among the 450 genes with the highest RPKM values was confirmed by a binomial test with correction for multiple comparisons. According to the statistical criterion, genes encoding products participating in the protein synthesis, transcription, and protein fate were significantly overrepresented, whereas genes coding for transport and binding proteins, as well as proteins with unknown functions and proteins participating in central intermediary metabolism, were significantly underrepresented. Overall, such a picture is typical for bacteria grown in the logarithmic phase, since frequent cell divisions require the machinery for transcription and synthesis of new proteins. In contrast, in nutrient-rich media, the expression of genes encoding transport and binding proteins and enzymes involved in central intermediary metabolism is not necessary.

**One third of the revealed transcriptional start points correspond to leaderless transcripts**

Visual inspection of the Artemis transcriptional profile led to the identification of abrupt increases in transcription levels, frequently located near the 5′ ends of annotated genes (Figure 2A). The positions of these spikes were considered as transcriptional start points (TSP). To map these points, we used a computer algorithm that accounted for the differences in transcription levels upstream and downstream of the spikes, the abruptness of increases, and the distance to the nearest upstream spike. All the mapped putative TSPs were manually verified. The spikes located upstream of the 5′ ends of annotated genes were considered to be their respective TSPs. In total, putative TSPs for 844 genes were mapped. Our method does not guarantee that all revealed TSPs are primary and not the products of degradation. To confirm their primary nature, we searched for –10 consensus sequences upstream of these TSPs using MEME software [34]. Such a consensus sequence was found to be located 5–8 nt upstream of TSPs of 652 genes and represents a hexamer, TA[GCA][GC][CG]T, with the first, second, and sixth nucleotide positions being the most conserved. Although many genes of mycobacteria are operated by sigma-factors alternative to SigA and therefore have another –10 consensus sequence, we discarded putative TSPs for genes that do not have properly positioned –10 sequences. To validate our method, 5 of 652 remained TSPs (MAV_1369, MAV_3090, MAV_4124, MAV_0525, and MAV_1669) were randomly selected and confirmed by RLM-RACE. TSPs revealed by inspecting the profile mapped ±3 nt from those revealed by RLM-RACE. A list of TSPs for 652 genes is summarized in Table S2. Interestingly, 33% of revealed TSPs mapped ±3 nt from the start codons of their respective genes; therefore, we consider that these genes are

---

**Figure 1. Functional classes of genes with the highest transcription.** Bars show the levels of over- or underrepresentation of each functional category among 450 genes with the highest expression. Functional categories significantly over- or underrepresented in this list of genes are indicated with asterisks.

doi:10.1371/journal.pone.0074209.g001
Leaderless genes are widespread in *Actinobacteria*, and constitute approximately 20% of genes in this phylum [35]. Leaderless mRNAs may have an important role in stress adaptation in bacteria [36], and their abundance in *M. avium* requires further studies.

We have shown that *M. avium* contains even higher proportion of leaderless genes, than the average for *Actinobacteria*. The 5′ UTRs of other mRNAs varied from 3 to 728 nt in length, with a mean value of 83 nt. These 5′ UTRs may contain sequences that play a role in the regulation of gene expression.

Ykok and SAM-IV riboswitches in *M. avium*

Visual inspection of long 5′ UTRs further identified sequences that possibly code for riboswitches, 4 of which matched to known riboswitch structures in the RFAM database (Table 3) [37]. The SAM-IV riboswitch was found upstream of MAV_4315. SAM-IV riboswitches specifically bind S-adenosylmethionine, a cofactor used in many methylation reactions [38], whereas MAV_4315 encodes O-acetylhomoserine sulfhydrylase, an enzyme implicated in cysteine and methionine metabolism.

Ykok leader sequences were identified upstream of 3 following genes: MAV_2915, encoding a proline-glutamate (PE) protein;
The presence of the Ykok leader reveals that this is not distinctly known, several of them are thought to be virulence or localized to the cell surface. While the function of these proteins between these genes suggests that they constitute a single operon. MAV_2912–MAV_2915 genes, which encode a PE protein (MAV_2915), 2 proline-proline-glutamate (PPE) proteins (MAV_2913 and MAV_2914), and a conserved hypothetical protein (MAV_2912). The close proximity between these genes suggests that they constitute a single operon. PE/PPE proteins, possessing conserved proline-glutamate or proline-proline-glutamate motifs at their N terminus, are secreted or localized to the cell surface. While the function of these proteins is not distinctly known, several of them are thought to be virulence factors [40]. The presence of the Ykok leader reveals that this operon is activated upon Mg\textsuperscript{2+} starvation, which could be observed inside phagosomes of macrophages. In M. tuberculosis, the Ykok leader has been found upstream of a putative operon encoding 4 PE/PPE genes, a conserved hypothetical protein, and a predicted magnesium transporter, MgtC [21]. These data indicate that the expression of PE/PPE proteins may be regulated by the presence of Mg\textsuperscript{2+}.

A putative riboswitch, which regulates the operon MAV_2907–MAV_2910, was revealed in the 5′ UTR of MAV_2910. Although its sequence did not match any known riboswitches in the RFAM database, the appropriate length of the 5′ UTR, as well as the high level of its transcription and decrease in transcription level downstream of it, suggest a bona fide riboswitch. This operon also encodes 2 PPE proteins and 2 proteins with unknown functions.

Multiple non-coding RNAs revealed and mapped in M. avium

According to the mapping results, non-coding transcriptome constitutes a significant part of mycobacterial RNA. Sequencing allowed us to reveal dozens of putative antisense RNAs, i.e., transcripts with TSPs located antisense to annotated features (Table S3) and several intergenic small RNAs.

**Antisense RNAs.** To search for TSPs of putative antisense RNAs, we implemented the same algorithm used for searching TSPs of annotated genes. We manually verified the identified candidate TSPs but did not filter them by the presence of downstream start-codons or upstream −10 sequences. In contrast to 5′ ends, the mapping of 3′ ends is arbitrary, because in most cases the transcription level is decreased significantly from 5′ to 3′ ends of cis-encoded ncRNAs. In other words, many putative antisense RNAs may represent series of transcripts with varying lengths; therefore, we manually assigned 3′ ends to the points where the transcription level was almost decreased to 0. The length of these transcripts varied significantly (the median, 281 nt), and some of them were localized to a single gene, as in the case of asMAV_0554 or asMAV_3303 (Figures 2B and 2C), whereas others spanned several adjacent genes, as in the case of asMAV_1569–1571 (Figure 2D). We checked all the detected asRNAs for ORFs longer than 70 nt. Potential polypeptides were subjected to blast searches in the “Non-redundant protein sequences” database. The results showed that none of the detected asRNAs had ORFs that code for protein homologs in the database. While several transcripts showed relatively high transcription levels, as in the case of asMAV_3303, asMAV_1569–1571 spanning 3 genes showed much higher transcription levels than antisense genes. The full list of identified putative antisense RNAs is provided in Table S3.

### Table 3. Predicted riboswitches.

| Start | End | Gene | Product | Type |
|-------|-----|------|---------|------|
| 4445976 | 4446194 | MAV_4315 | O-acetylhomoserine sulfhydrylase | SAM-IV |
| 2947367 | 2947501 | MAV_2915 | PE family protein | Ykok leader |
| 3338199 | 3338383 | MAV_3236 | hypothetical protein | Ykok leader |
| 5347431 | 5347601 | MAV_5197 | manganese/iron transporter | Ykok leader |
| 2941785 | 2942378 | MAV_2910 | PPE family protein | putative riboswitch |

Coordinates of the riboswitches are provided according to MAH1104 genome. Genes downstream of the riboswitches and regulated by them, products of these genes and types of riboswitches according to RFAM database are demonstrated.

doi:10.1371/journal.pone.0074209.t003

Intergenic sRNAs in M. tuberculosis and M. avium: common and different

The identified intergenic sRNAs of M. avium were subjected to a search for M. tuberculosis homologs. As a result, 6 of 10 sRNAs were found to have such homologs (Table 4). For igMAV_2868–2869, the sRNA is encoded by the intergenic locus that was flanked by genes conserved among M. avium, M.
are assumed to play an important role in the adaptation of conditions that do not allow aerobic respiration. These genes constitute the so-called DosR regulon and are activated under transcriptional regulator [20]. Genes regulated by DosRS phase and is regulated by the DosRS two-component.

Transcription of MTS1338 is strongly induced in the stationary oxidative stress, which mimics the intracellular environment. MTS1338, and MTS2975 [21]. Among them, MTS479 and M. tuberculosis found in M. avium tuberculosis underlines their role in the pathogenesis.

Expression of igMAV_0468–0469 and igMAV_0469–0470 in vivo

We selected 2 intergenic sRNAs with the highest expression level in vitro, igMAV_0468–0469 and igMAV_0469–0470, to assess by qRT-PCR their expression in the lung tissue of infected mice which previously were shown to be genetically susceptible and resistant to infection. Compared to I/St, B6 mice are considerably more susceptible to infection caused by M. avium in terms of bacterial multiplication and damage of the lung tissue: in B6, but not I/St mice prolonged infiltration with neutrophils leads to formation of necrotic lung granulomata and death [42]. The expression of igMAV_0468–0469 and igMAV_0469–0470 was very high in culture medium, only slightly lower in the lungs of susceptible, but dramatically reduced in the lungs of resistant mice (Figure 4). Interestingly, Arnvig et al. demonstrated that the transcript of MTS2293, the M. tuberculosis homolog of igMAV_0468–0469, accumulated at a high level comparable to that for 16S rRNA in the lungs of B6 mice which are relatively resistant to TB [20]. The different expression profiles of MTS2293 and ncMAV_0468–0469 in resistant mice may be explained either by differences in physiology between M. avium and M. tuberculosis or by differences in infection models. Regulation of MTS2293 expression has not been studied so far. It doesn’t contain known promoter sequences upstream of 5’ end [20], and further work on elucidation of regulatory mechanisms is needed. Although there are similarities in the infection models of M. tuberculosis (resistant B6 mice) and M. avium (resistant I/St mice), the 2 models are not identical. They differ in the regulation of the neutrophil response, which may be explained by different patterns of genetic control between 2 infections [43]. Specifically, the susceptibility of mice to M. avium infection depends on a single gene, Nramp1, which encodes a proton-coupled divalent metal ion transporter associated with bacterial phagosomes in macrophages [44]. In contrast, the susceptibility of mice to M. tuberculosis infection is always controlled by many genetic loci [45]. These differences may lead to different microenvironments inside host macrophages, thereby causing different transcriptional responses.

Conclusion

In summary, we performed massive sequencing on the Illumina platform to obtain a comprehensive description of M. avium
transcriptome. The qualitative and quantitative analysis reveals the abundance of non-coding RNAs, which even exceeds the amount of transcribed coding sequences. We found that 6 *M. avium* intergenic sRNAs were expressed at extremely high levels, compared to other sRNAs and the majority of protein-coding genes. Among them, the expression of *igMAV_0468–0469* and *igMAV_0469–0470* depends on the host genetics, which was demonstrated in *in vivo* experiments. These results may imply the involvement of these intergenic sRNAs in the adaptation of mycobacteria to the host defense system.

In recent years, many bacterial sRNAs have been shown to control the overall response to stress and act as substantial regulators of bacterial virulence [46,47]. sRNAs MTS479 and MTS1338, which are implicated in pathogenesis, are present in pathogenic *M. tuberculosis*, but not in conditionally pathogenic *M. avium*. Their expression may have an impact on pathogenesis.

The new era of sequencing technologies in the near future will provide researchers with a useful tool for acquiring transcriptional maps of many, if not all, bacteria. A systematic comparison of their expression profiles at different stages of infection will help assess the influence of sRNAs on different virulence phenotypes.

### Materials and Methods

#### Ethics statement

Mice of inbred strains I/StSnEgYCit (I/St) and C57BL/6YCit (B6) were bred and maintained under conventional, non-specific-pathogen-free conditions at the Animal Facilities of the Central Institute for Tuberculosis (Moscow, Russia) in accordance with guidelines from the Russian Ministry of Health (guideline 755) and the NIH Office of Laboratory Animal Welfare (assurance A5502-06). Water and food were provided ad libitum. Female mice aged 2.5–3.0 mo. at the beginning of experiments were used. All experimental procedures were approved by the Central Institute for Tuberculosis Institutional Animal Care Committee (IACUC), protocols 6, 7, 8, 11, 14 of March 6, 2012.

#### Strains and culture conditions

A previously characterized [48] chicken isolate *M. avium* subsp. *avium* strain TMC724 (ATGC25291) was a kind gift of T. Ulrichs, Max Planck Institute for Infectious Biology, Berlin, Germany. Bacilli were grown at 37°C in Dubos broth with addition of 0.05% Tween 20 for approximately 2 weeks until logarithmic growth phase (OD<sub>600</sub> ~ 0.6) was achieved.
Infection and quantitative real-time PCR
Two groups of 5 female I/St and B6 mice were infected with $1 \times 10^3$ viable CFU of *M. avium* in the aerosol chamber as described [42]. 13 weeks following challenge, lungs were homogenized in Trizol reagent (Life Technologies, Carlsbad, CA) and RNA was isolated according to a standard protocol using BeadBeater cell disrupter with 0.1 mm zirconia/silica beads (BioSpec Products, Bartlesville, OK). cDNA for qRT-PCR was made with random primers and Superscript II according to manufacturer’s instructions (Life Technologies). qRT-PCR was performed using Stratagene MX3005p machine (Agilent Technologies, Santa Clara, CA) and qPCRmix-HS SYBR reagent (Evrogen, Moscow, Russia). PCR primers are listed in Table S5. All amplifications were repeated in triplicates.

RNA isolation and mRNA enrichment
Three independently grown bacterial cultures were cooled rapidly on ice and centrifuged. RNA was isolated as previously described [49]. Briefly, cell pellets were resuspended in Trizol reagent (Invitrogen) and shaken 3 times for 30 sec. at BeadBeater cell disrupter with 0.1 mm zirconia/silica beads. RNA was isolated using phenol-chloroform extraction and re-suspended in mQ water. RNA isolated from 3 cultures was mixed and treated with Turbo DNase (Life Technologies) to remove traces of genomic DNA and purified with the RNeasy mini kit (Qiagen, Venlo, Netherlands). The absence of DNA contamination was verified by PCR with primers specific to 16S rRNA. RNA sample was depleted of 16S and 23S rRNA with MicrobExpress kit (Life Technologies) and ethanol precipitated. Depletion of 16S and 23S rRNA was verified by the agarose gel electrophoresis.

RNA processing and illumina sequencing
To prepare cDNA for sequencing we applied “Illumina RNA ligation protocol” described in [24] with slight modifications. Briefly, 200 ng of RNA was fragmented with RNA fragmentation reagent (Life Technologies) by heating at 70°C for 10 min. After ethanol precipitation, RNA was treated with Shrimp alkaline phosphatase (Fermentas, Vilnius, Lithuania) to remove 5’ and 3’ phosphate groups including threephosphate groups from the 5’ ends of RNA molecules. RNA was purified with RNeasy mini kit (Qiagen) according to modified protocol from “Directional mRNA-Seq Sample Preparation guide” issued by Illumina. RNA was 5’ phosphorylated using T4 Polynucleotide kinase (Fermentas), ethanol precipitated and dissolved in mQ water. 3’ and 5’-RNA adapters were ligated and cDNA was constructed using TruSeq™ Small RNA Sample Preparation kit (Illumina, San Diego, CA). After 12 cycles of PCR amplification, cDNA constructs were fractionated in polyacrylamide gel. 65–75-nt-long cDNA fragments were excised, eluted from the gel and sequenced by running 76 cycles on the Illumina Genome Analyzer IIx.

Processing of RNA-seq data
Before mapping of reads to the genome, terminal sequences of adapters were trimmed with Perl script. The reads were mapped to *M. avium* subsp. hominisuis 104 genome (GenBank accession number cp000479.1) with Bowtie 2 [50], setting parameters: -q – sensitive-local. Mapping statistics and RPKM values [33] for annotated genes and trans-encoded ncRNAs was calculated using custom Perl scripts. Reads overlapping a gene at least by one nucleotide were accounted for when calculating RPKM for this gene. Transcriptional profiles for forward and reverse strains of genome were generated, representing count of overlapping reads for every nucleotide of the genome. The transcriptional profile was visualized with Artemis genome browser [26].

![Figure 4. Expression of igMAV_0468–0469 and igMAV_0469–0470 in *M. avium* isolated from cultures or the lungs of resistant and susceptible mouse strains. The quantities of intergenic sRNAs relative to 16S rRNA are shown on a logarithmic scale. The relative quantities of intergenic sRNAs, as well as their confidence intervals, are summarized in the table below the figure. The CFU/lung values for I/St and B6 mice are also shown. doi:10.1371/journal.pone.0074209.g004](image-url)
To search for genomic loci, present *M. avium* hominisuis 104 but absent in *M. avium* subsp. *avium* TMC724, we aligned genomes of these strains with Nucmer package, a part of MUMmer 3.0 software [31]. Textual output was parsed with Perl script to retrieve coordinates of these loci. They were compared with non-transcribed regions of genome to reveal MAH-specific LSPs.

Abrupt increases of transcription level were considered to be putative TSPs. They were mapped with Perl script, which considered an increase to be a putative TSP, if it met several empirical criteria: (i) more than twofold rise of transcription in 3-nucleotide window upstream the increase (ii) transcription level at the highest point of the increase is higher than at every point in 60 nt window upstream of the increase (iii) transcriptional level at the highest point of the increase is more than 10. All found putative TSPs were manually verified. TSPs located in intergenic loci were considered to be TSPs of their respective genes. TSPs located in antisense orientation to the genes were considered to be TSPs of cis-encoded ncRNAs.

**Accession codes**

Sequence read data, transcriptional profile for visualization in Artemis genome browser and RPKM values for annotated genes have been submitted to the GEO database at NCBI under accession number GSE46281.

**5’ RACE and northern blotting**

For verification of the revealed TSPs five genes were randomly selected (MAV_1369, MAV_3090, MAV_4124, MAV_0525, MAV_1690) and 5’ RACE was made using FirstChoice® RLM-RACE Kit (Life Technologies). Products of 5’RACE were amplified with external and internal primers (Table S5) and sequenced.

Northern blotting for confirmation of presence of trans-encoded ncRNAs MAV_1415-1416, MAV_2868-2869 and MAV_1034-1035 was made using riboprobes generated with Riboprobe Systems (Promega, Fitchburg, WI). Primers for generation of DNA templates for riboprobe generation are listed in Table S5.

**References**

1. Heifetz L (2004) Mycobacterial infections caused by nontuberculous mycobacteria. Semin Respir Crit Care Med 25: 283–295.
2. Thorel MF, Kirschvsky M, Levy-Frebault VV (1990) Numerical taxonomy of mycobactin-dependent mycobacteria, emended description of *Mycobacterium avium*, and description of *Mycobacterium avium* subsp. avium subsp. nov., *Mycobacterium avium* subsp. paratuberculosis subsp. nov., and *Mycobacterium avium* subsp. silvaticum subsp. nov. Int J Syst Bacteriol 40: 254–260.
3. Turene C, Collins DM, Alexander DC, Behr MA (2008) *Mycobacterium avium* subsp. paratuberculosis and *M. avium* subsp. avium independently evolved pathogenic clones of a much broader group of *M. avium* organisms. J Bacteriol 190: 2479–2487.
4. Divorska I, Bull TJ, Bartos M, Matova L, Svastova P, et al. (2003) A standardised restriction fragment length polymorphism (RFLP) method for typing *Mycobacterium avium* isolates links IS901 with virulence for birds. J Microbiol Methods 55: 11–27.
5. Harris NB, Barletta RG (2001) *Mycobacterium avium* subsp. paratuberculosis in Veterinary Medicine. Clín Microbiol Rev 14: 409–512.
6. Horsburgh CR Jr (1991) Mycobacterium avium complex infection in the acquired immunodeficiency syndrome. N Engl J Med 324: 1332–1338.
7. Nightingale SD, Byrd LT, Southern PM, Jocusch JD, Cal SN, et al. (1992) Incidence of *Mycobacterium avium*-intracellulare complex bacteremia in human immunodeficiency virus-positive patients. J Infect Dis 165: 1082–1083.
8. Thegerstrom J, Marklund BI, Hoffner S, Axelsson-Olsson D, Kauppinen J, et al. (2012) Identification of small RNAs in Mycobacterium tuberculosis. Mol Microbiol 73: 397–408.
9. DeChira JM, Contreras-Martinez LM, Levy J, Smith D, McDonough KA, et al. (2010) Multiple small RNAs identified in *Mycobacterium bovis* BCG are also expressed in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. Nucleic Acids Res 38: 4076–4078.
10. Pellin D, Miotto P, Ambrosi A, Cirillo DM, Di Serio C (2012) A genome-wide identification analysis of small regulatory RNAs in *Mycobacterium tuberculosis* by RNA-seq and conservation analysis. PLoS One 7: e32723.
11. Miotto P, Forti F, Ambrosi A, Pellin D, Veiga DF, et al. (2012) Genome-wide discovery of small RNAs in *Mycobacterium tuberculosis*. PLoS One 7: e31950.
12. Arvinik KR, Young DB (2009) Identification of small RNAs in *Mycobacterium tuberculosis*. Mol Microbiol 73: 397–408.
13. Ignatov DV, Mefodieva LG, Majurov KB, Skvortsov TA, Ashikina TL (2012) Novel small RNAs from *Mycobacterium avium*. Russian Journal of Bioorganic Chemistry 38: 447–449.
14. Chao Y, Vogel J (2010) The role of Hfq in bacterial pathogens. Curr Opin Microbiol 13: 24–33.
15. Ignotov DY, Mefodieva LG, Majurov KB, Skvortsov TA, Ashikina TL (2012) Novel small RNAs from *Mycobacterium avium*. Russian Journal of Bioorganic Chemistry 38: 447–449.
16. Arnvig KR, Yousef DB (2009) Expression of small RNAs in *Mycobacterium avium* subsp. paratuberculosis subsp. nov., and *Mycobacterium avium* subsp. avium. J Bacteriol 190: 2479–2487.
24. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, et al. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 7: 709–715.

25. Filiatrault MJ, Stodghill PV, Bronstein PA, Moll S, Ländberg M, et al. (2010) Transcriptome analysis of Pseudomonas syringae identifies new genes, noncoding RNAs, and antisense activity. J Bacteriol 192: 2359–2372.

26. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28: 464–469.

27. Alexander DC, Turenne CY, Behr MA (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28: 464–469.

28. Bailly-Bechet M, Vergassola M, Rocha E (2007) Causes for the intriguing presence of tRNAs in phages. Genome Res 17: 1486–1495.

29. Gioffre A, Infante E, Aguilar D, Santangelo MP, Klepp L, et al. (2005) Mutation in mce operons attenuates Mycobacterium tuberculosis virulence. Microbes Infect 7: 325–334.

30. Papenfort K, Vogel J (2010) Regulatory RNA in bacterial pathogens. Cell Host Microbe 8: 116–127.

31. Arnvig KB, Young DB (2010) Regulation of pathogen metabolism by small RNA. Drug Discovery Today: Disease Mechanisms 7: e19-e24.

32. Papenfort K, Vogel J (2010) Regulatory RNA in bacterial pathogens. Cell Host Microbe 8: 116–127.

33. Arroyo KB, Young DB (2010) Regulation of pathogen metabolism by small RNA. Drug Discovery Today: Disease Mechanisms 7: e19-e24.

34. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28–36.

35. Zheng X, Hu GQ, She ZS, Zhu H (2011) Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. BMC Genomics 12: 361.