8-2012

Systems Biology Approaches to Probe Gene Regulation in Bacteria

Diogo F. Troggian Veiga

SYSTEMS BIOLOGY APPROACHES TO PROBE GENE REGULATION IN BACTERIA

By

Diogo Fernando Troggian Veiga, M.Sc.

APPROVED

_____
Gábor Balázsi, PhD

_____
Tim Cooper, Ph.D

_____
Ju-Seog Lee, Ph.D

_____
John N. Weinstein, M.D, Ph.D

_____
Ambro van Hoof, Ph.D

APPROVED

_____
Dean, The University of Texas Health Science Center at Houston
Graduate School of Biomedical Sciences

SYSTEMS BIOLOGY APPROACHES TO PROBE GENE REGULATION IN BACTERIA

A

DISSERTATION

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
and
The University of Texas
MD Anderson Cancer Center
Graduate School of Biomedical Sciences
in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Diogo Fernando Troggian Veiga, M.Sc.

Houston, TX

August 2012

Dedication

I dedicate this work to my first mentor, Luismar Marques Porto, who gave me the opportunity to grow in a fruitful scientific environment, and taught me the basics of scientific research. His vision and ideas inspired me to become a computational biologist, and to pursue scientific questions by multidisciplinary approaches.

I also dedicate this dissertation to my mother, Zenaide Salete Veiga, for her continuous support and unconditional trust during this period of my life. My mother is one of the most courageous persons I have ever known, and her courage and determination serve as a model to my own endeavors.

Acknowledgments

I would like to express my deep gratitude to my mentor, Gábor Balázsi, who accepted me in his lab, trusted on my ideas and gave me the freedom to pursue them. Gábor was an excellent mentor, always available when I needed, and the lab microenvinroment that he created provided the ideal conditions for my doctoral studies.

I thank all my colleagues for their invaluable help, suggestions and “hard time” during lab meetings, which were vital to improve my research. In special Dmitry Nevozhay, who taught me most of the molecular biology experiments performed in this dissertation.

As a foreign student staying for a long period abroad, I met many wonderful people going through similar experiences. Each of them enriched my life with their friendship. Many of these friends already went back to their countries, but they will always be remembered. In special, I would like to thank Tiago Paixão, my friend and roomate who shared with me much of this period. Tiago was the older brother that I never had, helping me through difficult moments. Of course, he also taught me to cook many excellent Portuguese dishes, as well as introduced me to very good music.

Finally, I was lucky to meet the love of my life, Manuela during this period. Meeting her was certainly the greatest gift I had while living in Houston, and her love and support were essential during this endeavor.

Abstract

Mechanisms that allow pathogens to colonize the host are not the product of isolated genes, but instead emerge from the concerted operation of regulatory networks. Therefore, identifying components and the systemic behavior of networks is necessary to a better understanding of gene regulation and pathogenesis. To this end, I have developed systems biology approaches to study transcriptional and post-transcriptional gene regulation in bacteria, with an emphasis in the human pathogen *Mycobacterium tuberculosis* (*Mtb*).

First, I developed a network response method to identify parts of the *Mtb* global transcriptional regulatory network utilized by the pathogen to counteract phagosomal stresses and survive within resting macrophages. As a result, the method unveiled transcriptional regulators and associated regulons utilized by *Mtb* to establish a successful infection of macrophages throughout the first 14 days of infection. Additionally, this network-based analysis identified the production of Fe-S proteins coupled to lipid metabolism through the alkane hydroxylase complex as a possible strategy employed by *Mtb* to survive in the host.

Second, I developed a network inference method to infer the small non-coding RNA (sRNA) regulatory network in *Mtb*. The method identifies sRNA-mRNA interactions by integrating *a priori* knowledge of possible binding sites with structure-driven identification of binding sites. The reconstructed network was useful to predict functional roles for the multitude of sRNAs recently discovered in the pathogen, being that several sRNAs were postulated to be involved in virulence-related processes.

Finally, I applied a combined experimental and computational approach to study post-transcriptional repression mediated by small non-coding RNAs in bacteria. Specifically, a probabilistic ranking methodology termed rank-conciliation was developed to infer sRNA-mRNA interactions based on multiple types of data. The method was shown to improve target prediction in *Escherichia coli*, and therefore is useful to prioritize candidate targets for experimental validation.

Table of Contents

List of Illustrations

List of Tables

List of Abbreviations

MΦ *Macrophage*

MFE *Minimum Free Energy*

Mtb *Mycobacterium tuberculosis*

NetReSFun Network Response to Step Functions

p.i. Post infection

sRNA small non-coding RNA

TG Target gene

TF Transcription factor

TRN Transcriptional Regulatory Network

WRP Weighted Rank Products

Chapter 1

General Introduction – The Systems Biology Workflow

Recent advances in experimental techniques are enabling the study of organisms and their components at an unprecedented level of coverage and detail. Nowadays, high-throughput assays are able to interrogate in a genome-wide fashion every aspect of biological systems, from the quantification of genes, RNAs and proteins, to metabolites and their interactions (1), as well as the landscape of phenotypes arising from perturbations of each component of the system (2). Yet, the community readily noticed that the ability to produce large quantities of data does not translate into immediate knowledge. Indeed, the mining for successful biomarkers for new therapies or vaccines is a difficult task - analogous to a needle in the haystack - given the multitude of candidates generated by genome-scale datasets.

The discipline of systems biology emerged during the paradigm shift in molecular biology from a data-limited to a data-intensive science. The systems biology workflow aims to transform high-throughput data into testable hypotheses, and ultimately into biological discoveries (Figure 1). This multistep workflow is often initiated by assembling a large-scale biological network that describe the parts of the biological system and their interactions. Biological networks are assembled from "*omics"* datasets using network inference methods, and can be used to describe many types of functional association between entities, including enzyme

catalysis (metabolic networks), protein phosphorylation and binding (signaling networks), and protein-DNA interactions (transcriptional regulatory networks). The problem of learning regulatory interactions from data (reverse engineering of networks) has been tackled by several computational methods (reviewed in Veiga *et al* (3) for regulatory networks). The most popular approaches included the use of similarity measures such as correlation (4), partial correlation (5), and mutual information (6) to detect dependency between molecular entities.

The inferred large-scale network provides a coarse-grained description of the biological system that can be further refined to generate testable hypotheses. Fine-tuning the model to specific settings includes integrating datasets to detect active modules of the network during specific experimental conditions or finding network modules enriched for molecular functions (Figure 1). Eventually, subnetworks may be encoded in formal mathematical models when detailed biochemical parameters such as synthesis and degradation rates of molecules are available. Finally, the analysis of the refined network generates hypotheses about the functioning, dynamics or connectivity of the biological system, and experiments are performed to validate the assumptions. For example, in terms of connectivity, the refined network inferred in a condition-specific manner guides the experimentalist to test regulatory interactions in a particular condition, leading to the identification of new regulatory connections in the system. In terms of dynamics, the study of the network behavior uncovers principles of network functioning and possibly indicates how such networks can be reliably controlled.

The described systems biology workflow that turns *omics* data into testable hypotheses has been applied to systematically study gene expression control in several organisms, including the model prokaryote *Escherichia coli*. For instance, Faith *et al* (7) developed CLR, a network inference approach based on mutual information to predict regulatory interactions between transcription factors (TFs) and target genes. The method detects interactions by computing a

background corrected mutual information between the TF and the putative target, based on gene levels extracted from hundreds of microarray experiments. Then, a network is constructed by thresholding corrected mutual information values for all TF-target pairs, and the interactions that survive a cutoff are included in the transcriptional regulatory network (TRN). The large-scale model of the *E. coli* TRN generated novel hypotheses of transcriptional regulation, and experimental validation using CHIP-qPCR confirmed 21 novel TF-target interactions.

As described in the workflow, large-scale models of gene regulation usually need to be refined before generating useful hypotheses (Figure 1). The application of CLR to study functional roles of small non-coding RNAs (sRNAs) in *E. coli* (8) illustrates this process of network refinement. The inferred large-scale sRNA network was composed of several modules controlled by specific sRNAs, and modules were linked to biological functions. The functionally annotated network lead to the hypothesis that two sRNAs, IsrA and GlmZ, were implicated in DNA damage response. Indeed, confirmatory experiments demonstrated that a strain lacking both sRNAs (Δ*isrA*Δ*glmZ*) was unable to activate DNA repair after exposure to damage. Several other studies applied systems biology approaches to study gene regulation. Noteworthy initiatives attempted to reverse engineering TRNs in human cells, which lead to the identification of major regulatory hubs that govern B cells (9) and the aggressive phenotype observed in glioblastoma cancer cells (10). It is now clear that the application of systems biology tools to analyze high-throughput data can generate meaningful hypotheses and accelerate the pace of biological discovery.

In my thesis, *I have applied the systems biology workflow to study gene regulation in prokaryotes*, with an emphasis in the human pathogen *Mycobacterium tuberculosis* (*Mtb*). In the following chapters, I present computational methods designed to construct large-scale network models of gene regulation, and I demonstrate how the analysis of such networks generated new

hypotheses regarding the regulatory networks that control gene expression in the pathogen.

In Chapter 2, I introduce a network response method geared towards understanding the behavior of *Mtb* TRN during parasitism of macrophages. The method built on an expanded large-scale TRN and identified transcriptional regulators critical for pathogen survival in the host.

In Chapter 3, I present a network inference method for constructing the sRNA-mediated post-transcriptional regulatory network in *Mtb*. The reconstructed network was utilized to predict functional roles for the multitude of sRNAs recently discovered in the pathogen, and several sRNAs were postulated to be involved in virulence-related processes.

In Chapter 4, I describe a network inference method for inferring sRNA-mediated post-transcriptional regulation relying on multiple data sources for learning regulatory interactions. The method was shown to improve target prediction of several *E. coli* sRNAs, and experimental validation of top predictions found evidence of novel genes that might be under post-transcriptional control in the bacteria.

In the final Chapter, I describe my attempt to discover sequence-encoded properties that drive target selection by sRNAs. By analyzing experimental results of MicC as well as other sRNAs, I found that accessibility and conservation properties of binding sites can be used to discriminate between targets and non-targets.

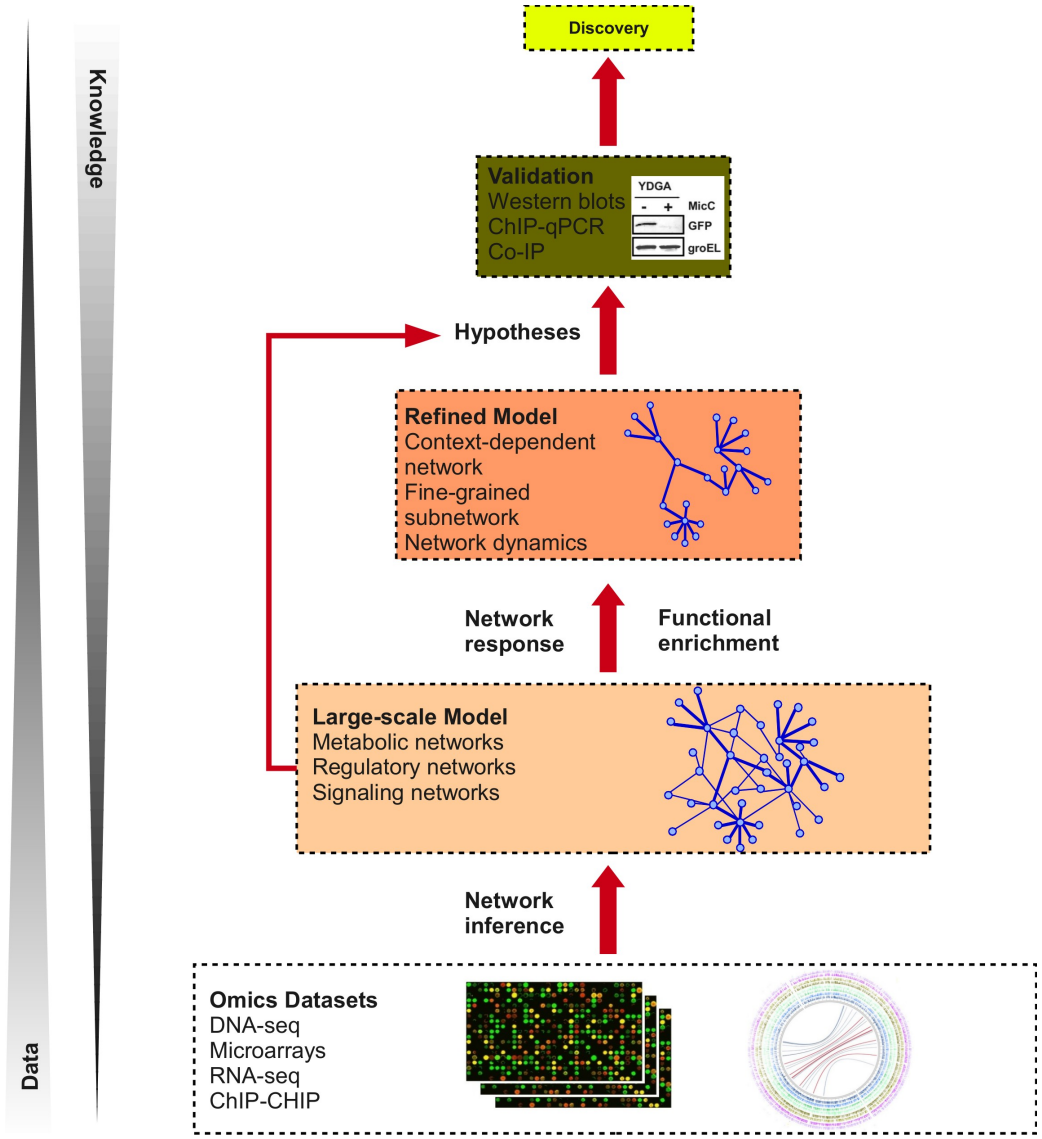

Figure 1: The systems biology analysis workflow is designed to move genome-scale data into meaningful biological hypotheses, thereby enabling biological discovery. Network inference methods are used to reconstruct large-scale biological networks from high-throughput data, while network response methods are used to evaluate network utilization in a specific context. Hypotheses for new experiments are induced from the the analysis of either a large-scale or a refined network model.

Chapter 2

Transcriptional Regulatory Network Response during Mycobacterial Long-term Infection of Macrophages

2.1 Overview of the tuberculosis infection

M. tuberculosis is a Gram-positive bacterial pathogen (genus Mycobacterium) that is the causative agent of tuberculosis in several mammals. Upon infection of the host, *Mtb* inhabit different microenviroments throughout the course of the disease. The formation of such microenvironments is largely dictated by the immune system response in the host (Figure 2). Tuberculosis infection initiates when inhaled droplets carrying the *Mtb* tubercle bacili enter the human body through the airways, and the pathogen establishes its infection niche in the alveolar tissue in the lungs. The first stage of infection elicits a response of the innate immune system, when alveolar macrophages (MΦs) are recruited to recognize the pathogen invasion and to trigger inflammation (11). Naïve MΦs recognize *Mtb* cells through cell surface receptors that bind to carbohydrates on the bacterial outer membrane. The recognition leads to phagocytosis and confinement of the pathogen in membrane-bound vacuoles known as phagosomes (Figure 2). Fusion of a phagosome with a lysosome creates a phagolysosome (P-L), and released lysozymes degrade bacterial cell walls. In the transition to the acquired immune response, the

production of inflammatory cytokines such as tumor necrosis factor γ (TNF-γ) and interferon γ (IFN-γ) by $CD8^+$ and $CD4^+$ T cells, as well as by dendritic (DN) cells induce MΦ activation and a higher rate of phagolysosome formation (12). In the persistent infection state, which corresponds to the chronic phase of the disease, *Mtb* is confined in granulomas, which are aggregates of phagocytic cells including B and T cells containing a hypoxic caseous necrosis at the center (12). The dormant *Mtb* characterized by a non-replicative phenotype might reside in such granulomas indefinitely, and the reactivation of the disease is often observed during conditions of immune suppresion (13).

Upon internalization in vacuoles, *Mtb* encounters an inhospitable environment and faces several killing mechanisms. For example, during the oxidative burst MΦs utilize NADPH oxidases bound to the phagosome membrane to produce reactive oxygen species (ROS) - the most important being hydrogen peroxide (H_2O_2), superoxide anion (O_2^-), and nitric oxide (NO) – in order to damage DNA, proteins and membrane lipids of the engulfed pathogen (14). ROS are potent microbicidal agents that oxidize and inactivate chemical groups, such as amino acids in proteins and fatty acids in lipids. NADPH oxidases also contribute to vacuole acidification by transporting H^+ protons into the phagosome cytosol. The resulting vacuole acidification increases the activity of lysozymes responsible for degrading peptidoglycans forming the cell wall. Additionally, MΦs produce lactoferrins that bind with high-affinity to iron, causing deprivation of this essential metal (14).

However, *Mtb* has evolved a number of defense mechanisms to counteract MΦ attacks. For instance, during the oxygen burst bacteria upregulate the production of oxidoreductases, in order to minimize damage caused by ROS and to reestablish the redox balance in the cell (15). *Mtb* is also able to arrest phagosome maturation by exporting the SapM phosphatase, which interferes with the PI3K signalling pathway in the host (16). Additionally, the secretion of ZmpA

downregulates the formation of the inflammasome by blocking the production of IL-1β and other interleukins (16). To fight iron deprivation, *Mtb* produces siderophores which are able to maintain adequate intracellular iron concentration (17). In the absence of oxygen, the bacteria utilize alternative respiratory pathways to carry out an anaerobic electron transport using different electron acceptors such as nitrate, nitrite and fumarate (18). *Mtb* has also acquired mechanisms that mediate acid resistance, including the upregulation of proton ATPase transporters to pump H^+ outside the cell (19), and the production of peptidoglycans and cell wall lipids such as lipoarabinomannan to decrease cell wall permeability at low pH (15). The bacteria is also able to degrade host lipids as an alternative energy source during starvation (20). Altogether, these defense mechanisms confer *Mtb* the ability to grow and survive within its host cell niche.

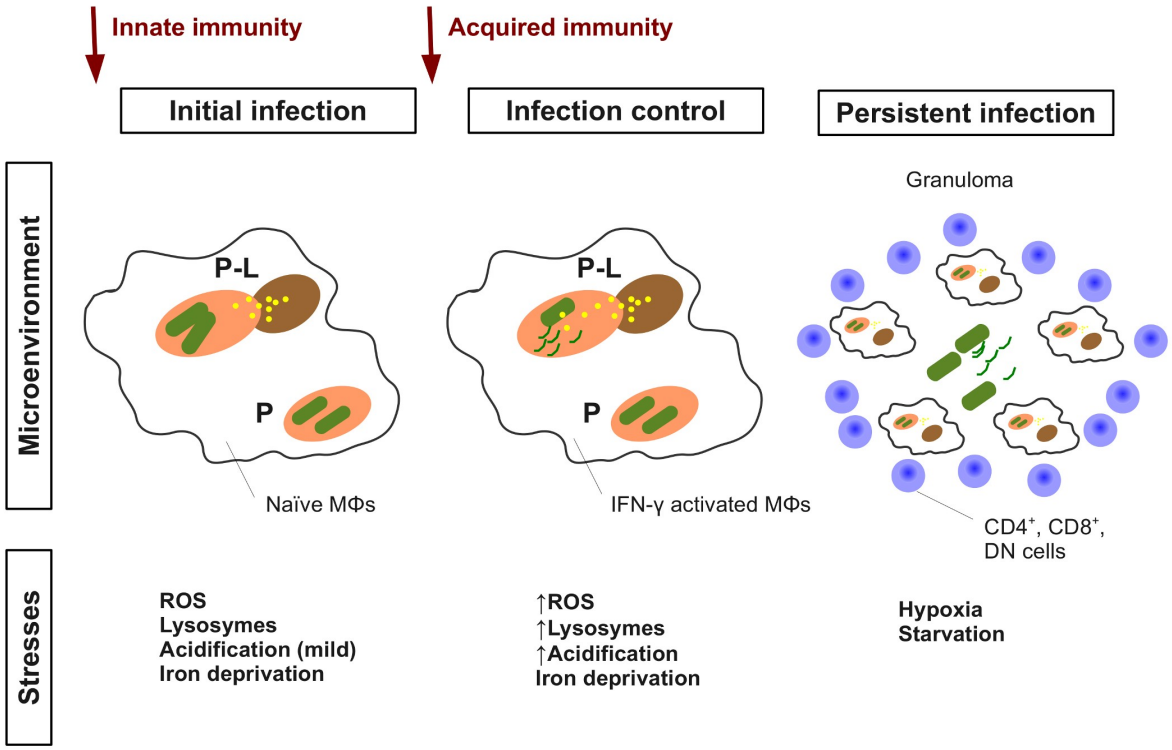

Figure 2: Progression of tuberculosis infection in the context of the innate and acquired immune system. *Mtb* resides in a variety of intracellular compartments (microenvironments) and copes with a series of stress conditions (triggered by the immune system) during the course of infection. Legend: Phagosomes (P), phagolysosomes (P-L), *Mtb* cells (green), lysozymes (yellow). See text for details.

2.2 Transcriptional profiling of *Mtb* during prolonged infection of macrophages

Transcriptome profiling of *Mtb* during MΦs infection constitutes a powerful tool to analyze pathogen-host interactions. Since the seminal work of Schnappinger *et al* (21), several microarray studies were conducted to investigate *Mtb* transcriptome changes during *in-vitro* infection of immune cells, such as human THP-1 cells (22), dendritic and natural killer cells (23), and bone marrow derived MΦs from mice (24). These studies found that upon phagocytosis, *Mtb* induces the expression of genes to fight iron deprivation, oxidative and nitrosative stresses, as well as to use fatty acids as an energy source. However, while these

studies helped to define the transcriptional signature of *Mtb* during infection, most of them focused on the early stages of the infection process.

Most recently, Rohde *et al* (25) used a combination of experimental techniques, including electron microscopy, CFU growth assays and microarrays to study the long-term behavior of the tubercle bacili during infection of naïve MΦs from mice. These multiple experimental platforms were used to follow the *Mtb* infection during a period of 14 days. The study found that *Mtb* undergoes different growth phases throughout the first 2 weeks of infection of naïve MΦs (Figure 3A). In the stress phase, which spans the initial 48 hours post infection (p.i.), a sharp decrease of ~50% of viable *Mtb* CFU was observed. This is followed by an adaptation phase - days 2 to 6 p.i. - characterized by a slight CFU increase, and a growth phase spanning days 6 to 14 p.i. Therefore, the growth curve indicates that, following an initial stress phase and decrease in CFU, *Mtb* appears to adapt in the phagosome environment before steady growth at late time points. Electron microscopy imaging found that *Mtb* may reside in different cellular compartments during the infection, and that the bacterial burden (number of *Mtb* cells per MΦ) increased in late time points (Figure 3B). The study also collected the total RNA of viable *Mtb* after 2 hr p.i., and every other 2 days until the end of 14 days, and subjected the samples to transcriptome profiling.

However, the molecular mechanisms that allow *Mtb* to establish a successful infection are still poorly understood. According to the systems biology paradigm, analyzing the temporal dynamics of gene expression in the context of a large-scale regulatory network may shed light on such mechanisms. Therefore, identifying the systems-level response of the regulatory network - the concerted action of the regulatory network during the prolonged infection - may be helpful to uncover key regulators and associated molecular mechanisms that allow *Mtb* to survive within macrophages. Moreover, this systems-level understanding of the transcriptional

response may lead to new insights into the biology of adaptation and growth of the pathogen during prolonged infection of MΦs, which in turn may guide the development of novel drugs for arresting *Mtb* infection.

In order to obtain a systems-level understanding of the *Mtb* transcriptional response during macrophage infection, I developed a network-based methodology to analyze the 14-day time-course microarray data collected during the prolonged infection model. The network response method identified subnetworks of the *Mtb* global transcriptional regulatory network that are utilized by the pathogen to counteract phagosomal stresses and ensure survival within resting macrophages. As a result, the method unveiled transcriptional regulators and associated regulons utilized by *Mtb* to establish a successful infection of macrophages throughout the first 14 days of infection. Additionally, this network-based analysis identified the production of Fe-S proteins coupled to lipid metabolism through the alkane hydroxylase complex as a possible strategy employed by *Mtb* to survive in the host.

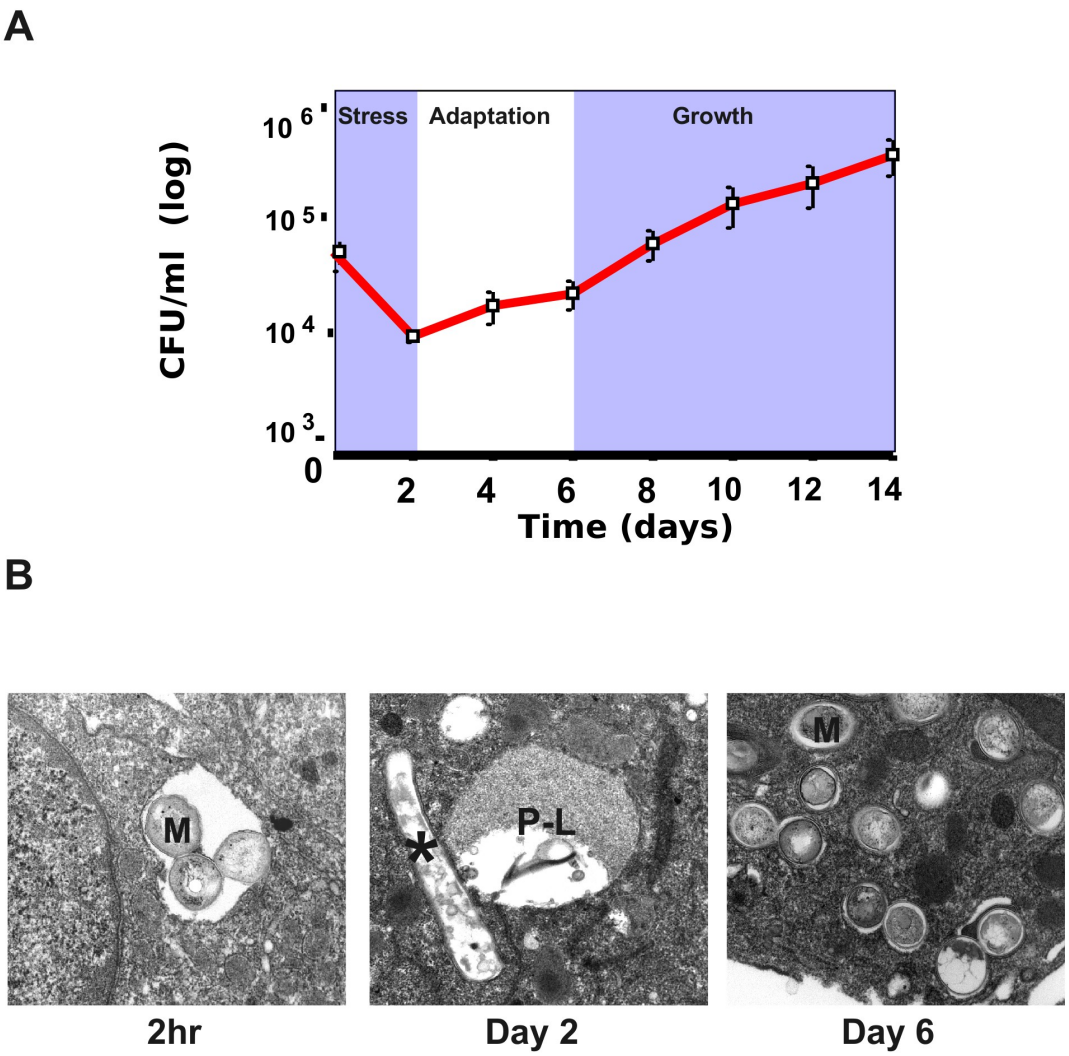

Figure 3: Long-term interactions of *Mtb* with Mφs. (A) Survival of *Mtb* CDC1551 isolated from resting bone-marrow derived Mφs at different times post infection. (B) Electron microscopy images depicting different compartments inhabited by *Mtb* during infection: (i) An intact *Mtb* (M) confined within a vacuole at 2hr, (ii) *Mtb* residing in a phagolysosome at day 2, and (iii) Rapid increase in the number of bacteria per cell, from 1.5 *Mtb*/cell at 2hr p.i. to more than 10 *Mtb*/cell at day 6. Experiments performed by Kyle Rohde, Cornell University. [Reprinted from (25)under the Creative Commons License].

2.3 Assembling a large-scale *Mtb* TRN

As described in the systems biology workflow (Chapter 1), the initial step to define the network response of *Mtb* during Mφ infection involves the construction of a large-scale transcriptional regulatory network (TRN). I assembled a large-scale *Mtb* TRN containing gene regulatory interactions extracted from both experimental and computational datasets. The previous available large-scale TRN (26) comprised 738 genes (%18 of the genome) and 937 links obtained from three sources: (*i*) literature mining; (*ii*) MtbRegList database (27); and (*iii*) inference from orthology with *E. coli* (28). For network expansion, I collected gene regulatory data from two additional sources: (*iv*) MycoRegNet database (29), and (*v*) regulatory interactions discovered by a new bacterial one-hybrid reporter system termed TB1Hybrid (30).

MycoRegNet is composed of predicted protein-DNA interactions transferred from *Corynebacterium glutamicum* to *Mtb* using a methodology that relies on orthology and conservation of TF binding sites. Such orthology-based transferring of regulatory interactions has been shown to be valid when applied to phylogenetically related bacteria, which is the case of *C. glutamicum* and *Mtb* (both species of Actinobacteria). Accordingly, a regulatory interaction was transferred from *C. glutamicum* to *Mtb* when both of the following were true: (i) a *C. glutamicum* transcription factor (TF) and its target gene (TG) had orthologs in *Mtb* (inspected by a BLASTP protein search using the e-value cutoff $e = 10^{-4}$); and (ii) the *C. glutamicum* TF recognizes a binding site that is conserved in *Mtb* and located upstream of the *Mtb* TG. The binding site profiles (motifs) for each TF were constructed based on experimentally mapped interactions in *C. glutamicum*.

I also performed operon-based network expansion, propagating a TF's regulatory effect to all members of the operon containing a given TG. Orthology-based regulatory links inferred

from the *C. glutamicum* dataset expanded the TRN considerably, by adding 425 new interactions. The high overlap with regulatory links supported by experimental literature deemed confidence to these interactions relying on orthologous TF-TG gene pairs in the two organisms (Figure 4A). The TB1Hybrid assay discovered 114 regulatory interactions, 31 of which were exclusively identified by this method (Figure 4A).

The final *Mtb* TRN, obtained after adding new regulatory interactions to the existing network and performing operon-based extension, contained 1133 genes (28% of the genome) and 1801 regulatory links, more than a half of which were experimentally determined (Figure 4B, 4C).

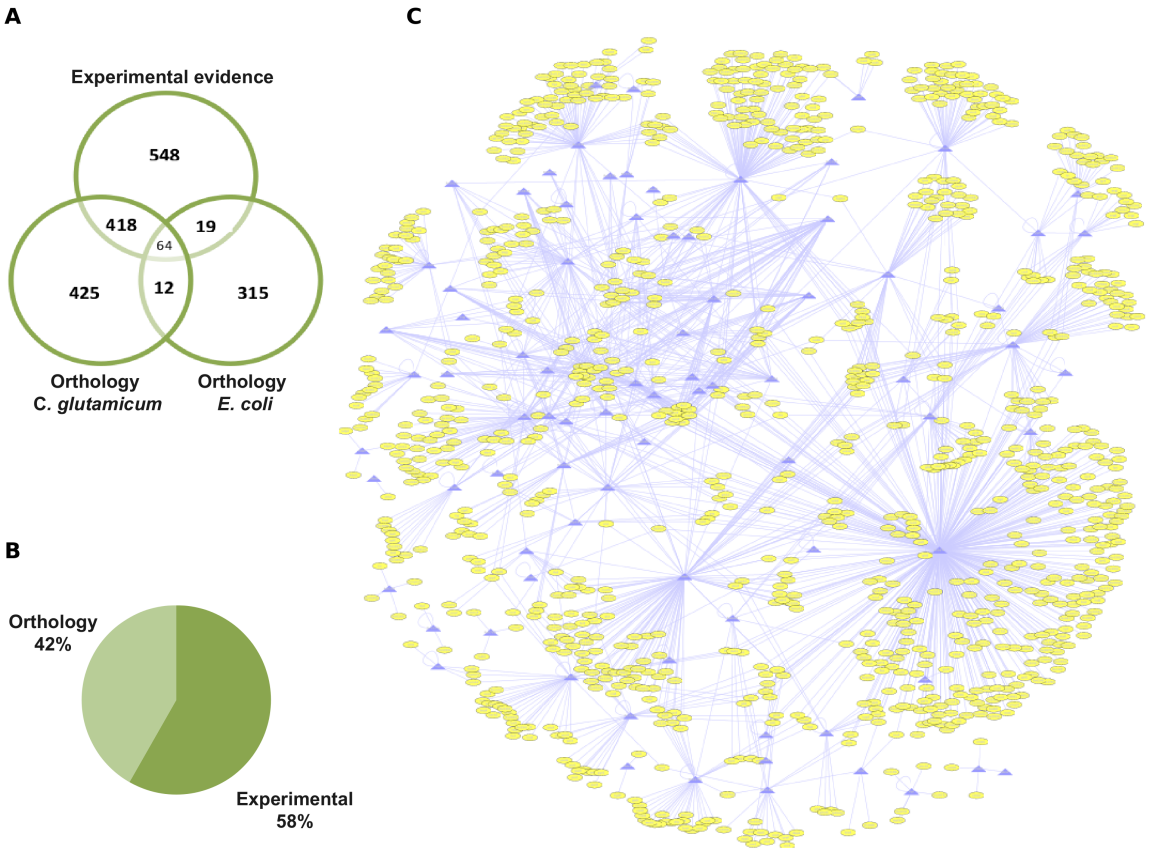

Figure 4: The expanded *Mtb* Transcriptional Regulatory Network (TRN). (A) Number of regulatory links from each data source and corresponding overlaps. Links with experimental evidence originate from literature, as well as from the MtbRegList database and the TB1H assay. (B) Distribution of interactions based on their inference method. (C) Overview of the TRN, depicting protein-DNA interactions as edges linking TFs (blue triangles) to TGs (yellow circles).

2.4 Regulon response identification using NetReSFun

The global TRN provides a static summary of all possible regulatory schemes that the bacteria may use when facing a broad spectrum of environments, ranging from normal to stressful conditions. However, previous work demonstrated that bacteria utilizes its TRN in an environment-dependent fashion, by modulating a hierarchy of transcription factors in response to changing conditions (31). Therefore, my goal was to identify which parts of the global *Mtb* TRN are differentially regulated to enable survival within macrophages. Specifically, I decomposed the global TRN into regulons, which are subnetworks composed of a transcriptional regulator and its immediate targets (Figure 5A). Then, to understand how *Mtb* utilizes its regulons from the TRN to respond to phagosomal stresses, I improved the earlier method called NetReSFun (Network Response to Step Functions) and applied this methodology to identify *Mtb* regulons that respond (*i.e* become transcriptionally active) during macrophage infection.

NetReSFun relies on covariance to quantify gene expression changes between consecutive time points (*t, t+1*). Initially, the method computes the individual response of every gene in the network as the covariance between the gene expression profile and a set of step functions. Each step function *t* "jumps" from 0 to 1 at time point τ, as shown below:

$$s(\tau,t)=\begin{Bmatrix}0,t<\tau\\1,t\geq\tau\end{Bmatrix} \quad (1)$$

Then, the individual gene response (Equation 2) is defined as the scaled covariance $\mathrm{cov}_i(\tau)$ between the expression profile $x_i(t)$ of gene *i* and a step function $s(\tau,t)$ that jumps at

time point τ:

$$\mathrm{cov}_i(\tau) = \left\langle [x_i(t) - \bar{x}_i][s(\tau,t) - \bar{s}(\tau)] \right\rangle / \sigma[s(\tau,t)] \quad (2)$$

where brackets denote averaging over genes, horizontal bar is averaging over time, and $\boldsymbol{\sigma}$ is the standard deviation.

Figure 5B illustrates how the methodology compares gene expression profiles to these pre-defined step functions using scaled covariance to detect whether a gene's expression peaked at time τ, i.e., changed in the interval [τ-1, τ].

Next, to detect regulon-wide change between subsequent time points, the method computes the combined response of regulon *I,* or *Cov-score,* which is defined as the mean of the absolute covariances of regulon genes

$$Cov_I = \left\langle \left| \mathrm{cov}_i(\tau) \right| \right\rangle_I \quad (3)$$

The method uses a non-parametric test to evaluate the significance of the *Cov-score* in order to identify responsive regulons from time course microarray data. This non-parametric test assigns a p-value to the observed regulon response Cov_I based on a reference cumulative distribution (c.d.f) created by a bootstrapping procedure. Specifically, the c.d.f F_I is constructed with *Cov-scores* drawn of 1,000 random regulons of the same size of regulon *I*, but assembled using nodes randomly chosen from the network. The density estimation of the c.d.f is performed by assigning probability mass 1/1000 for each random score, and linear interpolation is applied

to transform the discrete F_I into a continuous function in the interval [0,1]. Finally, a p-value of observing Cov_I by chance can be readily estimated as below.

$$\text{p - value} = 1 - F_I(Cov_I) \tag{4}$$

A significant *Cov-score* (i.e. p-value < 0.05) indicates regulon response, meaning that the expression levels of the subnetwork's gene members are either down- or upregulated between consecutive time points (t, $t+1$). Alternatively, from the TF "perspective", a simultaneous change of its direct target genes may also be a surrogate of the TF's activity. Under this assumption, NetReSFun may recognize regulators that become active through post-translational modifications, such as phosphorylation and metabolite binding, but do not show increased expression levels. Therefore, from the perspective of the regulator associated with each regulon, a significant *Cov-score* indicates that this particular TF might be active and hence regulating its target genes during (t, $t+1$).

To determine the direction of transcriptional change within the regulon, the method computes the deviation index for each regulon as the positive/negative ratio of individual covariances, as shown below

$$R_I(\tau) = \sum \text{cov}_i > 0 \Big/ \Big| \sum \text{cov}_i < 0 \Big| \tag{5}$$

The z-score of the deviation index, which relates the change of a given regulon *I* to all other regulons in the TRN, is given by

$$Z_I(\tau) = R_I(\tau) - \langle R(\tau) \rangle_A / \sigma[R(\tau)]_A \quad (6)$$

where A denotes all regulons extracted from the TRN, brackets denote averaging over all regulon deviations, and σ is the standard deviation.

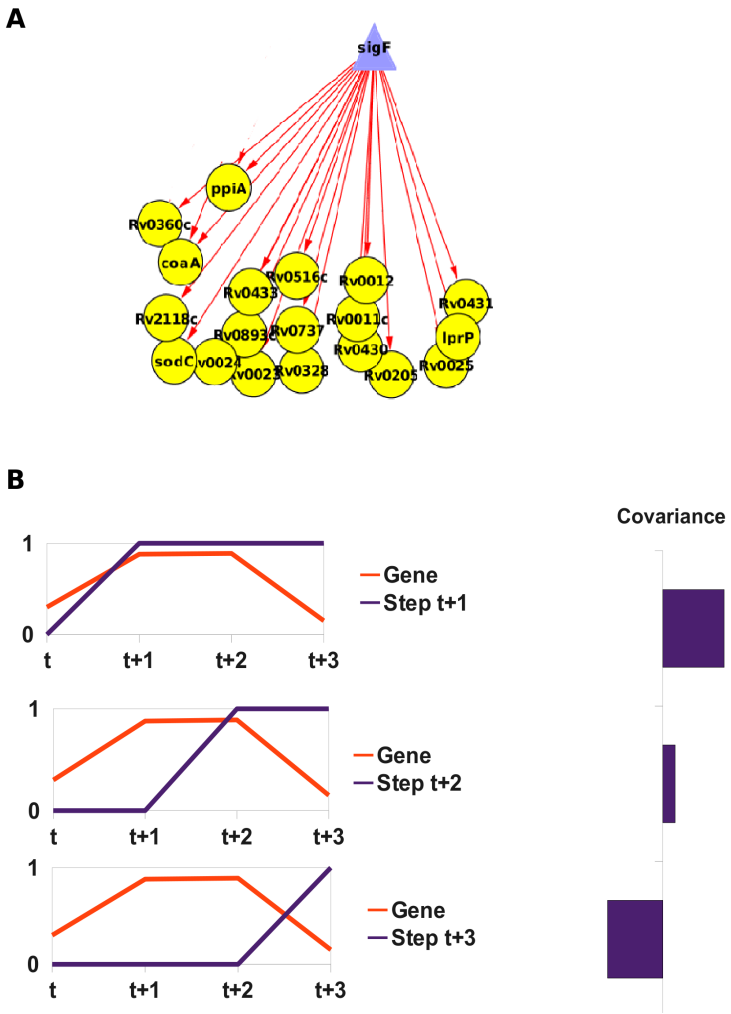

Figure 5: **Regulon response using NetReSFun.** (A) *sigF*-controlled regulon extracted from the TRN is composed of immediate direct targets. (B) Schematic illustration of how NetReSFun identifies peaks of gene activity during a time course based on covariance. In the interval (*t, t+1*), both the gene expression (red) and the step function (blue) increase sharply, resulting in high and positive covariance. Next, in the interval (*t+1*, *t+2*), the gene level exhibits only a smaller increase, causing the covariance to be marginally positive. In the interval (*t+2*, *t+3*), the covariance will be high and negative, indicating that the gene's expression and the step function change in opposite direction.

2.5 Transcriptional Regulatory Network (TRN) response to phagosome cues

I applied NetReSFun to analyze the 14-day macrophage infection time course data and thereby to identify regulons that are utilized by the bacteria to counteract phagosomal stresses during long-term infection of Mφs. The results of this analysis are summarized in the temporal map of network response (Figure 6A), which depicts responsive regulons during the time course, at a significance level of 0.05. The quantity plotted in the network response map is the deviation *z*-score of each responsive regulon as computed by Equation 6. Hence, the color indicates whether the overall trend of expression change within the subnetwork was positive (red) or negative (blue) at a given time interval, when compared to whole network. Intermediate colors denote regulons involving both up- and downregulated genes. The accompanying heatmap in Figure 6B indicates the source of regulatory links within the regulon (darker colors corresponding to higher fractions of links based on experimental evidence).

Strikingly, the map reveals that the dynamics of the TRN utilization occurs in a defined pattern that can be mapped to distinctive phases of intracellular growth, and involves (*i*) upregulation of transcription during stress phase, followed by a (*ii*) slow growth phase where *Mtb* utilizes its TRN to promote downregulation of transcription. Below I describe some of the regulons responding to the phagosomal cues in each phase, connecting these subnetworks to specific stress responses.

In the first 2 days p.i. - which corresponds to the stress phase in the growth curve of Figure 1A – a high number of responsive regulons was observed (20 out of 83), mostly exhibiting increased expression of involved genes. Among these were DosR, HspR, KstR, members of the WhiB family (WhiB3, WhiB4), two-component response regulators (Rv0260c, Rv0818, RegX3), and alternative sigma factors (SigE, SigK, SigM). The induction of a large

number of subnetworks is reminiscent of the general Environmental Stress Response (ESR) in yeast (32) and in *Bacillus subtilis* (33). Interestingly, a number of regulons were found to be acting in concert to counteract initial phagosomal stresses. This is the case of Rv0818, KstR, WhiB3, Rv0260c, Rv1359, Rv3557c and SigM, which together control the assembly of iron-sulfur (Fe-S) cofactors necessary for electron transfer and redox reactions in the sulphur assimilation system (SUF system), activated during iron starvation and oxidative damage (34). Also, KstR and Rv0818 are both implicated in the control of oxidoreductases, that may be used to fight reactive oxygen species produced by MΦs. The KstR regulon, strongly induced in the early infection, contains several thiolases involved in β-oxidation of fatty acids, which are known to be utilized by *Mtb* as an alternative energy source within the nutrient-depleted phagosomal compartment (20).

During the repressive phase, which initiates after ~6 days of residence inside Mϕs, the TRN was mainly utilized to repress target genes. This pattern is especially evident for a number of stress-responsive subnetworks that shift into downregulation during the slow growth phase, including HspR and DosR. This pronounced repression in transcription indicates that the bacteria reside in a niche that may still be stressful, but differs from the conditions encountered immediately after macrophage entry. For example, the SigH-controlled subnetwork displays strong negative response only late in the time course (~8 days and onwards). As a global regulator, SigH modulates the transcription of SigE and SigB, as well as its own promoter. Although SigH is not required for growth in Mφs, mutants lacking *sigH* caused reduced immunopathology and lethality in mice (35; 36). It is likely that these regulatory changes are associated with the reprogramming of the pathogen's metabolism to limited oxygen and nutrients encountered in later phases of the infection.

This downregulation can be observed in almost every late responder, with only a couple of exceptions. For instance, the FurB-controlled subnetwork displays an overall upregulation in the late time points. This large subnetwork (56 genes) comprises several secreted ESAT-6-like proteins that are induced in this period, such as EspC (Rv3615c), which has found to be immunodominant in patients with active and latent TB (37).

Importantly, the fact that some subnetworks respond throughout the entire time-course indicates their importance for establishing productive infection. The presence of *sustained responders* such as HspR and DosR can have two possible implications. First, the opposite trends in the early phase of the infection (primarily upregulation) and later phases of infection (primarily downregulation) may indicate that the stress to which these subnetworks respond initially is ameliorated at later time points. Alternatively, sustained responders may be necessary both to counteract initial phagosomal stress during the early phase of infection as well as for driving the dormancy phenotype encountered in later phases of infection. This is possibly the case of the extensive DosR regulon (124 members), which is responsible for transitions into and out of dormancy-like conditions, by promoting redox homeostasis, lipid metabolism and growth arrest (38). HspR also might be a critical regulator necessary both in early infection and dormancy stage. HspR activates a subset of the heat-shock general stress response upon Mφs invasion (39), but is also necessary in the persistent phase since Δ*hspR* strains exhibited attenuated growth in the chronic phase (40).

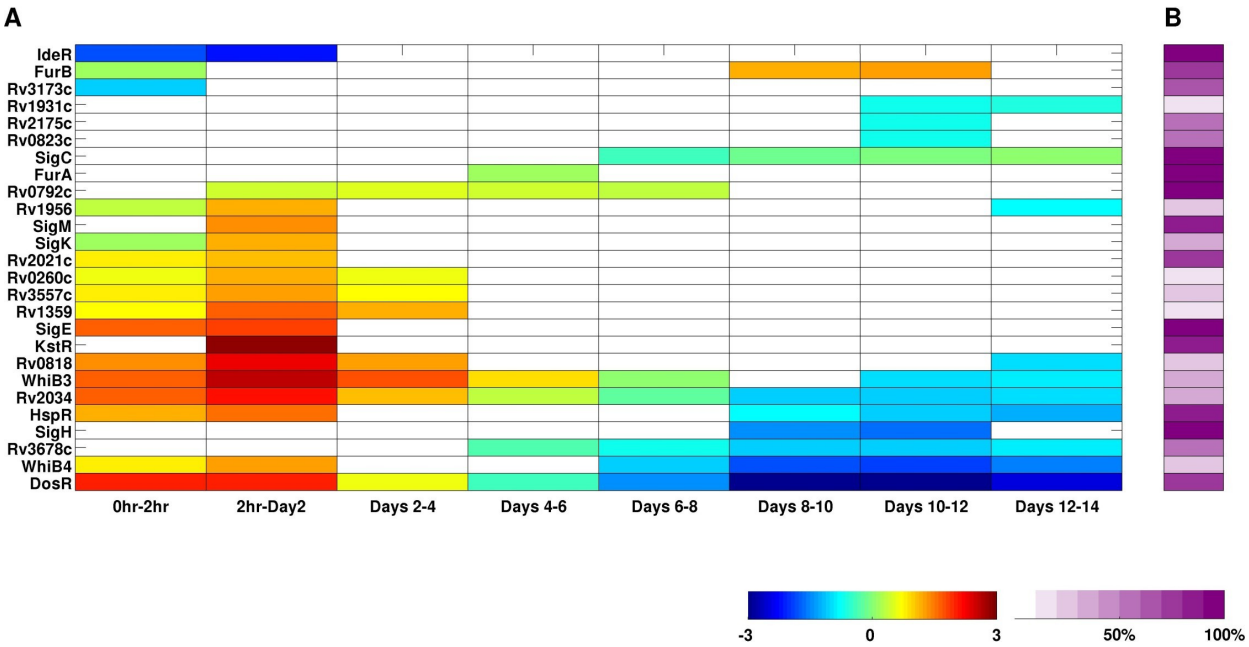

Figure 6: Temporal network response reveals the pattern of *Mtb* TRN utilization during prolonged infection of resting macrophages. (A) Most of TF-controlled regulons represented in the TRN exhibited peaks of upregulation early in the infection followed by a repressive phase in late time points. The color scale indicates the overall direction of the transcriptional change within the subnetwork, from positive (+3) to negative (-3). White color represents intervals where the regulon was not significantly responsive, considering p-value of 0.05. (B) Percentage of regulatory links within the subnetwork based on experimental evidence.

Finally, the network response suggests novel sustained responders that might be critical for *Mtb* adaptation within the intracellular compartment. For example, the Rv2034 regulon, mostly based on regulatory links inferred from *C. glutamicum* orthology, contains multiple *fadE* homologous implicated in β-oxidation of fatty acids and redox homeostasis. Notably, Rv2034 was recently characterized as an activator of *phoP* virulence gene in mycobacteria (41), which makes this regulator an interesting candidate for follow-up studies.

2.6 Assigning a role to rubredoxins synthesis during infection

Based on the analysis of the network behavior, I formulated the hypothesis that lipid metabolism mediated by the overexpression of rubredoxins (Rds) may be critical for pathogen survival in the

host. This hypothesis relies on the observation that several regulators that are active during the infection are controlling both the production of Fe-S (iron-sulfur) clusters as well as lipid metabolism. Rubredoxins are Fe-S proteins necessary for the breakdown of lipids in order to generate energy (via fatty acid oxidation), and therefore they establish a link between regulons controlling Fe-S clusters production and lipid metabolism, as explained below.

Several stress-responsive regulators control the *suf* operon (Rv1460-1466), which is up-regulated after 2 days p.i. (Figure 7A). The *suf* operon encodes the mycobacterial machinery necessary for the biogenesis of Fe-S clusters and their delivery to target apoproteins (34). Such Fe-S clusters are cofactors of wide range of proteins that are important in respiration, TCA cycle, DNA repair, gene regulation, among other biological processes (42).

By searching for candidate proteins that contain a domain for Fe-S cluster binding, I identified two Rds (*rubA* and *rubB*) that are able to acquire these newly synthesized clusters produced by the *suf* operon. Notably, these Rds are also under the control of four stress-responsive regulators, and are strongly up-regulated in the stress phase (Figure 7B). The first 3 components of the operon (*alkB*/*rubA*/*rubB*) encodes the alkane hydroxylase system, an enzymatic complex that catalyzes the hydroxylation of carbon chains (alkanes), converting octane to octanol. This is the step-limiting reaction of the octane oxidation pathway, which commits fatty acids for β-oxidation through the addition of coenzyme A (Figure 7C). Interestingly, a previous study showed that mycobacterial Rds cloned into *E. coli* mutants lacking functional Rds were able to restore growth when n-octanes were the exclusive carbon source (43)which demonstrates that *Mtb* Rds can metabolize n-octanes, possibly using the β-oxidation pathway. Here I propose that overexpression of rubredoxins, which are dependent on Fe-S clusters to become active, is used by *Mtb* for degrading n-octanes and to generate energy for intracellular viability. Based on these observations, follow-up studies should be conducted to

determine whether the inhibition of the alkane hydroxylase system constitutes a feasible approach for arresting *Mtb* infection.

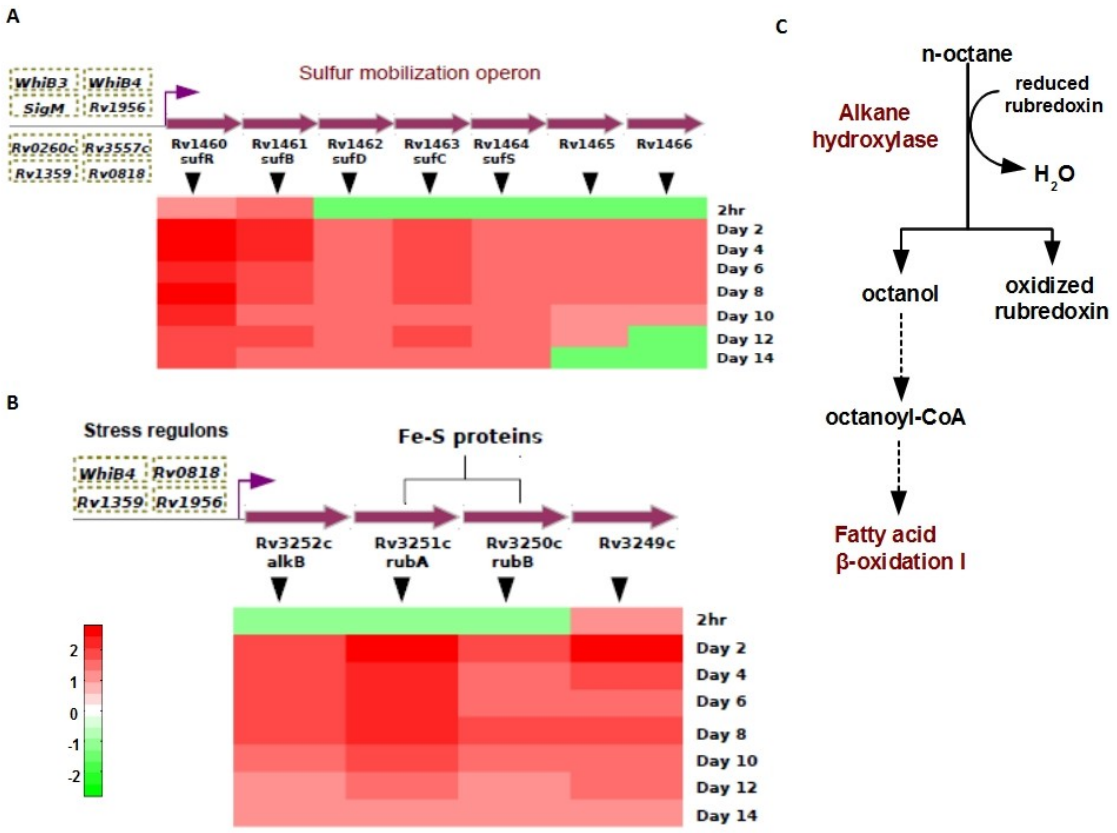

Figure 7: Network response suggests that rubredoxins link Fe-S clusters to lipid metabolism during MΦs infection. (A) Large number of stress regulons controlling the suf-operon, induced early in the time course. (B) The alkane hydroxylase complex (*alkB*/*rubA*/*rubB*) is being up-regulated after 48 hr post-infection, and it is under control of stress-responsive regulators WhiB4, Rv0818, Rv1359 and Rv1956. Heatmap shows $\log_2$ transformed expression values. (C) Rubredoxins are Fe-S proteins required for the utilization of n-octanes as energy source through the fatty acid β-oxidation pathway.

2.7 Discussion

In the clinic, *Mtb* infections are diagnosed as either active or latent. The active form of the

disease is characterized by replicating *Mtb* cells causing lesions in tissues (44). In the latent form, the immune system successfully controlled the bacterial burst and the tubercle bacili are found in a dormant (non-replicative) state. Individuals harboring latent TB are asymptomatic and free of tissue damage (44). Based on the skin test for tuberculosis, the estimates are that 1/3 of the world's population is carrying the bacteria without symptoms (45). The reactivation of the disease is at increased odds when the immune system is compromised by other diseases or repressed by drugs, such as during organ transplants. Reactivation of latent TB is the leading cause of death among HIV-infected patients (13).

Nevertheless, the mechanisms that allow the pathogen to persist in the human host are still poorly understood. The system-level analysis of the network response during infection shed light on such mechanisms. Accordingly, the extensive regulation of the Fe-S proteins during infection, and its possible connection to lipid metabolism through the alkane hydroxylase complex may configure an important mechanism for *Mtb* adaptation and persistence in the host. *Mtb* realigns its intracellular metabolism to include lipids as a carbon source, as denoted by the upregulation of genes involved in lipid metabolism in early and chronic phases of infection (20). For instance, the pathogen is able to mobilize, import and metabolize cholesterol present in the center of granulomas during chronic infection (46). Interestingly, the fatty acid β-oxidation pathway – identified by the network analysis – constitutes one of the routes used to generate energy from cholesterol breakdown (20). Fitness experiments with strains deficient in the alkane hydroxylase complex would reveal whether this enzyme is necessary for MΦ survival and/or growth using lipids as unique carbon source.

Besides identifying DosR, the regulator that is considered to be essential for counteracting phagosomal stresses, the analysis identified novel sustained responders such as WhiB4 and Rv2034. The importance of these regulators for establishing a successful infection

may also be studied by knockout experiments, seeking to evaluate whether the lack of these regulators attenuate *Mtb* growth in the host.

The current network analysis is limited to the ~ 30% of the genome for which transcriptional regulatory information is available. However, the survival mechanisms employed by *Mtb* are likely to involve additional genes and other layers of gene regulation. Large-scale models of the pathogen's interactome and metabolism were assembled. For example, a genome-scale metabolic map of the bacteria encompassing 661 genes and 939 reactions has been constructed (47), as well as a version of the *Mtb* proteome interaction map, containing physical and functional interactions, which was assembled by text mining and other protein linkage methodologies (48). With a few modifications, NetReSFun can be applied to identify active subnetworks on a expanded network that incorporates functional and regulatory interactions from multiple *omics* datasets. Overlaying temporal gene expression data on such integrated network could provide novel insights about the biological mechanisms underlying survival and persistence of *Mtb* in the host.

Chapter 3

Inferring the sRNA-mediated Regulatory Network in *Mycobacterium tuberculosis*

3.1 Small non-coding RNAs carry out post-transcriptional regulation in bacteria

Small non-coding RNAs (sRNAs) are post-transcriptional regulators involved in a variety of bacterial functions, including stress response (49), quorum sensing (50) and differentiation (51). A recurrent theme among reasonably characterized sRNAs is their involvement in the initiation of protective responses to various forms of stress, being that some sRNAs have been implicated as key regulators of bacterial response to iron limitation (52), anoxia (53), and osmotic stress (54). Differently than riboswitches and other regulatory RNA elements that are transcribed as part of the messenger RNA (mRNA), an sRNA constitutes an autonomous transcript containing features such as its own upstream promoter, start codon and terminator sites (55). Typically sRNA sizes in bacteria range from 50 to 200 nucleotides (55), and they are often divided between antisense and *trans*-encoded, according to their localization in the genome with respect to targets (Figure 8A). Antisense sRNAs (asRNAs) are transcribed from the opposite strand of protein-coding regions, and commonly overlap a large portion of their target genes (56). Such asRNAs were originally found in extra-chromosomal DNA (phages, plasmids and transposons), and constitute the first examples of antisense regulation in bacteria (57). On the other hand,

trans-encoded sRNAs originate from intergenic regions, and are able to recognize several targets by short and imperfect base pairings (58).

The base pairing of an sRNA to a region of complementarity in mRNA targets interferes directly or indirectly with the messenger translation, depending on the binding location. Often, sRNAs block translation initiation by competing for the ribosome binding site and preventing the ribosome loading on the Shine-Dalgarno region (Figure 8B). Translational arrest is also observed when sRNAs binds downstream of the start codon within the coding sequence (59). Additionally, binding in other regions of the messenger may cause decrease in mRNA stability and/or change of transcript conformation (Figure 8B). The resulting double-stranded RNA region (duplex) formed by the sRNA-mRNA interaction attracts endonucleases such as RNAse E and RNAse III which promote cleavage and degradation of the mRNA (55). Alternatively, in some cases sRNA binding can facilitate translation by exposing the ribosome binding site originally sequestered in the transcript (56).

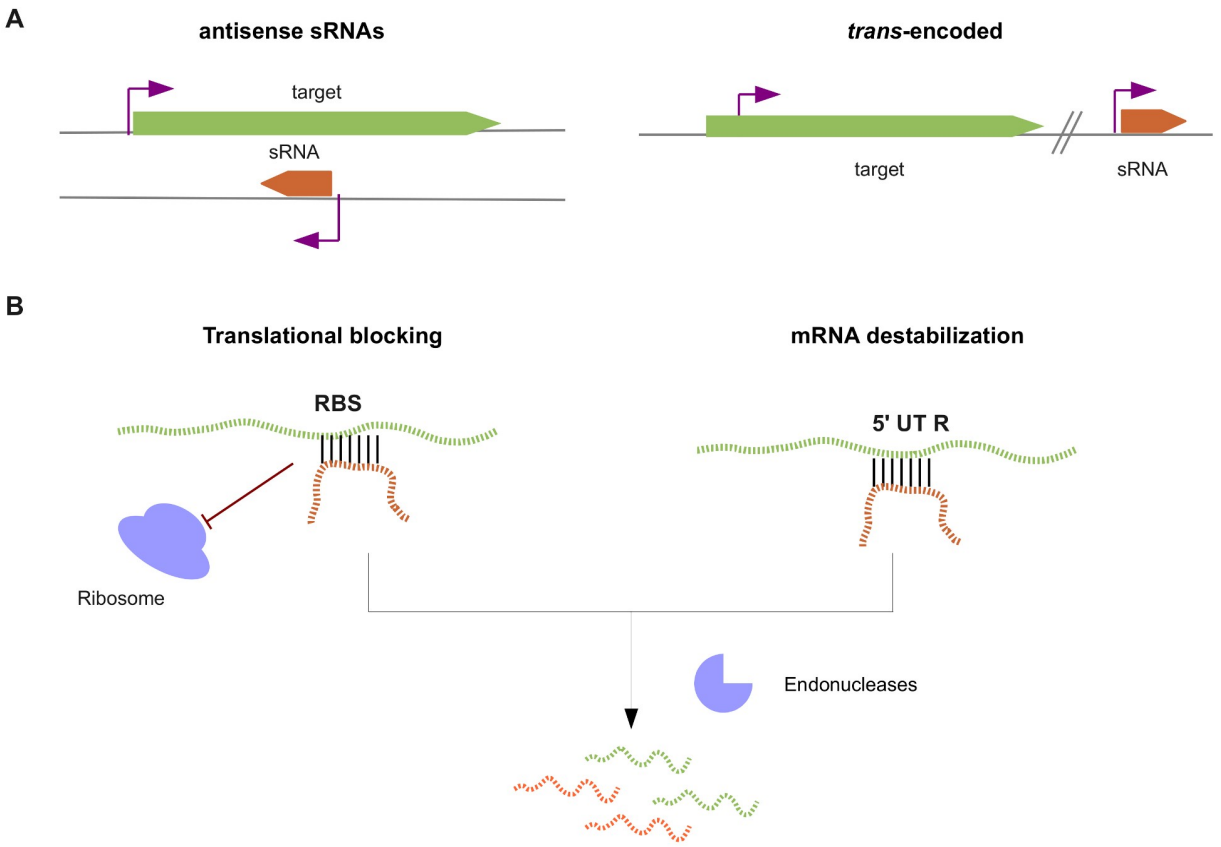

Figure 8: Types of bacterial sRNAs and their effects on gene expression. (A) Antisense sRNAs are transcribed opposite to coding regions while *trans*-encoded sRNAs are located in intergenic *loci* of the genome. (B) Mechanisms of gene repression triggered by sRNAs recognition of targets include translational blocking and mRNA destabilization.

The elucidation of the physiological function of several sRNAs in *Escherichia coli* revealed their importance in the context of larger regulatory networks. For example, the regulatory interplay between FurB regulator and the RyhB sRNA guarantees iron homeostasis in the bacteria. During normal conditions, Fe^{2+}-bound FurB represses the RyhB promoter, while during iron starvation this repression is alleviated and RyhB downregulates a multitude of non-essential genes that bind to iron (60). sRNAs also participate in the regulation of the nutrient homeostasis in the cell. Together with the GcvA regulator, the GcvB sRNA responds to the high levels of intracellular glycine by inducing the repression of several ABC transporters on the

membrane, decreasing the uptake of aminoacids of cells growing in rich media conditions (61).

3.2 Experimental identification of mycobacterial sRNAs

In mycobacteria, the first sRNAs were discovered by inspection of cDNA libraries prepared from low molecular weight RNA (62). Also, using a combination of cloning of cDNA libraries from RNA isolates and computational prediction to find putative candidates, DiChiara *et al* (63) reported the finding of 34 sRNAs in mycobacteria. Nevertheless, the adoption of deep-sequencing technologies as a tool to measure whole transcriptomes significantly increased the pace of sRNA findings, and large-scale screenings using RNA-seq platforms are discovering hundreds of sRNAs in bacterial species (64–67).

Recently, the first transcriptome-wide search for sRNAs in *Mtb* using RNA-seq combined to microarrays was conducted (68). Total RNA from *Mtb* H37Rv and BCG strains growing during exponential phase were fractioned by size (< ~200 nt) and depleted of ribosomal RNA. Next, RNA samples were subjected to Illumina sequencing, and the short reads (~100 nt-long) produced by the platform were analyzed by a custom data analysis pipeline in order to define the position of sRNA candidates in the genome. This pipeline was implemented by a series of steps, including: (i) mapping of reads to the reference genome allowing a maximum of 2 mismatches, (ii) selection of reads that match intergenic sequences only, (iii) filtering of reads based on abundance and genome conservation, and finally (iv) joining of contiguous read "peaks" to define sRNA candidates. Figure 9 illustrates the outcome of this mapping process for two candidate sRNAs – *c_1275611* and *c_3800005* – that were also detected by Northern blots. Figure 9A depicts the mapping of reads - that satisfy the abundance and conservation criteria - around the genomic region containing the candidate sRNA *c_1275611*. The boundaries of the

candidate sRNA were determined by joining two adjacent peaks in the region, which corresponds to a region of 62 nt. Northern blot validation confirmed the expression of the sRNA *c_1275611*, detecting a transcript ~40 nt-long (Figure 9B). Also, the putative sRNA *c_3800005* (mapped in a region of 87 nt), was detected in two different *Mtb* strains, in both exponential and stationary phases as a ~170 nt-long transcript (Figure 9C, 9D).

Additionally, microarrays containing custom-designed probes were used to confirm the expression of all sRNAs identified by RNA-seq, and a consensus list of candidates was generated. Altogether, the combined deep-sequencing and microarray analysis of the *Mtb* transcriptome revealed 466 asRNAs and 122 *trans*-encoded sRNAs encoded in the bacterial genome (68).

As a result, the genome-wide screening revealed that *Mtb* contains a vast post-transcriptional layer of gene regulation, mediated by a multitude of non-coding RNAs. However, little is known about the functional roles of these new sRNAs, including their contribution to *Mtb* virulence. To address this question, I developed computational methods to characterize the function of the newly found sRNAs in the regulation of gene expression in the bacteria. In order to achieve this goal, I studied separately the function of antisense sRNAs using enrichment analysis, and *trans*-encoded sRNAs using a target prediction methodology. The enrichment analysis identified several KEGG pathways and biological functions that are extensively regulated by antisense sRNAs.

On the other hand, the target prediction method was applied to infer a post-transcriptional regulatory network in *Mtb*, where regulatory links represent *trans*-encoded sRNAs base-pairing to mRNA targets to modulate their translation. Based on the knowledge of the network, *trans*-encoded sRNAs were predicted to be involved in several biological

functions, including nutrient transport and DNA transposition.

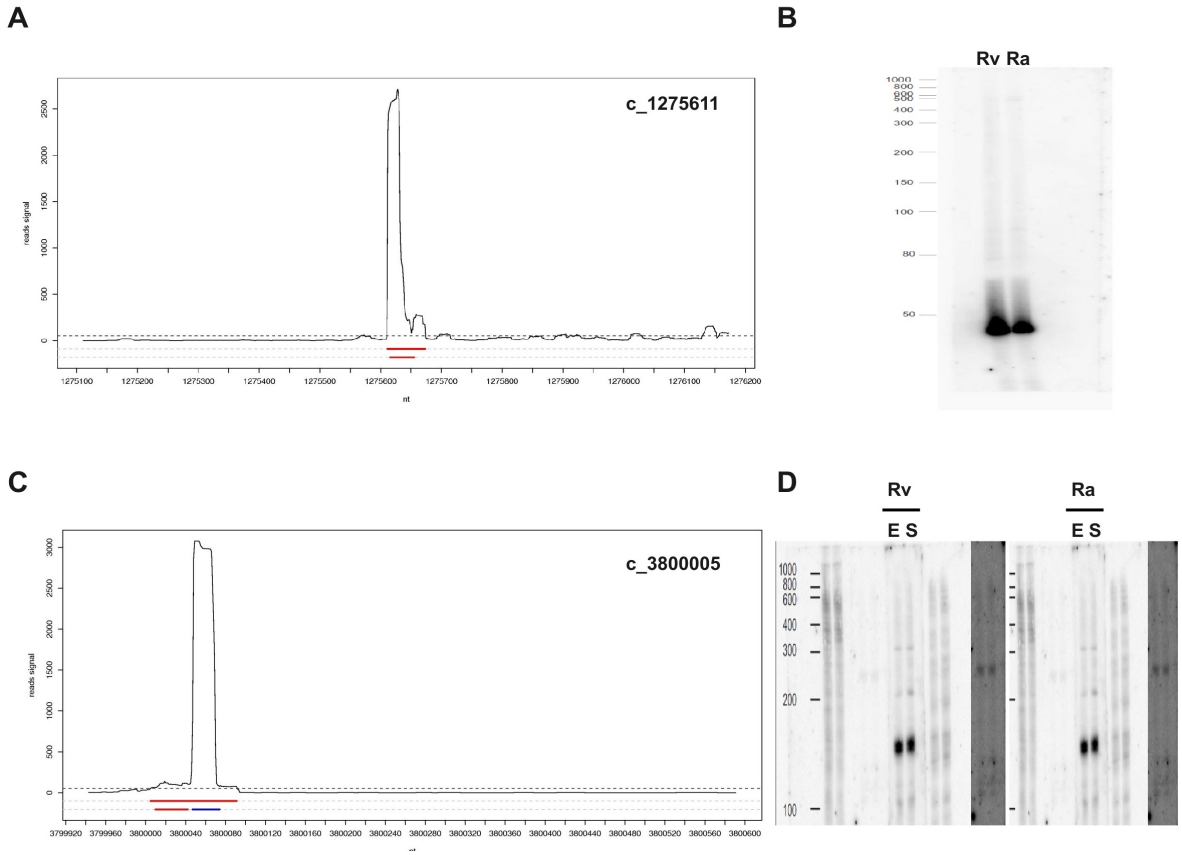

Figure 9: Genome-wide identification of Mtb sRNAs using RNA-seq and validation of selected candidates by Northern blots. (A) Genomic locus containing adjacent peaks of RNA-seq reads is harboring the candidate sRNA c_1275611. (B) Northern blot results confirmed c_1275611 in both H37Rv and H37Ra Mtb strains. (C) Genomic locus harboring the c_3800005 sRNA. (D) Northern blots confirmed expresion of c_3800005 in H37Rv and H37Ra in exponential and stationary phase (E and S, respectively). Experiments performed by Paolo Miotto, San Rafaelle Institute, Italy. Adapted from Miotto *et al* (68) (submitted for publication).

3.3 Current approaches for computational prediction of sRNA regulons

Understanding the effects of the sRNA-mediated regulation in mycobacteria requires the

identification of targets followed by the characterization of metabolic and signaling pathways that are under the control of sRNAs. Differently than asRNAs, *trans*-encoded sRNAs may act as key regulatory nodes in biological networks since they are capable of directly controlling the translation of multiple genes in the cell. For instance in *E. coli,* the RyhB sRNA that controls over 10 genes (52) and the sRNA GcvB that controls more than 20 genes (61), demonstrate the widespread effects of a single *trans*-encoded sRNA in bacterial cells.

The problem of predicting sRNA targets can be tackled in at least two distinct ways. In the hybridization-driven methods, represented by tools such as RNAhybrid (69) and TargetRNA (70), the tendency of base pairing between the sRNA and the target mRNA is quantified by the minimum free energy (MFE) of hybridization, disregarding the secondary structure of the interacting partners. These methods find the base-pairing region between the sRNA and mRNA sequences that yields the lowest hybridization energy (energetically most favorable). In structure-driven approaches, which include tools such as RNAup (71), RNAplex (72) and IntaRNA (73), the base pairing potential is measured by the MFE of binding, which is dependent on the secondary structures of the interacting molecules. Therefore, the binding MFE is computed as the sum of the hybridization energy (negative) and the energy input (positive) necessary to open both the sRNA and the target. In structure-driven approaches, exposed regions in secondary structures which are not internally base paired, such as hairpin loops and other single-stranded portions of the RNA molecule, are more likely to participate in the duplex formation since they require less energy input to become binding sites. These approaches were motivated by the the observation that the efficiency of mammalian siRNA-based repression is proportional to the exposure of the binding site in the particular mRNA conformation, being that binding sites buried in the mRNA reduced siRNA repression (74). In order to model this phenomenon, structure-driven approaches search for interactions considering that both sRNA

and mRNA interact using their most accessible binding sites, as determined by their secondary structures.

However, sRNA target prediction has been shown to be a challenging task prone to identify a high number of false-positives for several reasons. First, as explained above, a successful sRNA-mRNA interaction depends on the secondary structures of both the sRNA and mRNA molecules. Yet, the thermodynamical rules of RNA-RNA binding are poorly known, and therefore the estimation of binding energies are only approximate. Another difficulty arises from the fact that *trans*-encoded sRNAs recognize targets by short and imperfect base-pairings, often containing non-Watson Crick pairings and bulge loops (analogous to gaps in sequence alignment). Ultimately, this flexibility in base-pairing rules lead to thousands of potential binding sites in the transcriptome, increasing the difficulty in finding true sRNA targets.

In such scenario, computational methods are necessary to help select the most promising sRNA-mRNA base-pairings out of all putative base-pairings in the transcriptome. To address this problem, I developed a target prediction method that implements three steps for selecting the most promising sRNA-mRNA base-pairings: (*i*) enumeration of all possible binding sites of the sRNA in the transcriptome, (*ii*) scoring of each binding site based on the MFE of binding and location, (*iii*) selection of top-scoring binding sites by filtering out interactions that are not better than random. The interactions that survive the non-random cutoff are selected to composed the final post-transcriptional regulatory network. Below I provide a detailed description of the computational method.

3.4 Inferring targets for the novel *trans*-encoded *M. tuberculosis* sRNAs

I developed a methodology for sRNAs target prediction in *Mtb* that integrates *a priori*

knowledge of possible binding sites with structure-driven identification of binding sites. The method starts by applying RNAup to find binding sites between a particular sRNA and all annotated genes in *Mtb*. RNAup predicts regions of base pairing based on structure accessibility, and its folding algorithm allows RNA molecules to assume most of the structural motifs found in secondary structures. Each target is represented by a sequence of 300 nucleotides (100 bp upstream + 200 bp downstream of the start codon), which contains 5' UTR elements involved in sRNA recognition. Then, given the *trans*-encoded $sRNA_j$, RNAup finds the minimum free energy (MFE) of binding between $sRNA_j$ and every possible target sequence in the *Mtb* H37Rv genome (4047 sequences, representing all genes).

Subsequently, each sRNA-mRNA interaction is evaluated by the binding score (*b-score*) defined as

$$b-score\left(sRNA_j, mRNA_k\right)=MFE \ * \left\langle f_{bs}\left(a_1,\ldots,a_n\right)\right\rangle \tag{7}$$

where MFE is the negative total free energy of the $sRNA_j - mRNA_k$ binding, $a_{1,}\ldots a_n$ corresponds to the location of the binding site in the mRNA (relative to the translation start site), and $\left\langle f_{bs}\left(a_1,\ldots,a_n\right)\right\rangle$ is the average frequency of binding to positions $a_{1,}\ldots a_n$. The nucleotide binding frequencies were obtained from a list of experimentally verified mRNA binding sites from *E. coli* and *Salmonella* (Table 1, histogram of Figure 10A).

Thus, the *b-score* involves an energy component that quantifies how energetically favorable is the duplex formation at the given region (the lower the MFE, the more favorable is the dimer formation), and a location component based on information about already validated

binding sites (Figure 10A). This scoring scheme was designed to prioritize sRNA-mRNA interactions with low MFE of binding that occur in regions of the mRNA target known for their tendency to interact with sRNAs.

The p-value associated with a specific *b-score* is estimated by comparing the observed *b-score* for the $sRNA_j - mRNA_k$ interaction to a reference cumulative distribution generated using 1000 random target mRNA sequences. The cumulative distribution F_{sRNA_j} is constructed by assigning equal probability to each random *b-score*, and linear interpolation is used to transform the discrete distribution into a continuous function in the interval [0,1]. The random mRNA sequences were generated using the dinucleotide frequencies (AT, AA, AC, and so on) learned from the real set of 5' UTRs extracted from the *Mtb* H37RV genome.

Using the cumulative distribution F_{sRNA_j} , I obtained the p-value of associated to a random *b-score* as given below:

$$p-value = 1 - F_{sRNA_j}\left[b\left(sRNA_j, mRNA_k\right)\right] \tag{8}$$

The method correct p-values for multiple hypothesis testing by computing their correspondent false discovery rates (FDR) using the Benjamini-Hocheberg approach (75). Finally, the $sRNA_j$ regulon is formed by interactions that satisfy the specified FDR cutoff, when applied to the set of all possible bindings (Figure 10B).

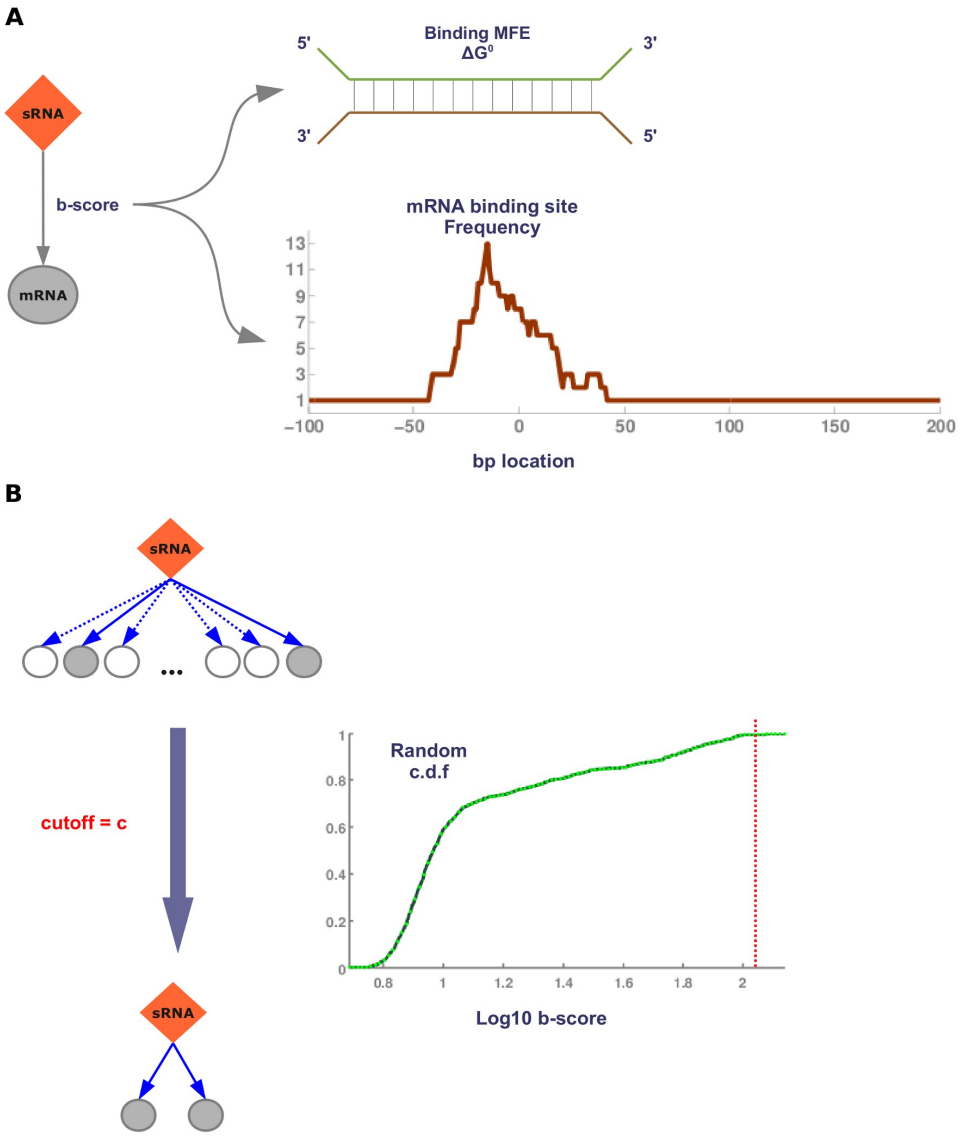

Figure 10: Computational approach to infer sRNA targets. (A) The *b-score* combines two critical aspects of the sRNA-mRNA interactions, namely MFE of binding based on binding site accessibility (top) and location of binding based on *a priori* information of binding sites (bottom). Position 0 refers to the translational start site. (B) sRNA regulons containing only interactions with significant *b-scores* (bold blue lines) are selected by applying a FDR cutoff (red line) to the cumulative distribution of binding scores between a particular *trans*-encoded sRNA and random target sequences.

sRNA	mRNA target	Region of interaction
DiscF	FtsZ	-28 to +2 (a)
DsrA	Hns	+7 to +19 (a,b)
DsrA	RpoS	-119 to -97 (a)
MicC	OmpC	-41 to -15 (a); -30 to -15 (b)
MicC	OmpD	+23 to +26 (c)
MicF	OmpF	-16 to -10 (a,b)
OxyS	FhlA	-15 to -9 (a); +34 to +42 (a)
RprA	RpoS	-117 to -94 (a)
RyhB	Sdh	-42 to -3 (a)
RyhB	SodB	-17 to +9 (a); -4 to +5 (b)
Spot42	GalK	-19 to +39 (a); -19 to +21 (b)
MycA	OmpA	-21 to -6 (b)
SgrS	PtsG	-28 to +4 (b)
GcvB	DppA	-31 to -14 (b)
GcvB	OppA	-8 to +16 (b)

Table 1: Validated mRNA binding sites bound by *trans*-encoded sRNAs. Extracted from (a) Tjaden *et al* (70), (b) Tafer and Hofacker (72), and (c) Pfeiffer *et al* (59).

3.5 Functional enrichment of sRNA sub-networks

I studied the function of the inferred sRNA regulons by finding over-represented biological functions within these subnetworks, by the means of the Fisher's Exact Test (FET). I downloaded the functional annotation for *Mtb* genes from the DAVID database (76), which included information about GO terms and INTERPRO protein families and domains. Then, I applied the FET to determine whether the regulon is enriched for a specific term, based on the contingency table shown in Table 2.

Under this setup, the FET uses a hypergeometric distribution to compute the probability to observe a sub-network containing $a+b$ members of which exactly a genes are annotated with the term j, being that genes are drawn from the total N genes within the category (either GO or

INTERPRO). Also, I slightly modified the FET by removing one gene from the list of interest (i.e. *a-1*) before computing the Fisher exact probability; this modification removes the effect of terms based on single genes only (77). The FET p-value is found by summing up probabilities of the observed table and every other stronger table increasing *a* and decreasing *b*, until *b* reaches 0.

Term j	(+)	(-)	Total
in-regulon genes	a	b=a'-a	a+b
out-regulon genes	c	d=c'-c	c+d
Total	a+c	b+d	N

Table 2: Contingency table used in functional enrichment analysis of predicted sRNA regulons. *a*=number of genes in the regulon annotated with term *j*; *a'*=number of regulon genes annotated within the category of term *j*; *b*=number of regulon genes not annotated with term *j*. The quantities *c*, *c*' and *d* are similarly defined for genes outside the sub-network. *N* is the total number of genes for which annotation is available.

3.6 sRNA-mediated post-transcriptional regulatory network

The *b-score* methodology was used to predict targets for the 122 *trans*-encoded *Mtb* sRNAs discovered by deep-sequencing and confirmed by microarrays. Figure 11A depicts an overview of the inferred sRNA-mediated regulatory network, produced using the FDR cutoff of 0.01. At this level of significance, the network comprises 941 interactions involving 96 sRNAs and 696 unique target genes. In this bipartite post-transcriptional network, regulatory links represent the structure-driven base pairing of sRNAs to their mRNA targets, in order to inhibit their translation. The complete sRNA-mediated regulatory network is available for download at the following website: *https://docs.google.com/spreadsheet/ccc?key=0Aue_AzEGF-5SdHdhWnZEZU0yVmFYU014SVI3T0doVFE.*

To gain insight into the function of the newly predicted sRNA-mRNA interactions, I looked for molecular processes (Gene Ontology terms) and protein domains (INTERPRO terms) over-represented within sRNA regulons. The results of the enrichment analysis are shown Figures 12 and 13. The analysis found that 18 sRNA regulons (out of 96 subnetworks), were associated with a specific GO or INTERPRO term using a stringent level of significance (0.01). For example, the *c_3557057* sRNA was predicted to regulate two glutamine transporters *Rv0073* and *glnQ*, the last one belonging to the ABC transporter family (Figure 11B). The annotation of these transporters include the terms IPR002373 (cAMP/cGMP−dependent protein kinase) and GO:0008603 (cAMP−dependent protein kinase regulator activity), since these transporters contain a kinase domain regulated by cAMP levels. Another interesting case is the *c_806185* sRNA (Figure 11C), which was predicted to control four genes (*Rv1756c*, *Rv3380c*, *Rv1369c*, and *Rv1764*) involved in transposition of insertion sequences (terms GO:0032196 and IPR001584). Such bacterial transposases belong to mobile genetic elements known as transposons, and are capable of promoting genomic diversity by self-inserting themselves in random positions of the genome (78). Additionally, the transposases *Rv1756c* and *Rv1764* are encoded in the same locus of the genome – the RD152 (region of differentiation) -, which is deleted in hypervirulent Beijing strains (79). Therefore, the putative down-regulation of these transposases by the *c_806185* sRNA may also cause a virulent phenotype.

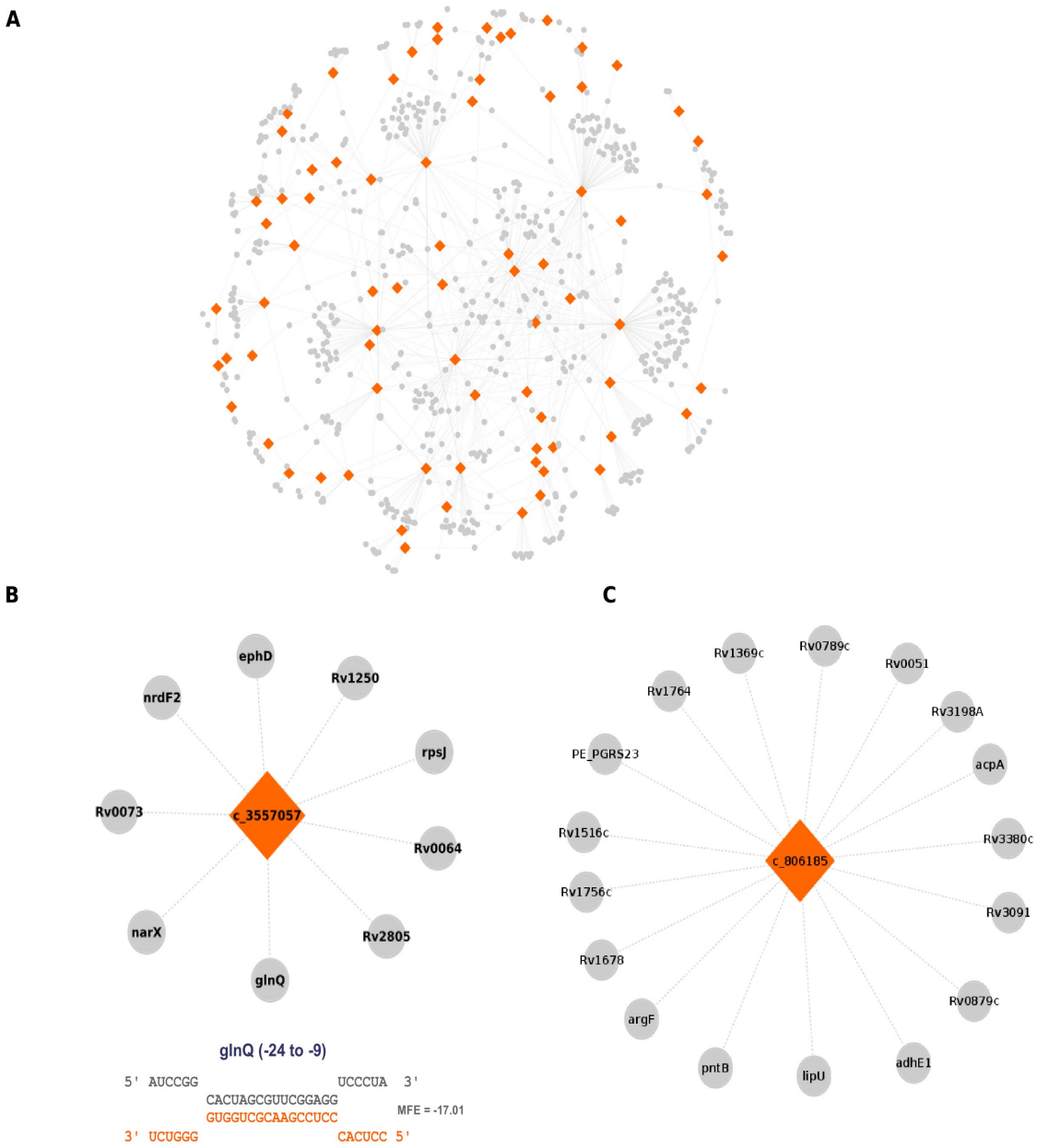

Figure 11: Predicted sRNA-mediated post-transcriptional regulatory network. (A) Overview of the network at FDR 0.01, involving 96 sRNAs (orange nodes) and 696 target genes (gray nodes). (B) The c_3557057 sRNA regulon contains 9 target mRNAs, including two glutamine transporters Rv0073 and glnQ. The binding between c_3557057 and glnQ is predicted to occur upstream of the translation start site (-24, -9) in the target (bottom). (C) c_806185 regulon is enriched in bacterial transposases.

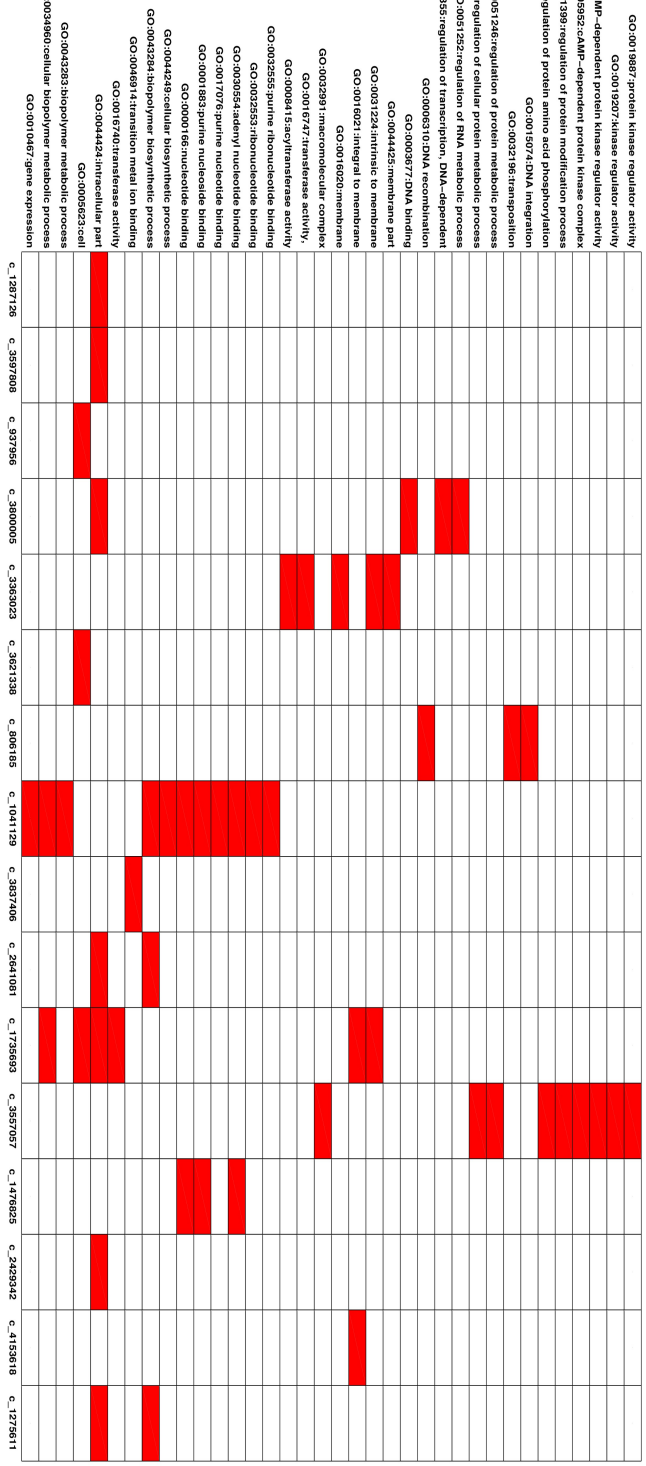

Figure 12: Over-represented GO terms within predicted sRNA regulons (p-value 0.01, Fisher's exact test).

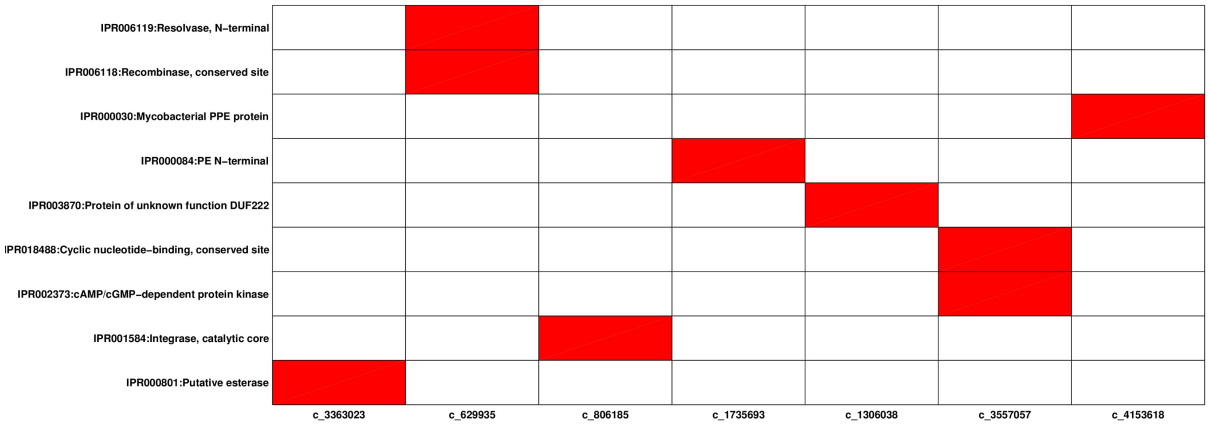

Figure 13: Over-represented INTERPRO terms (protein domains) within predicted sRNA regulons. Fisher's exact test (P=0.01).

3.7 Functional analysis of *M. tuberculosis* genes regulated by antisense sRNAs

The RNA-seq data revealed the existence of 466 asRNAs that are transcribed in a genomic region opposite to *Mtb* genes. These asRNAs overlap 407 unique genes (10.05% of the *Mtb* genome), which are distributed throughout several Tuberculist functional classes (Figure 14). A large proportion of asRNAs are overlapping conserved hypothetical proteins, metabolism and respiration and cell wall related genes.

I conducted an enrichment analysis to uncover KEGG pathways that are subjected to asRNAs regulation. Specifically, overrepresented KEGG pathways were obtained by a Fisher's exact test, using the contingency table shown in Table 3. The analysis unveiled that several key pathways, such as the respiratory pathway (oxidative phosphorylation, p-value 0.0005), are extensively composed of asRNAs-regulated genes (Table 4). It is known that *Mtb* can adapt its respiration (energy production) in response to the host environment. In the absence of oxygen, the bacteria utilize alternative respiratory pathways to carry out an anaerobic electron transport

using different electron acceptors such as nitrate, nitrite and fumarate. At the onset of acquired (Th1) immune response, the bacteria downregulates the aerobic components of the electron transport chain (including the ATP synthase complex), and upregulates genes related to microaerobic respiration (such as *cydA*), as well as genes involved in nitrate reduction and transport due to availability of nitrate (18; 80). Remarkably, four subunits of the ATPase synthase complex (*atpA*, *atpB*, *atpD*, *atpG*) appear to be regulated by asRNAs. Hence, this suggests that asRNAs may be involved in the switch among respiratory states that confer the ability of intracellular survival. Further studies could address whether such respiratory asRNAs are induced in response to the immune system. Interestingly, ribosome components were also enriched for asRNA regulation (p-value 0.03, Table 4). This may indicate that the bacteria might induce production of ribosome asRNAs in conditions of growth arrest through translational inhibition, and therefore may be important in dormancy-like conditions.

Furthermore, many asRNAs were found to be overlapping global regulators such as sigma factors (*sigA*, *sigB*, *sigH*), anti-sigma factors (*rshA*) as well as trancription factors (*ideR*, *dosR*, *rv0275c*, *rv0792c*, *rv2657c*, *rv3291c*, *rv3583c*, *rv3736*, *rv1830*). Other noteworthy biological processes that are under asRNA-mediated regulation are sensor and signal transduction (*e.g. mprA*, *devS*, *senX3*), cell division control (*ftsK*, *rv0284*, *rv1784*), DNA repair (*polA*, *rv3586*, *uvrA*, *uvrB*) and bacterial secretion system (*rv3291c*, *secE*, *secF*).

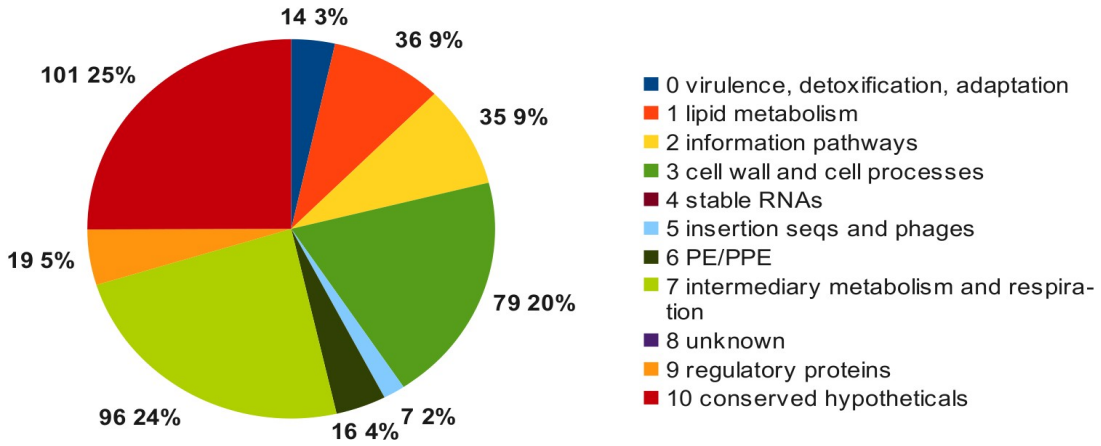

Figure 14: Classification of antisense overlapped genes in *Mtb* among Tuberculist categories.

Antisense overlapped	(+)	(-)	Total
in-pathway	a	b	a+b
out-pathway	c	d	c+d
Total	a+c	b+d	N

Table 3: Contingency table used in the pathway enrichment analysis. *a*=number of antisense overlapped genes belonging to pathway *j*; *b*=number of non-antisense overlapped genes annotated in pathway *j*. The quantities *c*, *d* are similarly defined for out of pathway genes. *N* is the total number of genes for which KEGG annotation is available.

Pathway	Genes	Total genes	P-value (Fisher's exact test)
mtu00190 Oxidative phosphorylation	*cydA,Rv1812c,cydB,ndh,ppa,sdhA,atpG,atpD,atpA,atpB,qcrB,ctaE,qcrA,qcrC,nuoC*	15	0.0005
mtu00903 Limonene and pinene degradation	*cyp125,echA3,cyp126*	3	0.0065
mtu00632 Benzoate degradation via CoA ligation	*sdhA,fadA2,fadA4,echA3*	4	0.0286
mtu03010 Ribosome	*rpmI,rpsA,rplT,rpsT,rpsF,rpsB,rpsD,rpsS,rplV,rpmC,rplP,rpsQ,rrl*	13	0.04
mtu00350 Tyrosine metabolism	*Rv0089,adhB*	2	0.04
mtu00626 Naphthalene and anthracene degradation	*Rv0089,cyp125,cyp126*	3	0.0425
mtu00340 Histidine metabolism	*hisH,Rv0089,*	2	0.04
mtu00280 Valine, leucine and isoleucine degradation	*lpd,fadA2,fadA4,echA3*	4	0.06
mtu00380 Tryptophan metabolism	*fadA2,fadA4,echA3*	3	0.07
mtu02020 Two-component system	*mprA,pepD,narG,pstS2,kdpD,fadA2,fadA4,senX3,narH,devS,dosR*	11	0.08

Table 4: KEGG pathways enriched in genes regulated by antisense sRNAs.

3.8 Discussion

I presented herein a computational approach to infer sRNA targets that combine two aspects that may be critical for establishment of successful sRNA-mRNA interactions. First, the method utilizes structure-driven identification of binding sites, as predicted by RNAup. As mentioned earlier, it has been shown that eukaryotic siRNA preferentially target exposed regions in the secondary structures of mRNAs. Assuming that RNA interaction mechanisms is structure-driven, this mode of recognition may also be important to predict interactions in prokaryotic mRNAs. Furthermore, bacterial sRNAs are longer than siRNAs and can assume complex secondary structures. Then, the accessibility-based selection of the binding region within the sRNA may be relevant for target inference, even though no experimental study demonstrated yet that the folded structure of bacterial sRNAs drives target recognition.

The location of binding also appears to be important for the establishment of sRNA-mRNA interactions. Analysis of several validate binding sites curated from the literature suggested that *trans*-encoded sRNAs tend to bind to a region covering ~40 nucleotides around the start codon. This underscores the hypothesis that *trans*-encoded sRNAs compete with ribosomes to bind to the Shine-Dalgarno sequence (55), and hence location of binding is critical for their roles of blocking and/or interfering with translation. The method used this piece of information to attribute weights to each base pairing identified by RNAup, and the *b-score* was utilized to select the most promising interactions.

However, one major problem of using data generated by RNA-seq assays is the lack of precise boundaries of the sRNA transcript. The lengths obtained after mapping of RNA-seq reads usually disagree with the ones observed in Northern blots (as shown in Figure 9). Clearly, the lack of precise knowledge of the size of novel sRNAs impact downstream applications such as target prediction. A possible solution is to use large-scale mapping of transcript ends such as high-throughput RACE (81) to define the 5' and 3' ends of candidates generated by deep-sequencing.

The deep-sequencing of *Mtb* transcriptome revealed a staggering number of non coding RNAs, and the characterization of the physiological roles of these sRNAs will be a time-consuming process. Despite all limitations, the proposed method can be used as a preliminary tool to quickly study the role of sRNAs, and later prioritize candidates for more detailed studies based on their involvement in key processes such as virulence and host survival.

Chapter 4

Data Integration by Rank-Conciliation Applied to Small RNA Target Inference

4.1 Current limitations for studying small RNAs regulation

Even though new sRNAs can be discovered in a high-throughput fashion, the rapidly expanding list of sRNAs does not advance our understanding of their regulatory role. Both experimental and computational problems are hampering the study of sRNA regulation in bacteria. On the experimental side, what is missing is a large-scale approach to validate direct mRNA targets. Experiments such as transient sRNA over-expression followed by whole transcriptome or proteome quantitation (53; 82) generate an extensive list of potential targets, but the proposed interactions require further characterization to prove direct binding. Hence, target validation needs to be performed case by case, for instance by seeking decreased target levels in Northern blots after sRNA induction, and eventually showing the specificity of target binding by compensatory mutations. However, this process is time-consuming when hundreds of potential targets need to be tested for each new sRNA.

One remedy may be to shorten the list of potential targets by computational approaches, but target prediction has also proved to be difficult. One difficulty arises from the fact that *trans*-encoded sRNAs interact via short and imperfect base pairings, leading to thousands of putative

targets in the transcriptome. Also, as mentioned before, the sRNA-mRNA interaction *in vivo* depends on the secondary structures of both RNA molecules. Recent computational approaches consider the accessibility of secondary structures, and are able to predict binding sites with higher accuracy (71–73). Still, the lack of base pairing rules and the fact that base pairing is dependent on the accessibility of the folded structures make the prediction of sRNA-mRNA interactions a challenging task, generating a considerable number of false positives.

Additional resources, such as publicly available gene expression compendia may contain useful information to improve target predictions. A hallmark of the sRNA-mediated regulation is the over-expression of the sRNA compared to its targets during inducible conditions. For example, upon iron starvation, the RyhB promoter activity is ~4 times higher than its mRNA targets, while this ratio is even higher (~18) between Spot42 sRNA and its target galK when glucose is present (83). Given this unbalanced production, one may expect to observe anti-correlation due to the accumulation of sRNAs compared to mRNA levels, even considering that typically sRNAs are co-degraded with targets (84). Then, microarray data spanning a sufficiently large number of conditions may reveal this interdependence between sRNA and mRNA levels.

Furthermore, knowledge of binding site position within the sRNA and mRNA sequences may be informative to assess putative interactions. Bacterial sRNAs are longer than their eukaryotic counterparts (on average ~ 100 nucleotides) and fold into complex secondary structures containing active regions that are responsible for target binding. The existence of an autonomous domain sufficient for target recognition has been characterized for RybB (85). Later, a theoretical study based on large collection of binding sites demonstrated that these binding regions tend to be more conserved and accessible than the rest of the sRNA molecule (86). Therefore, restricting the base pairing to binding regions may improve the predictability of

true interactions. On the other side, binding site position within the mRNA sequence may also be an indicator that the interaction is occurring *in vivo*. The majority of *trans*-encoded sRNAs preferentially bind in the vicinity of the start codon of the target (55), suggesting that transcript-binding in this location is necessary for blocking translation (by preventing ribosome loading onto the Shine-Dalgarno region) or interfering with translation initiation (by binding shortly downstream of the translation start site). Consequently, the position of binding regions on both molecules may determine whether the sRNA-mRNA interaction is effective, and this information can be used to predict novel meaningful interactions.

Methods that integrate multiple data types may facilitate the discovery of post-transcriptional regulation in bacteria, by overcoming the limitations of target prediction relying on a single computational tool. Here I propose a computational method to infer sRNA targets – called rank-conciliation -, that integrates three different data types for prediction: MFE of sRNA-mRNA interactions, negative correlation in large microarray compendia and the location of the predicted binding site relative to the translation start site. At a glance, the method creates rankings for each data type, and a final ranking - containing the most likely targets – is produced by merging the information from all rankings. As detailed below, I used this method to rank-order potential targets of several *trans*-encoded sRNAs in *E. coli*. The method was shown to improve the quality of sRNA target prediction for most sRNAs.

4.2 Types of evidence for sRNA target prediction

The approach developed integrates multiple data types (herein called evidences) in order to rank-order sRNA-mRNA base pairings in the transcriptome. Below I describe the three types of evidence used for target prediction.

The first type of evidence considered was the MFE of interaction resulting from the interaction between a given sRNA and target mRNA (Figure 15A). I used different RNA-RNA prediction tools – RNAup (71), RNAplex (72), and RNAduplex (87) - to find the MFE of interaction between a given sRNA and every mRNA target sequence with mapped transcription start site in *E. coli*. Each tool utilizes a distinct energy model for computing the MFE of interaction. For instance, both RNAup and RNAplex compute the MFE of interaction based on the accessibility within the folded mRNA, while RNAduplex computes the MFE after hybridization, which is not based on secondary structures.

The second type of evidence was the negative Pearson correlation between a particular sRNA and each target gene, as found in gene expression data. This evidence was included based on recent results by Mitarai *et al* (83) demonstrating an inverse relationship between sRNA synthesis rate and mRNA levels is a characteristic of antisense regulation. They showed that the dynamics of the sRNA-based regulation is dictated by two parameters, namely the ratio of sRNA to mRNA production (α) and the rate of base pairing (i.e. speed of target inactivation, γ). I tested this model to gain insights on how varying levels of sRNA expression may affect mRNA levels, by simulating both sRNA and mRNA levels in two scenarios, assuming that the sRNA is: (i) expressed at arbitrary level ($\alpha > 0$), or (ii) always over-expressed ($\alpha > 1$). Figure 15B shows the distribution of cross-correlation (computed between sRNA and mRNA levels at steady state) for these two scenarios. The results show anti-correlated sRNA-mRNA expression during sRNA induction when $\alpha > 1$ (median=-0.39, IQR=0.08), but also for arbitrary sRNA expression, when $\alpha > 0$ (median=-0.30, IQR=0.02). This indicated that anti-correlation should be observable in large expression datasets that would correspond to $\alpha > 0$ (since it includes conditions where the sRNA is not induced). I order to check this property, I computed the correlation between every

known sRNA-target gene pair, using gene levels extracted from a large compendium of *E. coli* microarray data spanning 466 conditions (88). Indeed, as shown in Figure 15B (boxplot at bottom), the expression profiles of known sRNA-mRNA pairs tend to be anti-correlated (median=-0.18, IQR=0.36). The anti-correlation computed from microarray data is less pronounced, and corresponds to the the situation where $\alpha > 0$ (microarray includes conditions where the sRNA is uninduced). Figure 15C depicts anti-correlation between the expression of sRNAs *micC* and *rybB* and their known targets *ompC* and *tsx*, respectively. Altogether, it is reasonable to assume that anti-correlation computed from gene expression datasets may serve as evidence for interaction between the sRNA and its putative mRNA target.

The third piece of evidence was the likelihood of the binding site position (predicted by RNAup) on the mRNA target sequence relative to the translation start site. To evaluate how frequently specific positions in the mRNA may be targeted, a list of 53 experimentally validated mRNA-binding regions was collected from the literature (86), and used to construct an empirical probability density function ($p.d.f$) for the binding site location (Figure 15D). The binding probabilities of each nucleotide in a region of 100 nt around the start codon was performed using kernel density estimation (normal distribution with width 1.2). Considering that the predicted base pairing covers the positions n_1 to n_l , the binding site likelihood is given by the area under the probability density function spanning the region $n_{1,}\ldots n_l$

$$L(n_{1,}\ldots,n_l)=\int_{n_1}^{n_l} p.d.f \qquad (9)$$

Therefore, this *p.d.f* contains the probability of each nucleotide position in the mRNA to be targeted by sRNAs. This binding site distribution reinforces previous claims (55) that productive binding (i.e. binding that leads to mRNA degradation) occurs preferentially within or near the ribosome binding site (peaking around the -10 box) as well as in the beginning of the coding sequence (~ 25bp downstream TSS). However, it is also useful to ascribe probabilities for less common base-pairings, such as in the (+25,+50) region, farther downstream from the start codon.

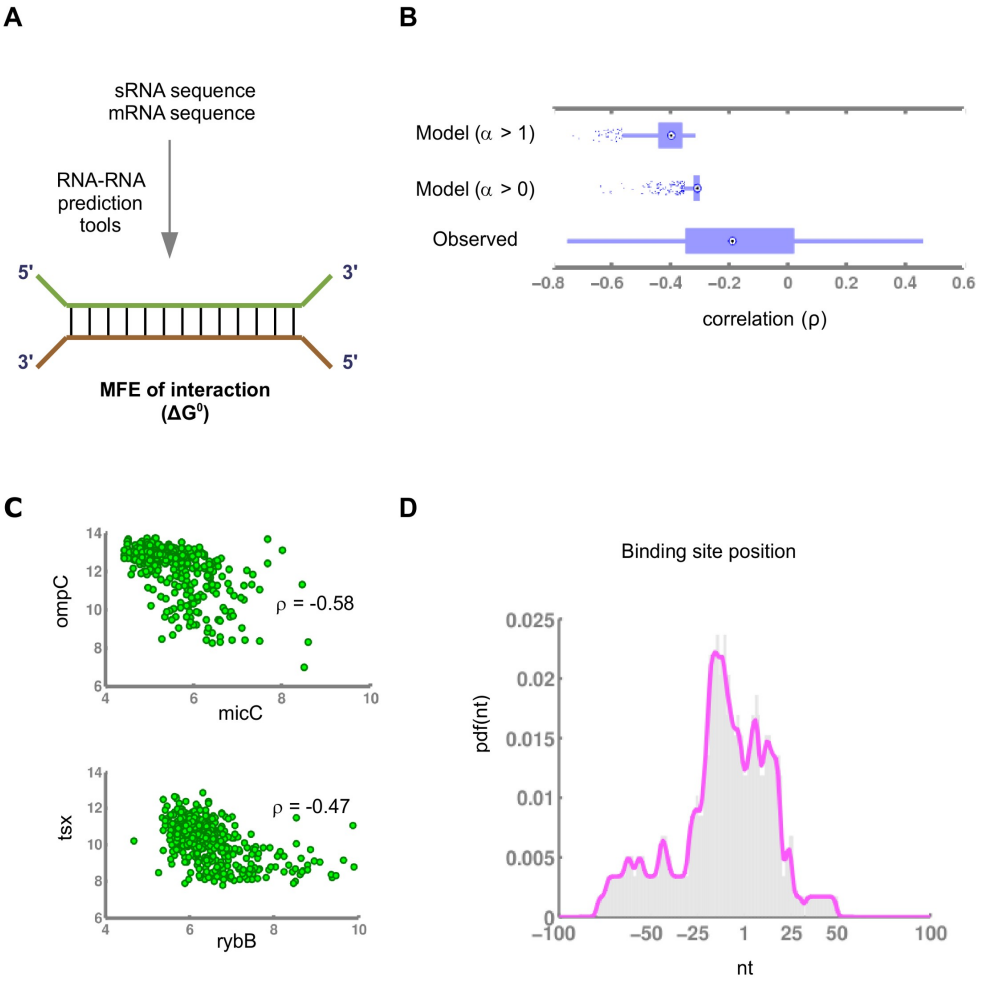

Figure 15: Types of evidence used to evaluate whether sRNA regulation occurs in vivo. (A) RNA-RNA prediction tools find the optimal base pairing between sequences by minimizing the MFE of interaction; the lowest the MFE the more likely is the sRNA-mRNA duplex formation. (B) Boxplots denote distribution of correlations between sRNA and mRNA levels, computed from simulated data and gene expression data. (C) Strong negative correlation between known sRNA-target pairs micC-ompC (p-value < 10^{-5}) and rybB-tsx (p-value < 10^{-5}) across hundreds of microarrays. (D) mRNA binding p.d.f estimates the probability of each nucleotide position in the mRNA to be involved in the base pairing, as found in a catalog of repression interactions experimentally validated in *E. coli* and *Salmonella*. No interactions are documented outside the (-100,100) boundary (position 1 refers to TSS).

4.3 Rank-conciliation to integrate multiple evidences

In order to integrate all types of evidence described above for target prediction, the methodology starts by transforming each evidence in a *rank-ordered* list of potential sRNA targets. The

ranking is generated on the basis of a specific criterion - such as lowest MFE of interaction – such that the most likely targets are expected to get top ranks.

However, various types of evidence will most likely assign different ranks for the same target, and therefore different types of evidence will disagree among each other in their ranking of targets. Moreover, the various types of evidence will also differ in their relevance for sRNA target prediction. Therefore, given a set of validated interactions, the method implemented a rank-conciliation approach that (*i*) attributes weights to each type of evidence utilized, suggesting their relative importance for predicting direct sRNA-mRNA interactions, and (*ii*) assign a reconciliated rank for each sRNA-mRNA interaction, which conveys the weighted contribution of each evidence (data type). Then, to create an integrative ranking that includes all evidences with appropriate weighting, the method proceeds as follows.

First, a ranking for each type of evidence is created by sorting sRNA-mRNA base pairings (from the best to the worst). Thereby three independent ranks were obtained for each base pairing: the rank associated with the MFE of interaction (lowest to highest), the rank associated with correlation (lowest to highest), and the rank associated with the likelihood of the mRNA binding site position (highest to lowest). Second, the method computes the weighted rank products (WRP) for the sRNA-mRNA interactions and used them to conciliate all the ranks associated with various types of evidence.

The rank products were introduced as a metric to rank order significantly expressed genes across independent replicates (89). Here I apply a modified rank products to rank order sRNA base pairings. Considering that the particular sRNA-mRNA base pairing occupy positions $r_{1,}\ldots r_k$ in k evidence rankings, the WRP is given by the weighted geometric mean

$$RP_w(sRNA, mRNA) = \prod_{i=1}^{k} r_i^{w_i} \quad (10)$$

where $w_1, \ldots w_k$ are the weights of each evidence rank, and $(w_1 + \ldots + w_k) = 1$. The WRP preserve the internal ordering of the ranking but assigns a different importance for the rank.

The weights represent the contribution of each evidence for predicting known targets, and are learned from data (see below). Finally, the WRP values are sorted in ascendent order to generate an overall output rank, which defines the final rank of each sRNA-mRNA interaction.

This rank-conciliation approach has an analogue in the field of multicriteria decision making, in situations when several judges are designated to rank alternatives, and a combined ranking must be produced. Considering that judges have variable competence to assess these alternatives, each judge's ranking carries an importance (weight), and the combined ranking is produced using weighted geometric mean (90).

Figure 16A illustrates the rationale of the approach in a list of three base pairings (depicted in green, yellow, and red), separately ranked by three independent criteria. In a conventional rank products (RP) ranking, the final position is given by the geometric mean of each rank position, and consequently each ranking contributes equally to the final ordering. Thus, the red base pairing in Figure 16A (ranks 3rd, 1st, 1st and RP equals to 1.44) appears in the top of the unweighted RP-based ranking. On the other hand, in the WRP sorting, rankings contribute differently to the final ordering. For instance, by assigning weights 0.8, 0.1, and 0.1 to rankings A, B, C respectively, the green base pairing occupies the top of the list (ranks 1st, 3rd, 3rd and WRP equals to 1.24). Assuming that the green base pairing is known to occur, the WRP-based ordering is able to identify ranking A as the most effective criterion to recover targets. Therefore, by adjusting the rank weights the method is able to find the configuration that best

classifies known targets, and then apply weights to rank order the rest of the base pairings.

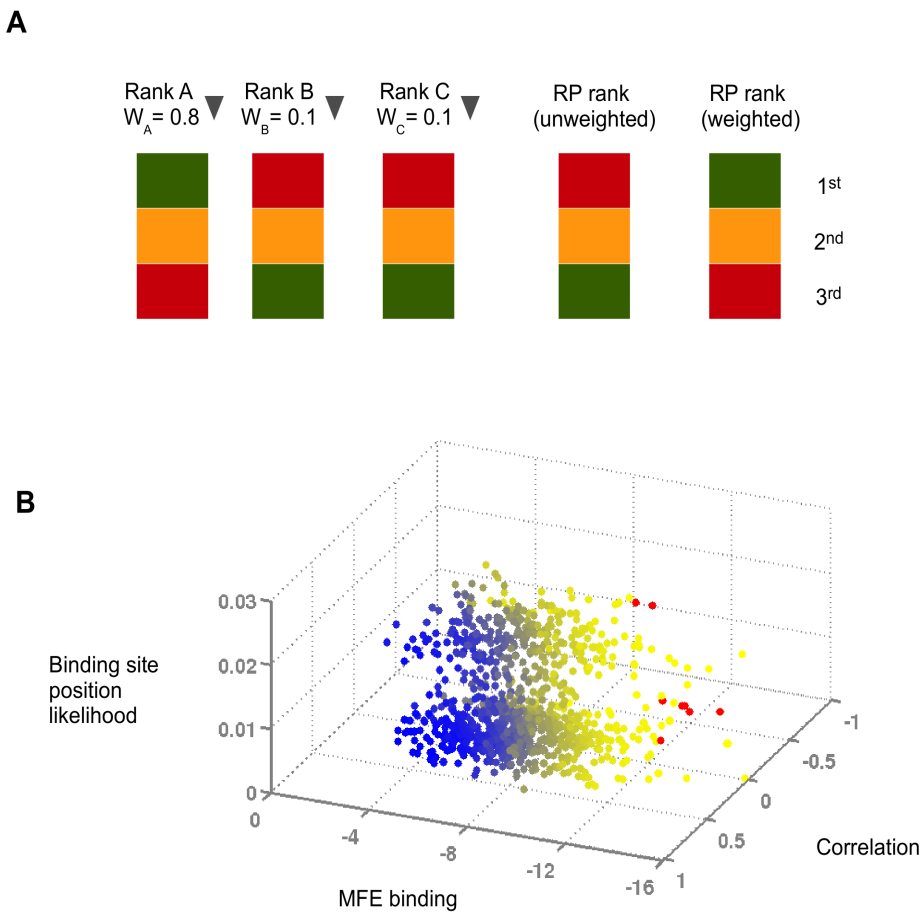

Figure 16: Rank-conciliation approach to combine evidences for predicting sRNA-mRNA regulation. (A) Scheme depicting how rankings are combined using the unweighted/weighted rank products statistic; the final ordering is determined by weights assigned to each rank. **(B)** Visualization of RybB interactions along with ranking information after rank-conciliation. Interactions are represented by their correspondent MFE of interaction (calculated by RNAup), correlation and mRNA binding site location probability, and are color-coded according to its final rank position, ranging from yellow to blue (top to bottom positions). Red dots indicate known RybB targets. Top-ranking base pairings approximate targets in terms of features, while the similarity worsens for lowest ranks. The rank-conciliation utilized evidence weights that optimally recover targets for all considered *E. coli* sRNAs (Table 9).

4.4 Determining the weights associated with various types of evidence

To determine the weights for various types of evidence, a set of known targets are needed for at least one sRNA. One of the best characterized sRNA is RybB, for which several targets have been validated, along with the specific RybB binding site locations (86). I searched for the weights that could optimally recover RybB targets using particle swarm optimization (PSO, (91). Given that RybB targets are ranked $r_{1,}\ldots r_{p}$ after rank-conciliation using a particular weight configuration $W=(w_{1,}w_{2,}w_{3})$, the observed ranksum given W is

$$S_{obs,W}=(r_1+\ldots+r_p) \tag{11}$$

The fitness of the particle W is defined as

$$D_w=S_{obs,W}-S_{ideal} \tag{12}$$

where S_{ideal} is the ideal ranksum where all RybB targets occupy the top positions of the list. Then the goal of the PSO search is to find W that minimizes *D*.

In order to search for optimal weights W , the PSO algorithm described in Nedjah and Mourelle (91)was implemented with modifications. The PSO search initiates by generating 50 random solutions (weights); then each solution ("particle") traverses the search space by following the best solution (optimal fitness) in its immediate neighborhood (swarm). After a certain number of iterations, swarms converge to a local minimum. Aiming to find a global rather than a local minimum, I explored the search space using 10 swarms containing 5 particles each.

Over repeated iterations, the PSO converges to several weight configurations W that minimize the ranksum of RybB targets. In the next step, rank-conciliation chooses, among all

possible W found in the PSO search, the solution that minimizes the ranksum for all experimentally verified sRNA-target interactions *E. coli*. This validation step assures that the optimal weights determined using RybB targets are also applicable to remaining sRNAs.

Figure 16B shows that, by applying rank-conciliation to all possible RybB base pairings using PSO-learned weights, it is possible to observe that the top positions of the conciliated rank are occupied by sRNA-mRNA base pairings that resemble (in terms of characteristics) interactions between RybB and its targets. In general, the ranking seems to recapitulate target properties, since the top-ranking, lowest weighted RP values (in yellow) tend to be closer to known targets while the lowest-ranking ones (blue) deviate from them.

4.5 Rank-conciliation based on three types of evidence improves target prediction

To find out whether the incorporation of novel data types improved target inference, I compared the results of target prediction using ranking based on MFE only (traditional method) versus rank-conciliation based on multiple types of evidence. The search for targets was done for a list of 17 *trans*-encoded sRNAs responsible for repressing at least one gene in *E. coli*, and for which gene expression profiles were available. sRNAs were represented by the sequence corresponding to their active domains that are assumed to be responsible for target binding (see Table 9).

Several genes are transcribed from different promoters because of alternative sigma factors. For this reason, a different target sequence was included for each annotated 5' UTR of the gene. The complete list of targets comprised 1697 analyzed mRNA targets (corresponding to 1305 genes). The list of validated *E. coli* 5' UTRs were obtained from RegulonDB release 6.8 (92). The fact that each annotated 5' UTR was treated as a separate potential target can lead to

better accuracy on estimated mRNA accessibility energies. Each mRNA target sequence was represented by the complete 5' UTR and 200 bases nt of coding sequence.

The comparison between rank-conciliation and traditional target prediction was performed using three different tools for MFE calculation. The programs RNAup, RNAplex and RNAduplex were used to compute the MFE for the sRNA-mRNA interaction, using default parameter settings, except *w* (maximal length of hybridization in RNAup), which was set to 24 based on the literature (86). Base pairings involving at least 7 nucleotides were selected for the comparison.

The optimal weights for each data type integrated by rank-conciliation were found by the PSO procedure described above. Independently of the tool used for computing the MFE, the PSO optimization found that the MFE rank has the highest importance (weight) for ranking targets, followed by anti-correlation and mRNA binding site position likelihood (Table 5).

	Weights		
RNA-RNA prediction tool	MFE of interaction	Negative correlation	Likelihood of mRNA binding site position
RNAup	0.5634	0.2274	0.2092
RNAplex	0.4689	0.1601	0.3709
RNAduplex	0.4532	0.2642	0.28

Table 5: Optimal weights used for rank-conciliation.

Next, both rank-conciliation and the traditional method were applied to predict sRNA-mRNA interactions. Figure 17 compares the results of target prediction obtained by both methods. In general, rank-conciliation can move known targets up in the list of predictions, since the approach finds smaller target ranksum for the majority of tested sRNAs. Specifically, the ranksum was improved for 11 sRNAs (out of 16) when applying rank-conciliation to RNAup interactions, and 10 sRNAs when using RNAplex and RNAduplex. The ranksum remained the

same for MicC targets among all comparisons. The positive correlation between ChiX and RydC sRNAs and their target genes result in a worse ranksum when using rank-conciliation, implying that the microarray compedium does not include informative conditions for inferring targets for these specific sRNAs.

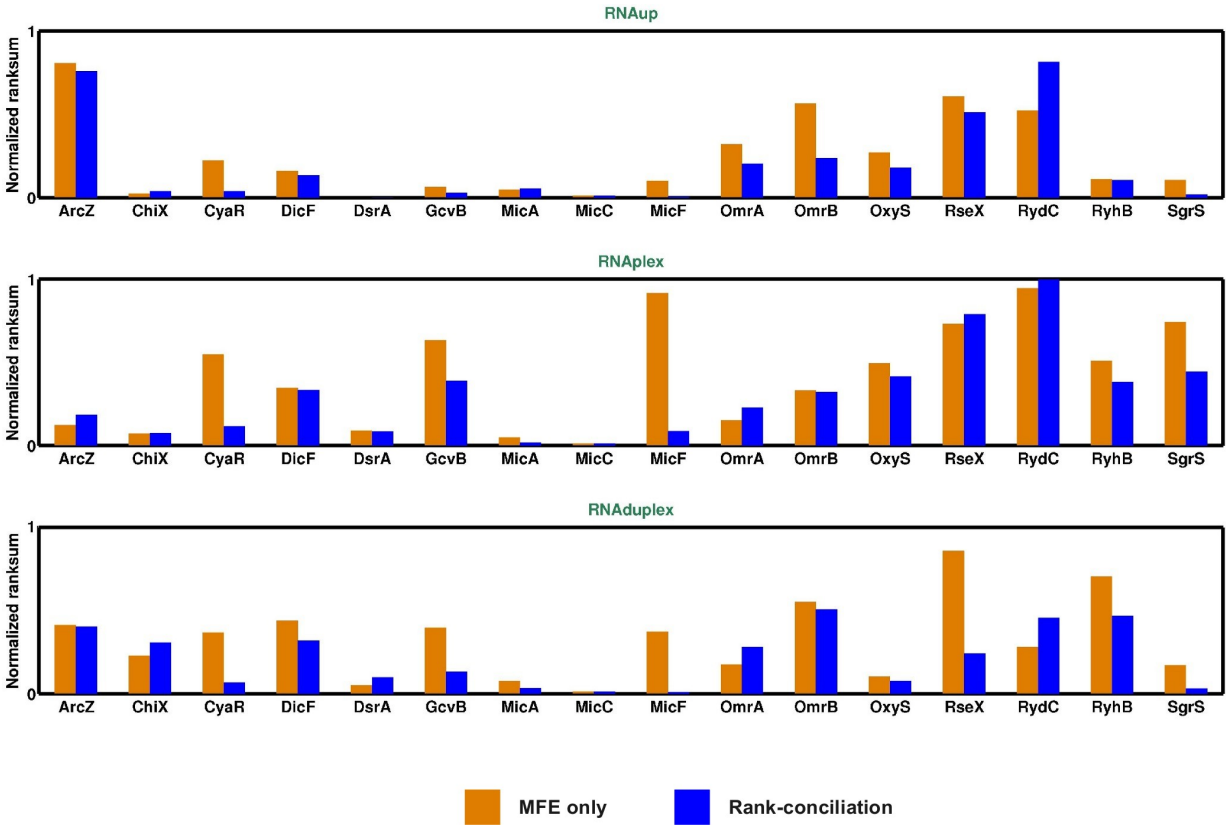

Figure 17: Rank-conciliation improves target recovery. Target ranksum obtained by ranking sRNA-mRNA interactions based on the MFE of interaction only (traditional approach, orange) was compared to the ranksum obtained using the rank-conciliation approach (blue). MFE of sRNA-mRNA base pairings were computed using RNAup, RNAplex and RNAduplex as indicated. Target ranksums are transformed into the interval [0,1] for comparison purposes, and are calculated as the sum of known targets positions within the ordered list of predictions.

The quality of the ranking was also evaluated using two metrics, namely the ranksum expected value (*e-value*) and the recall. Consider that targets occupy positions $r_{1,}\ldots r_p$ in the final ranking of length *N*, and that $S_{obs}=(r_1+\ldots+r_p)$ is their correspondent ranksum. Then, the ranksum *e-value* is the expected probability of finding a better ranksum in random orderings of the same length *N*, and is calculated as follows

$$e=1-c.d.f(S_{obs}) \tag{5}$$

where the cumulative distribution function (*c.d.f*) is constructed by ranksums collected from 10,000 random permutations of target positions. Also, the recall is defined as the proportion of recovered targets among the predictions.

4.6 Function of predicted mRNA targets

I performed an in-depth analysis of the sRNA-mRNA interactions predicted by RNAup and subsequently ranked by rank-conciliation. Figure 18 shows the top 100 base pairings obtained for each sRNA, discriminating the rank position of known targeted transcripts (green). For several sRNAs, including ChiX, GcvB, MicA, MicC, RybB, RyhB and SgrS, the rank-conciliation effectively retrieved targets in the uppermost ranks. For example, most GcvB-regulated transcripts were recovered between positions 1 and 18 (recall 0.9, e-value ~ 0). However, the approach could not find targets in the top 100 positions for ArcZ, DsrA, RseX and RydC (4 out of 17 sRNAs).

To gain some insight in the biology underlying the predictions, I performed GO enrichment analysis using the lists of predicted regulated genes among the top 100 base pairings

shown in Figure 18. As expected, for the group of sRNAs with high recall, the GO terms were in agreement with their repressed functions (see Figure 19). For instance, GcvB is known for inhibiting amino acid transporters during rich growth conditions (enriched for GO:0006865 amino acid transport, and GO:0016597 amino acid binding). Remark ably, many predicted GcvB targets are related to these terms: *gltI* (ranked 8th), *artJ* (26th) are amino acid transporters (GO:0006865), whereas *asnB* (4th) and *serA* (10th) are enzymes responsible for amino acid synthesis (GO:0016597). Also, RyhB was linked to its role of maintaining iron homeostasis during deprivation by down-regulating non-essential Fe-binding proteins (GO:0046872 metal ion binding, GO:0005566 iron ion binding, GO:0051539 iron sulfur cluster binding). Notably, several inferred targets are involved in these biological themes: *dmsA* (ranked 1st), *napF* (23th) were annotated as iron binding proteins (GO:0005566), while *hemN* (8th), *grxD* (9th), *nrfA* (10th), *sodA* (19th), *napF* (23th) and *dinB* (34th) are described as non-covalent metal binding proteins (GO: 0046872). Thus, rank-conciliation was capable of recovering known function, suggesting that it should be able to infer novel interesting targets, which in turn motivated experimental validation.

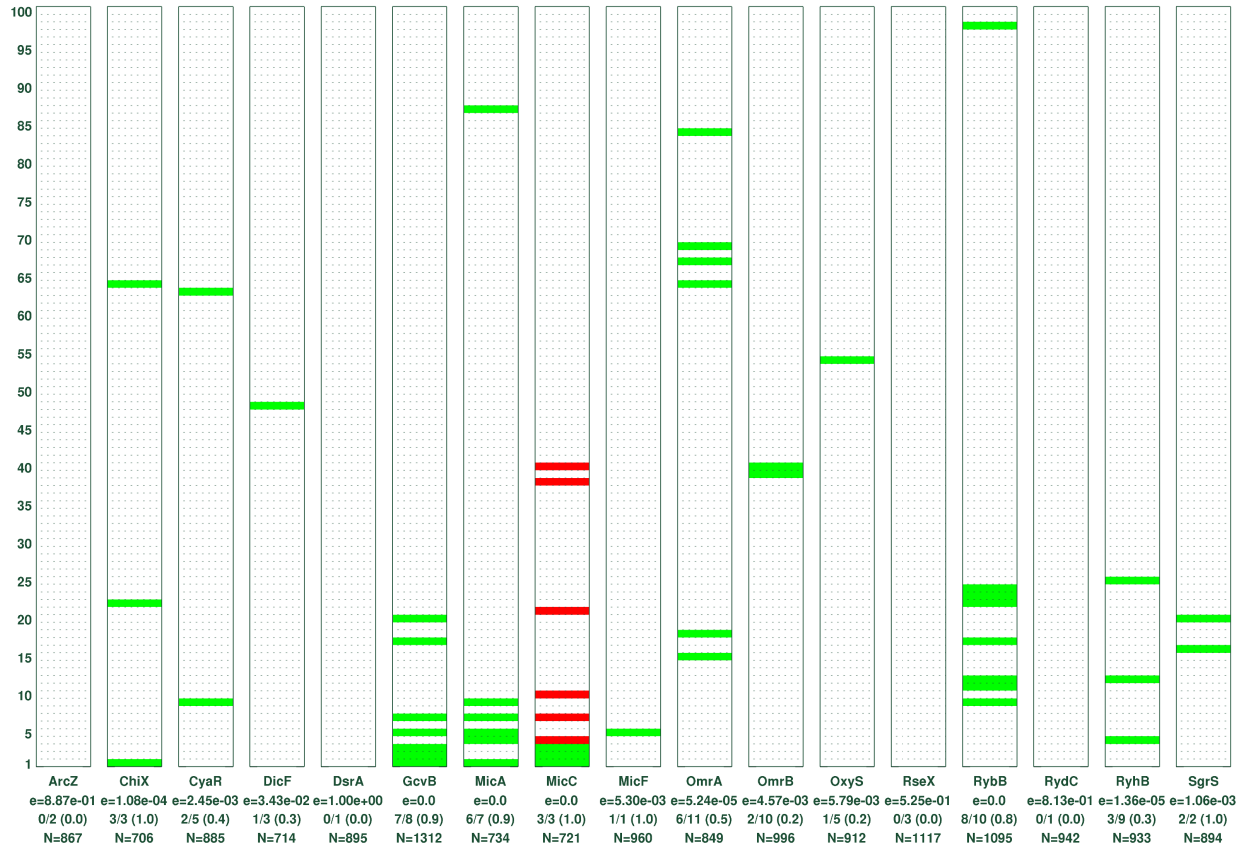

Figure 18: Rank-conciliation of *E. coli* sRNAs base pairings based on three data types. Targets were predicted using MFE of interaction (computed by RNAup), negative correlation and likelihood of binding site position. The ranking ability in recovering known targeted mRNAs (marked in green) is assessed by the *e-value* of the observed ranksum, which estimates the probability of finding a better ranksum in random orderings with the same length N. The recall R (proportion of recovered targets) is based on the top 100 interactions. MicC base pairings chosen for experimental validation are depicted in red. The complete list of rank-conciliation predictions is available at: *https://docs.google.com/spreadsheet/ccc?key=0Aue_AzEGF-5SdGZaZkU1YXJqNGdTblhqZE84ajZyUWc*

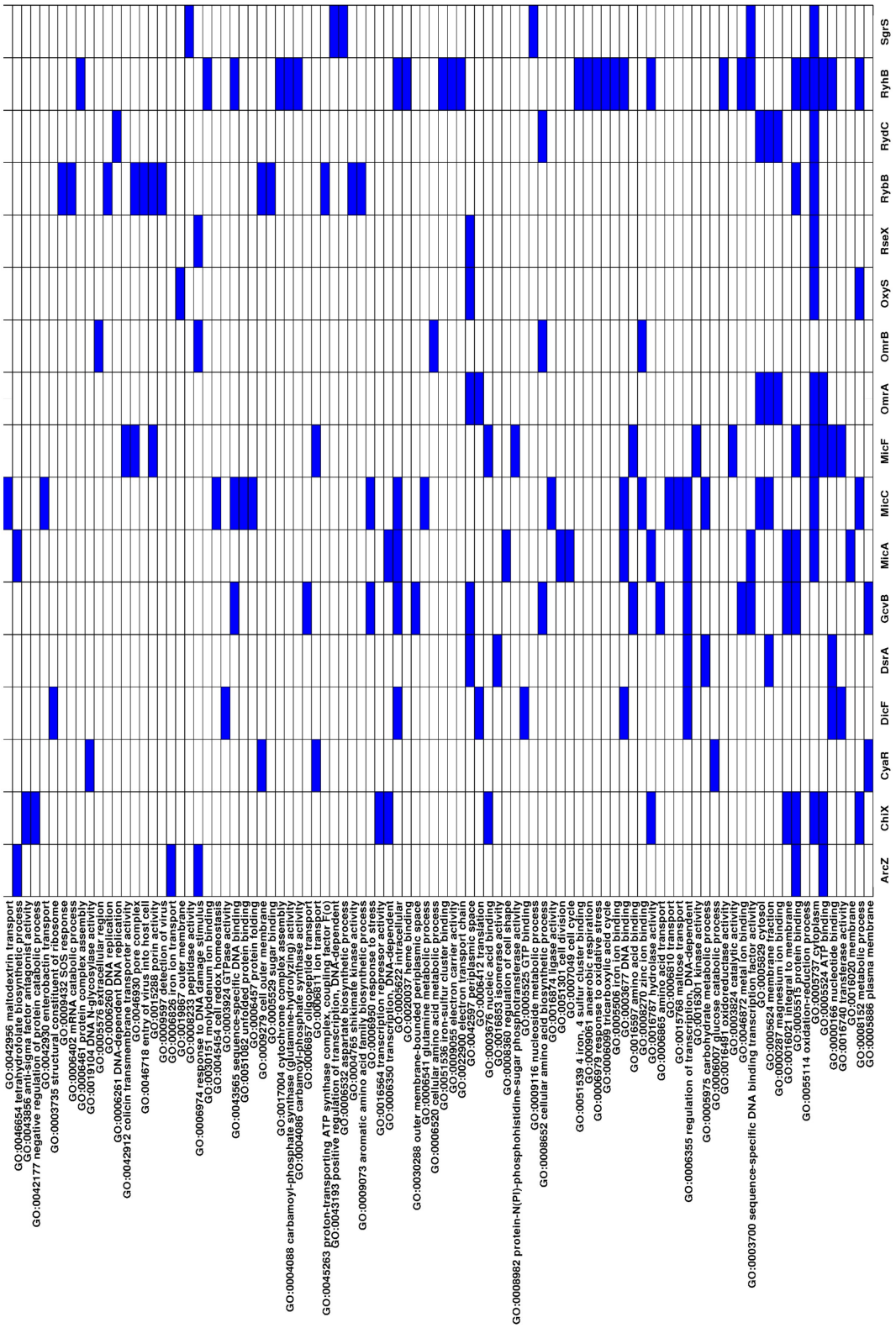

Figure 19: Over-represented GO terms among top 100 predictions made for *E. coli* sRNAs. (Fisher's exact test, $p < 0.05$).

4.7 Analysis of MicC secondary structure

The MicC sRNA was selected for follow-up experiments. As discussed before, bacterial sRNAs are longer than eukaryotic counterparts, and fold into complex secondary structures. Recently it has been shown that only single-stranded regions of the folded sRNA are responsible for target binding (86). Indeed, restricting the base-pairing search to single-stranded regions has found to improved target prediction (93). Therefore, to search for MicC targets using rank-conciliation, I used the 5' end of MicC sRNA, which is likely to be involved in target recognition for several reasons that I explain below.

Using a set of homologous genes from Enterobacterial species, I found that the MicC consensus secondary structure has a conserved single-stranded region (~ 15 nt) in the 5' end (see Figure 20). Experimental data also suggest that MicC utilizes its 5' end for target selection. For example, the determination of MicC secondary structure in Salmonella using *in vitro* probing found this region to be single-stranded (59), and mutational analysis showed that the 5' end binds to *ompC*, the only reported target of MicC in *E. coli (94)*. Therefore, the search for MicC targets (described in the sections above) was done using the 5' end of MicC, which is the region potentially able to recognize mRNAs.

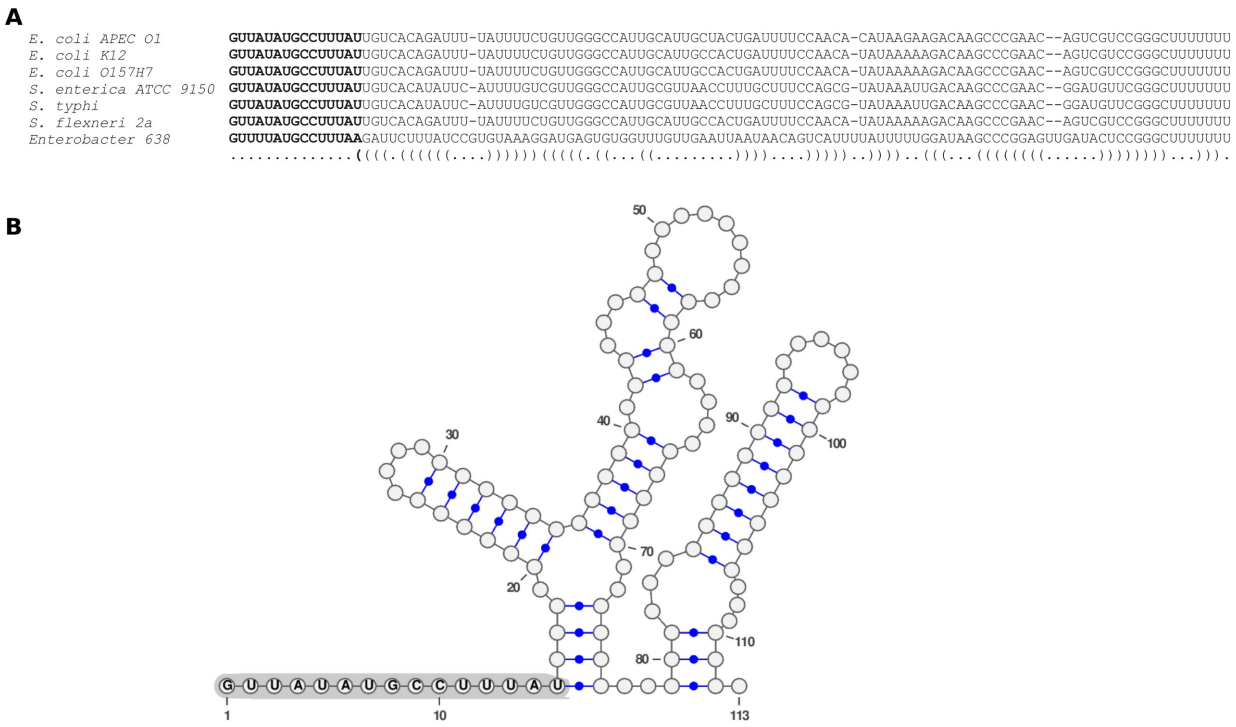

Figure 20: Secondary structure analysis identified the 5' end as the target recognition domain of MicC. (A) Consensus secondary structure predicted by LocaRNA based on the alignment of orthologous MicC genes in Enterobacteria. The structure is shown in a dot-bracket notation where dots "." represent unpaired nts, and parenthesis "**(,)**" represent internally-paired nts. (B) Consensus folding highlighting the15-nt long unpaired and conserved region of MicC used for target searching.

4.8 Experimental validation of MicC predictions

The validation of top-ranking MicC predictions was performed in two rounds. In the first round, 7 candidates were selected for validation: *ompC* (ranked 1st), *eno* (ranked 4th), *glnA* (7th), *tolC* (10th), *valS* (22th), and *manX* (39th). The targets are depicted in Figure 18 (highlighted in red). OmpC is the only known target of MicC, and was included as a positive control.

To determine the effect of MicC on the target mRNA sequence (TS), I used a GFP reporter assay developed to study sRNA regulation in bacteria (95). The reporter assay requires

the use of two plasmids, namely (*i*) the high-copy number sRNA plasmid, expressing MicC under the control of the constitutive PLlacO-1 promoter, and (*ii*) the low-copy number target plasmid constitutively expressing the target sequence fused to GFP (TS:GFP). As negative controls, cells were transformed with (*i*) a high-copy number plasmid expressing a scrambled sRNA, and (*ii*) the target plasmid expressing the TS:GFP fusion.

For testing MicC effects on the TS, *E. coli* cells were transformed with both the sRNA and target plasmids, and the abundance of GFP was measured on western blots. Therefore, lower levels of GFP (when compared to negative controls) indicate that MicC binds to mRNA sequence and prevents translation of GFP. Additional details of the reporter system are described in in the section Materials and Methods.

I constructed translational fusions of the TS and GFP (TS::GFP) and performed cloning to create the target plasmids. Table 6 provides details of the constructed GFP fusions, including 5' UTR coordinates, regions of base pairing and other properties of the tested sRNA-mRNA interactions. Often, the cloned TS contained only the the 5' UTR of target genes, and additional codons were included when the interaction was predicted to cover the coding region.

The western blot analysis was performed after overnight growth of *E. coli* TOP10 cells constitutively expressing both the GFP fusion and MicC (described in Material and Methods). As shown in Figure 21A, Eno, GlnA, ManX, GFP fusions do not appear affected by MicC regulation. Conversely, TolC::GFP exhibited a moderate inhibition, whereas GFP production from OmpC, ValS and YdgA fusions were strongly inhibited by MicC. These results suggest that MicC exerts its post-transcriptional control through recognition of the 5' UTR region of *ompC* (-29 to -15), *tolC* (-101 to -92) and *ydgA (-58 to -48),* while affecting *valS* by CDS recognition (+6 to +18). The base-pairings between MicC and tested genes are shown in Table 8.

Rank[a]	5' UTR	Target gene	Host plasmid	Fusion name[b]	5' end[c]	Fused codon	Base pairing[d]	MFE binding	Correlation	Repression by MicC[e]
1	ompCp	*ompC*	pSK-003	OmpC::GFP	-81	12	(-29,-15)	-18.12	-0.58	strong
4	enop	*eno*	pDV-Lenol	Eno::GFP	-76	67	(+183,+189)	-12.25	-0.51	none
7	glnAp2	*glnA*	pDV-LglnAl	GlnA::GFP	-73	63	(+175,+186)	-11.97	-0.29	none
10	tolCp2	*tolC*	pDV-LtolCl	TolC::GFP	-103	1	(-101,-92)	-9.83	-0.55	moderate
22	valSp2	*valS*	pDV-LvalSl	ValS::GFP	-82	9	(6,18)	-9.55	-0.36	strong
35	ydgAp	*ydgA*	pDV-LydgAl	YdgA::GFP	-58	17	(-58,-48)	-9.25	-0.34	strong
39	manXp	*manX*	pDV-LmanXl	ManX::GFP	-115	20	(-16,-10)	-7.79	-0.45	none

Table 6: Overview of tested GFP fusions in the first round of target validation, along with base pairings properties and MicC repression effects.

[a]Position after rank-conciliation based on MFE of interaction (RNAup), negative correlation and likelihood of mRNA binding site location.

[b]Construct name used in the text

[c]5' end of the promoter relative to the start codon

[d]Interaction site predicted by RNAup, relative to the start codon

[e]Negative change in GFP fusion levels due to MicC over-expression measured in western blots (see Figure 21A)

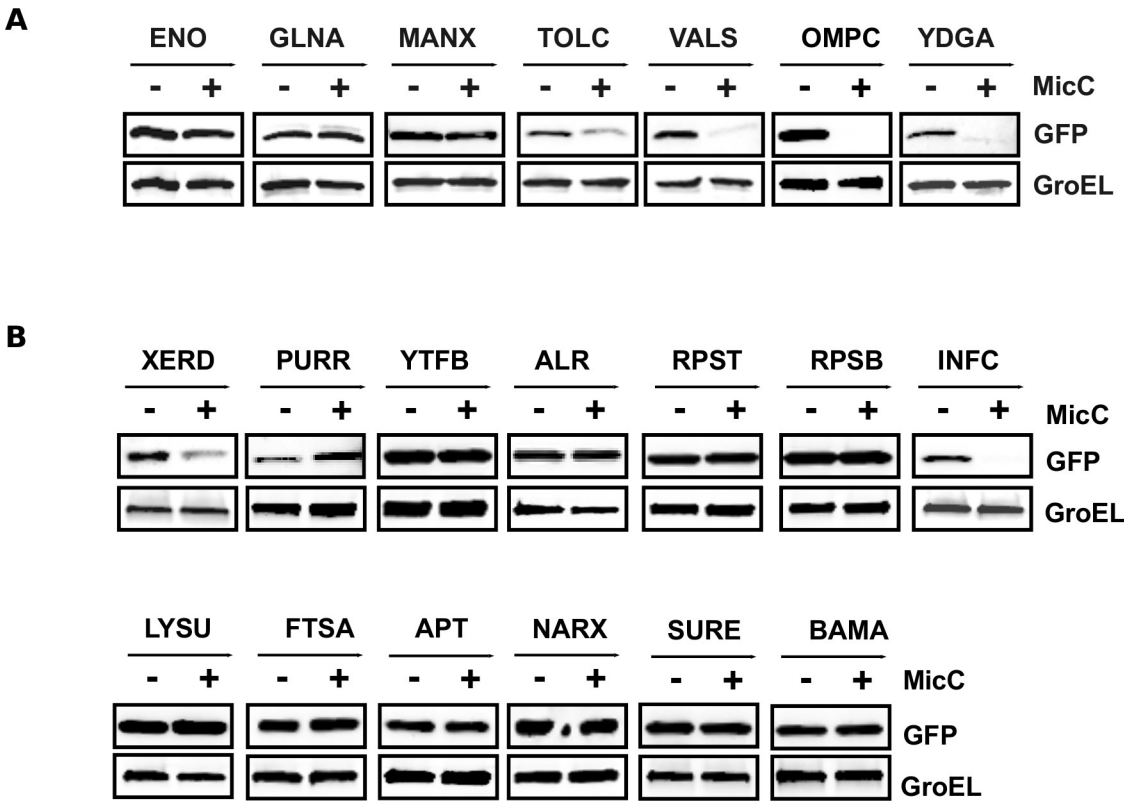

Figure 21: Immunobloting using anti-GFP antibody measured the levels of translational GFP fusions to target sequences of selected genes, in the presence of MicC regulation (+) and during non-specific regulation by the shuffled control sRNA (-). (A) Genes tested in the first round of validation. (B) Genes tested in the second round of validation. After overnight grow, cell lysates were prepared and resolved by SDS-PAGE, transferred to nitrocellulose membrane and probed with anti-GFP and anti-GroEL (loading control) antibodies.

4.9 Second round of validation of MicC predictions

The western blot analysis show that MicC failed to inhibit Eno (ranked 4th) and GlnA (ranked 7th) genes (Figure 18), despite their top MFE of interaction and anti-correlation values. RNAup predicted that these base pairings should occur more than 50 codons downstream from ATG, in a region where the ribosome translates with high efficiency (96). Ribosome and sRNAs compete for mRNA binding, and the fact that the MicC is not able to regulate targets in regions too far from the start codon indicates that the search for targets should be confined within the

space of low translational efficiency. This region of low translational speed corresponds approximately to the first 30 codons in *E. coli* (96).

For this reason, in the second round of validation the target prediction was performed using mRNA target sequences composed of the complete 5' UTR and 30 codons of coding sequence. With modified sequence inputs, rank-conciliation also predicts *ompC, valS, ydgA,* and *tolC* as top candidates (among 10 first predictions), as well as novel interesting candidates. From the list of top predictions, 13 target genes were selected for experimental validation (details described in Table 7). Western blot analysis of selected genes suggested that *infC* and *xerD* are negatively regulated by MicC (Figure 21B). MicC was predicted to base-pair in the Shine-Dalgarno region of *infC,* and the 5' UTR of *xerD* (Table 8), indicating that the sRNA interferes with translation by preventing ribosome binding. Considering both rounds of validation, from 20 tested genes, only 6 genes appeared to be affected by MicC, and therefore may be considered putative MicC targets.

Rank[a]	5' UTR	Target gene	Host plasmid	Fusion name[b]	5' end[c]	Fused codon	Base pairing[d]	MFE binding	Correlation	Repression by MicC[e]
12	ftsAp1	*ftsA*	pDV-004	FtsA::GFP	-421	1	(-13,-5)	-8.72	-0.42	none
15	lysUp2	*lysU*	pDV-006	LysU::GFP	-80	1	(-73,-64)	-8.06	-0.51	none
16	aptp	*apt*	pDV-007	Apt::GFP	-102	1	(-12,-6)	-8.97	-0.17	none
24	narXp	*narX*	pDV-009	NarX::GFP	-159	1	(-131,-123)	-8.19	-0.41	none
25	xerDp	*xerD*	pDV-010	XerD::GFP	-108	1	(-40,-33)	-8.49	-0.28	moderate
27	surEp2	*surE*	pDV-011	SurE::GFP	-412	1	(-215,-205)	-7.05	-0.5	none
28	bamAp	*bamA*	pDV-012	BamA::GFP	-107	1	(-11,-2)	-6.26	-0.52	none
29	purRp	*purR*	pDV-013	PurR::GFP	-155	1	(-63,-56)	-8.69	-0.04	none
31	ytfBp	*ytfB*	pDV-015	YtfB::GFP	-102	1	(-86,-79)	-8.49	-0.20	none
33	rpsBp	*rpsB*	pDV-016	RpsB::GFP	-162	1	(-36,-30)	-7.42	-0.40	none
34	infCp2	*infC*	pDV-017	InfC::GFP	-180	1	(-5,+2)	-7.93	-0.24	strong
35	alrp1	*alrP*	pDV-018	Alr::GFP	-193	1	(-178,-171)	-8.24	-0.16	none
38	rpsTp1	*rpsT*	pDV-020	RpsT::GFP	-132	1	(-55,-46)	-8.15	-0.16	none

Table 7: Overview of tested GFP fusions in the second round of target validation, along with base pairings properties and MicC repression effects. See Table 6 for description of the columns.

Gene	Base-pairing site
ompC	1 GUUAUAUGCCUUUAU 15 -15 CAAUAUACGGAAAUA -29
tolC	UA 1 GUUA UGCCUUUAU 10 -92 CAAU ACGG -101 GC
apt	5 GUUAUAUGCCUUUAU 11 -6 AUACGGA -12
ftsA	AUA 1 GUU UGCCUUUAU 11 -5 CAA ACGGA -13 G
purR	7 GUUAUAUGCCUUUAU 14 -56 ACGGAAAU -63
ytfB	8 GUUAUAUGCCUUUAU 15 -79 CGGAAAUA -86
xerD	5 GUUAUAUGCCUUUAU 12 -33 AUACGGAA -40
alr	6 GUUAUAUGCCUUUAU 13 -171 UACGGAAA -178
narX	A 2 GUUAU UGCCUUUAU 10 -123 AAUA ACGG -131 A
lysU	A 2 GUUAU UGCCUUUAU 11 -64 AAUA ACGGA -73 G
infC	6 GUUAUAUGCCUUUAU 12 +2 UAUGGAA -5
rpsB	5 GUUAUAUGCCUUUAU 11 -30 AUACGGA -36
rpsT	UU 1 G AUAUGCCUUUAU 10 -46 C UAUACGG -55 CU
valS	C 1 GUUAUAUG CUUUAU 13 6 CAAUAUAC GAAA 18 A
surE	AU 1 GUU AUGCCUUUAU 10 -210 CAA UACGG -215 GCU
bamA	1 GUUAU-AUGCCUUUAU 9 -2 CAAUAAUACG -11
manX	9 GUUAUAUGCCUUUAU 15 -10 GGAAAUA -16

Table 8: Predicted base-pairings between MicC and tested target genes. The top line indicates the base-pairing position in MicC while the the bottom line denotes the region on the target gene.

4.10 Biological insights gained from MicC putative target genes

Little is known about MicC regulatory capabilities, except the fact that this sRNA is induced in low temperatures and in minimal media supplemented with glycerol (94). When expressed, MicC contributes to replacing the porin composition of the outer membrane (from OmpC to OmpF) by repressing ompC mRNA (94). OmpF is a general diffusion porin with relatively large diameter that allows non-selective passage of solutes (97), and is preferred during low nutrient and non-toxic conditions (94).

One of the putative targets, TolC is an outer membrane porin (like OmpC), and a required subunit of several efflux transport systems that expel drugs to the extracellular environment. TolC associates with other proteins to form tripartite efflux pumps, such as TolC:AcrA:AcrB that confer resistance to multiple antimicrobials, dyes, bile salts and detergents (98). Interestingly, cells lacking TolC exhibit metabolic shutdown, inhibition of NADH dehydrogenases and growth arrest (99). This phenotype suggests that the negative regulation of TolC might be triggered during starvation accompanied by low toxins levels, therefore when MicC is induced and TolC-based pumps are not required.

YdgA is a protein of unknown function that is present as a dimer in the inner membrane (100). Even tough the molecular function of YdgA has not been determined, YdgA-deficient *E. coli* cells are unable to swarm, a phenotype compatible to low energy cells during starvation (101).

Additionally, both *infC* and *valS* are carrying out functions related to protein translation in cells. Specifically, *valS* codes for the valyl-tRNA synthetase, the enzyme that catalyzes the covalent linking of valines to uncharged $tRNA^{Val}$ molecules during translation, at expense of

ATP (102). On the other hand, InfC is a translation initiation factor that binds to the 30S ribosomal subunit, and mediates pairing of the tRNA anticodon to the mRNA start codon, thereby promoting translation (103).

Based on the function of putative target genes identified herein, MicC might be involved in the cellular response to low nutrient conditions (starvation) by acting in two distinctive ways. On one side, MicC increases the permeability of the bacterial envelope through repression of OmpC and TolC, therefore enhancing nutrient intake. On the other side, MicC contributes to arresting translation in cells with low ATP levels, by repressing ValS and InfC.

4.11 Rank conciliation in another scenario: combining custom microarrays and MFE of binding

The methodology to rank-order sRNA targets described in this study can be adapted to several contexts, when different sets of experiments are available to probe sRNA regulation. To demonstrate this point, I applied rank-conciliation approach to infer targets for FnrS, a recently characterized sRNA that belongs to the Fnr regulon in *E. coli*.

The Fnr transcriptional regulator responds to low intracellular oxygen and mediates the transition from aerobic to anaerobic growth (104). During anoxia, Fnr activates the expression of FnrS sRNA, which in turn participates in the anaerobic shift by post-transcriptionally repressing genes related to aerobic growth (53; 82) and Figure 22A.

Boysen *et al* (82) designed two microarray experiments aimed at finding FnrS targets. In the first experiment, mRNA levels were measured after transient overexpression of FnrS (*fnrS*$^{++}$$^{+}$), and compared to a mutant background (*fnrS*$^{-}$), in aerobic conditions. Then, a ranking was created based on this dataset, by sorting genes from the lowest to highest fold-change; this

ranking captures the impact of FnrS induction on transcriptome degradation (Figure 22B). In the second experiment, mRNA levels between the mutant (*fnrS*$^{-}$) and wild-type (*fnrS*$^{+}$) were compared after 30 min of incubation without O_2. In this case, genes were ranked from the highest to lowest fold-change, in order to capture the effects of relaxed FnrS regulation during the anaerobic shift (Figure 22B). A third ranking was done using the MFE of binding (computed by RNAup), i.e. FnrS and all possible target sequences in *E. coli*, as detailed before.

Next, rank-conciliation was applied to integrate all rankings, which comprised two steps, (i) computing weights based on the ranksum optimization and using FnrS validated targets; recovered weights were *0.4491*, *0.4271*, and *0.1238* for experiment 1, experiment 2 and binding MFE respectively, and (ii) generating an output rank of interactions calculated on the basis of the weighted geometric mean.

Figure 22C depicts the top 100 target transcripts inferred by rank-conciliation. A functional enrichment analysis of top predictions found an over-representation of genes implicated in oxidation-reduction (GO:0055114, p-value 0.01), including *hmpA* (ranked 8th), *dapB* (15th), *cueO* (16th), *gltB* (24th), *norV* (31th), *nfsB* (32th), *fdnG* (33th), and *ahpC* (38th). Several of these genes (*hmpA*, *dapB*, *gltB*, *nfsB*, *fdnG*) code for NAD reductases, which transform NAD into NADH. In normal aerobic respiration, NADH is the major substrate (electron donor) entering the oxidative phosphorylation pathway to produce energy. During anaerobic conditions, NADH is replaced by other electron donors (such as nitrate); therefore it is plausible that FnrS may be targeting these NAD reductases for degradation. Finally, the *yceJ* gene (ranked 1st) is a cytochrome b561 homologue that is predicted to be involved in cellular respiration (http://www.ncbi.nlm.nih.gov/gene/945628).

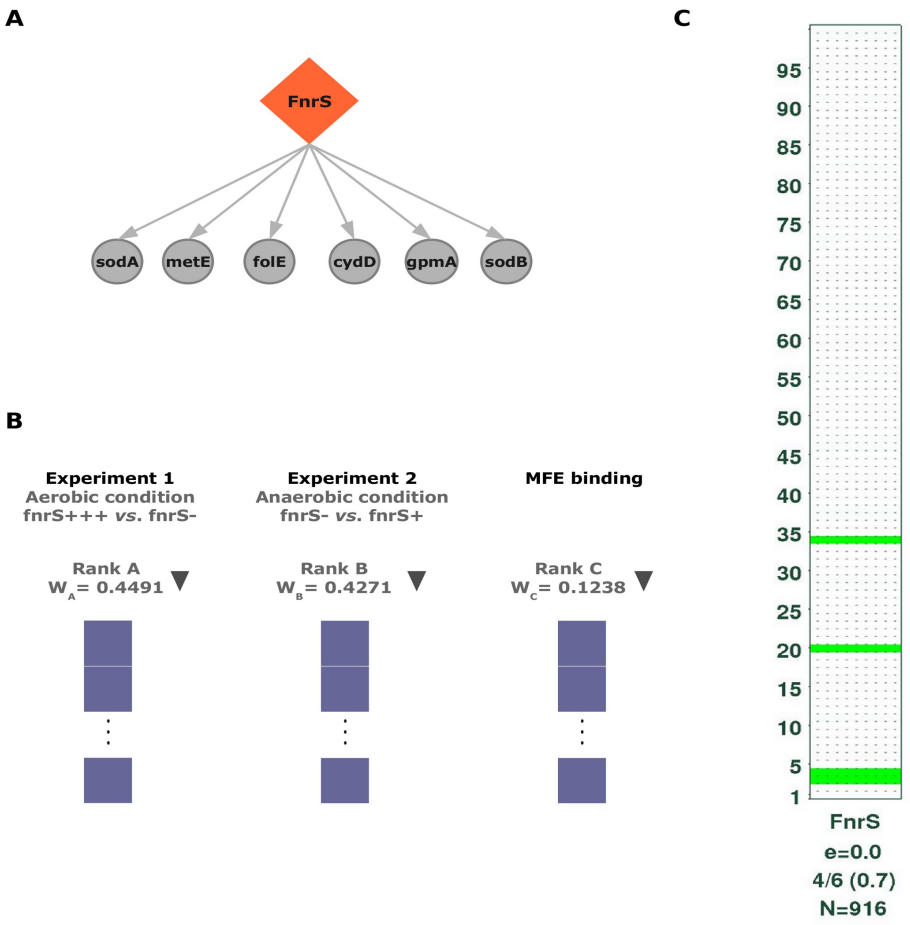

Figure 22: Rank-conciliation applied to FnrS target prediction. (A) Set of FnrS targets with annotated 5' UTR used for ranksum optimization. (B) Target genes were ranked based on microarrays generated in Boysen *et al* and MFE of binding as computed by RNAup (as performed in Figure 18); optimal weights are indicated next to each ranking. (C) Top 100 FnrS regulated transcripts (known targets are highlighted in green). FnrS is predicted to be regulating several enzymes related to aerobic respiration (discussed in the text). Recall and e-value are defined as in Figure 18.

4.12 Discussion

Prediction of sRNA targets is routinely performed by ranking potential targets according to the energy score of the interaction computed by RNA hybridization tools. However, these

approaches are limited because RNA binding mechanisms are non-trivial. Here I demonstrated that evidence for sRNA-mRNA regulation may exist in multiple datasets, which can be combined to predict novel targets. To this end, I developed rank-conciliation, a methodology for target prediction that can be viewed as a data integration framework seeking to combine multiple relevant data types for sRNA target inference. The experimental validation of some predictions generated by rank-conciliation revealed new post-transcriptional regulatory interactions in *E. coli*.

Since several data types can be considered simultaneously for inference of sRNA-mRNA interactions, the question I asked was how to combine these data types in a meaningful way to produce a priority list of putative new targets. Here, the methodology proposed uses a modified version of the rank products metric to address this problem. The method creates ranks for each data type (or evidence), the final ranks is found by merging the contributions of all data types. Specifically, the final rank of a given target is the (weighted) geometric average rank of this target across all rankings.

Additionally, not all rankings might have the same importance for target prediction. This assumption reflects the fact that some data types are better predictors of direct targets, and therefore should be rewarded. To tackle this issue, rank-conciliation utilizes (i) a weighted version of the rank products, and (ii) an optimization procedure aimed at finding out the contribution of each ranking on target prediction. In this context, the weight represents how well the criterion is able to rank order known targets in the early positions (i.e. minimize the ranksum of known targets).

The optimal weights recovered for *E. coli* sRNAs indicate that the ranking based on MFE of interaction is the most efficient in recovering direct targets, when compared to other

data types. The lower weight for the Pearson correlation ranking indicates that, even though negative correlation is a property that characterize sRNA-mRNA interactions, the use of this measure of association in a genome-wide context retrieves many indirect associations. Most importantly, it was demonstrated that the combination of these multiple informative clues about sRNA regulation achieve a better target inference, as shown by the improvement in target recovery (ranksum) for most of *E. coli* sRNAs analyzed in this study.

Rank-conciliation is also suitable for target prediction based on other combinations of computational/experimental datasets generated to probe sRNA regulation. As a proof of principle, I applied rank-conciliation to infer targets for the FnrS sRNA, by integrating base pairing prediction along with gene expression data collected in two relevant conditions. FnrS was predicted to be regulating several enzymes related to aerobic respiration, which is consistent with its role of repressing genes related to aerobic growth.

Additionally, variations of current rankings could be easily incorporated. For instance, novel algorithms for RNA interaction prediction based on more accurate underlying models and/or thermodynamical parameters are continuously being developed (105), and they can be used to compute MFE for interactions as they become available.

Additional experiments are necessary to confirm MicC targets

The western blot analysis conducted for target validation was done using three technical replicates, since it was repeated three times using protein samples extracted from one individual bacterial population. However, when I repeated the western blots using biological replicates (three protein samples from three independent clones were analyzed), I could not confirm that MicC is indeed affecting the translation of *infC*, *ydgA*, *valS*, *tolC*, and *xerD*. For example, in the case of *infC*, one replicate showed a complete knock-down of the GFP fusion in the presence of

MicC, while in the other replicates the levels of GFP were unchanged.

It is still unclear what is causing these contradicting results. One possibility is that long-term overexpression of MicC may be toxic for cells, especially because the sRNA is expected to arrest translation by repressing genes such as *infC* and *valS.* Then under constitutive MicC expression, some of the clonal populations that are collected for experiments may be overtaken by single cells that acquired mutations and are not expressing MicC. Indeed, another study that has used the same GFP reporter system to find targets for GvcB found that constitutive sRNA expression cause pleiotropic effects and affect readouts (61). They found that short-term overexpression of the sRNA from an inducible promoter could circumvent this problem, since this technique was able to identify a high number of GcvB targets, but just a fraction of them could be validated using the GFP reporter system.

Therefore, even though there is supporting evidence for MicC regulation, I consider that *infC*, *ydgA*, *valS*, *tolC* and *xerD* are putative MicC targets, and further experiments are necessary to prove regulation.

4.13 Materials and Experimental Methods

Plasmid construction and bacterial strains

Cloning of GFP fusion plasmids was performed as described in Urban and Vogel (95). *E. coli* TOP10 competent cells (Invitrogen) were chemically transformed with (i) low-copy plasmid constitutively expressing the putative target sequence fused with GFP (TS:GFP), and (ii) high-copy plasmid pSK-017 harboring MicC, under the control of the constitutive $P_{LlacO-1}$ promoter. In the control, cells were transformed with (i) the TS:GFP fusion plasmid and (ii) the high-copy plasmid pJV300 producing a shuffled nonsense sRNA.

Western blotting

Transformed *E. coli* cells were inoculated in 1 ml LB medium containing ampicilin (100 μg/ml) and chloramphenicol (20 μg/ml), and grown overnight at 37ºC. Total proteins were extracted by resuspension in lysis buffer (Cell Signaling Technology, Danvers, MA, USA), sonication (briefly up to 3X), followed by incubation on ice for 30 minutes. Lysates were centrifuged at 13,000 g for 20 min at 4°C and the protein concentration in the supernatants was determined using a colorimetric assay (BioRad DC Protein Assay; BioRad, Hercules, CA, USA). A total of 30 μg of protein (100 μg for YdgA::GFP fusion) was denatured in reducing sodium dodecyl sulphate (SDS) loading buffer (Cell Signaling Technology) at 95°C and separated by SDS-PAGE. Proteins were then electrotransferred onto nitrocellulose membranes and submitted to immunodetection with the relevant antibody, namely GFP mouse monoclonal antibody #2995 (Cell Signaling) and Anti-GroEL rabbit antibody G6532 (Sigma-Aldrich, St. Louis, MO, USA). Membrane-bound secondary antibodies (HRP-conjugated goat anti-rabbit or anti-mouse, BioRad) were detected using SuperSignal West Dura Extended Duration Substrate (Pierce Chemical Co., Rockford, IL).

Small RNA	Active regions[a]
RybB	**GCCACTGCTTTTCTTT**GATGTCCCCATTTTGTGGAGCCCATCAACCCCGCCATTTCGGTTCAAGGTGATGGGTTTTTTGT
GcvB	ACTTCCTGAGC**CGGAACGAAAAGTTTTATCGGAATGCGTGTTCTGGTGAACTTTTGGCTTACGGTTGTGATGTTGTGTTGTTGTGTTTGCAATTGGTCTGCGATTCAGACCATGGTAGCAAAGCTACCTTTTTTCACTTCCTGTACATTTACCCTGTCTGTCCATAGTGATTAATGTAGCACCGCCTAATTGCGGTGCTTTTTTTT**
OmrA	**CCCAGAGGTATTGATTGGTGAG**ATTATTCGGTACGCTCTTCGTACCCTGTCTCTTGCACCAACCTGCGCGGATGCGCAGGTTTTTTTT
OmrB	**CCCAGAGGTATTGATAGGTGAAGTCAACTTCGGGTTGAGCACATGAATTACACCAGCCTGCGCAGATGCGCAGGTTTTTTTT**
RyhB	GCGATCAGGAAGACCCTCGCGGAG**AACCTGAAAGCACGACATTGCTCACATTGCTTCCAGTATTACTTAGC**CAGCCGGGTGCTGGCTTTT
CyaR	**GCTGAAAAACATAACCCATAAAATGCTAGCTGTACCAGGAACCACCTCCTTAGC**CTGTGTAATCTCCCTTACACGGGCTTATTTTTT
DsrA	**AACACATCAGATTTCCTGGTGTAACGAATTTTTTAAGTGCTTCTTGCTTAAG**CAAGTTTCATCCCGACCCCCTCAGGGTCGGGATTT
MicC	**GTTATATGCCTTTAT**TGTCACAGATTTTATTTTCTGTTGGGCCATTGCATTGCCACTGATTTTCCAACATATAAAAAGACAAGCCCGAACAGTCGTCCGGGCTTTTTTT
FnrS	**GCAGGTGAATGCAACGTCAAGCGATGGGCGTTGCGCTCCATATTGTCTTACTTCCTTTTTTGAATTACTGCATAGCACAATTGATTCGTACGACGCCGACTTTGATGAGTCGGCTTTTTTTT**
MicA	**GAAAGACGCGCATTTGTTATCATCATCCCT**GAATTCAGAGATGAAATTTTGGCCACTCACGAGTGGCCTTTTTCTTTT
RseX	**TTTTTATTATTCTGTGTCATGATGCTTCCGTTATTAGCCTTTTATCGTCTTGTTTATATTTTTTGGGCCGGCATGATGCCGGCTTTTTTTT**
OxyS	GAAACGGAGC**GGCACCTCTTTTAACCCTTGAAGTCACTGCCCGTTTCGAGAGTTTCTCAACTCGAATAACTAAAGCCAACGTGAACTTTTGCGGAT**CTCCAGGATCCGCT
ArcZ	GTGCGGCCTGAAAAACAGTGCTGTGCCCTTG**TAACTCATCATAATAATTTACGGCGCAGCCAAGATTTCCCTGGTGTTGGCGCAGTATTCGCGCACCCCGGTCTAGCCGGGGTCATTTTTT**
ChiX	ACACCGTCGCTTAAAGTGA**CGGCATAATAATAAAAAAATGAAATTCCTCTTTGACGGGCCA**ATAGCGATATTGGCCATTTTTTT
MgrR	GATTCGTTATCAG**TGCAGGAAAATGCCTGTTAGCGTAAAAGCAAAACACAAATCTATCCATGCAAGCATTCACCGCCGGTTTACTGGCGGTTTTTTTT**
SgrS	GATGAAGCAAGGGGGTGCCCCATGCGTCAGTTTTATCAGCA**CTATTTTACCGCGACAGCGAAGTTGTGCTGGTTGCGTTGGTTAAGCGTCCCACAACGATTAACCATGCTTGAAGGACTGATGCAGTGGGATGACCGCAATTCTGAAAGTTGACTTGCCTGCATCATGTGTGACTGAGTATTGGTGTAAAATCACCCGCCAGCAGATTATACCTGCTGGTTTTTTTT**
RydC	**CTTCCGATGTAGACCCGTATTCTTCGCCTGTACCACGGGTCGGTTTTAGTACAGGCGTTTTCTT**
MicF	**GCTATCATCATTAACTTTATTTATTACCGTCATTCATTTCTGAATGTCTGTTTACCCCTATTTCAACCGGATGCCTCGCATTCGGTTTTTTTT**
DicF	**TTTCTGGTGACGTTTG**GCGGTATCAGTTTTACTCCGTGACTGCTCTGCCGCCC

[a]mRNA-target binding regions of the sRNA. These regions were predicted by a Bayesian classifier and were collected from Peer and Margalit (86), except for Ryb and MicC which have been extracted from Papenfort *et al* (85).

Table 9: Active binding regions of *E. coli* sRNAs used for target prediction (denoted in bold).

Chapter 5

A Hypothesis on Sequence Determinants of *trans*-encoded sRNA Regulation in *Escherichia coli*

5.1 Motivation

Messenger RNAs carry in their sequence two types of binding sites that allow their selection by sRNAs, namely base-pairing sites and Hfq sites. Base-pairing sites are located in regions of sequence complementarity between sRNA and mRNA sequences. Often, the quality and extent of complementarity within the base-pairing site determines the hybridization potential for the formation of the sRNA-mRNA duplex. Also, most of trans-encoded sRNAs circulate in the cell bound to Hfq, an RNA-binding protein that increases sRNA stability by preventing ribonuclease cleavage (106). Besides sRNA binding, the Hfq protein contains a region for mRNA binding (106). Recently, it has been shown that the presence of Hfq binding sites in the mRNA contributes to target recognition, since Hfq serves as a docking platform that promotes the encounter of an sRNA and its cognate mRNA (106).

Regarding sequence properties, both evolutionarily conservation and accessibility of binding sites in the secondary structure have been shown to be important for target recognition in eukaryotes. For example, sequence conservation is a signature of binding sites in eukaryotes, being that most miRNAs recognize conserved sites in the 3' untranslated region of mRNAs (107). On the other hand, mRNA folding may affect the accessibility of base-pairing and Hfq sites, and thereby impact sRNA recognition. In eukaryotes, the efficiency of miRNA-based

repression is proportional to the exposition of the base-pairing site in the particular mRNA conformation, being that sites buried in the mRNA reduced or abolished miRNA repression (108). This structural-dependent recognition is also observed for natural and synthetic siRNAs in mammalian cells (74; 109). In general, base-pairing sites within accessible regions in secondary structures, such as hairpin loops and other single-stranded portions of the mRNA, are more likely to be functionally active binding sites.

However, the importance of accessibility and conservation for the recognition of mRNA binding sites in bacteria has not been addressed yet. Here, I investigated the accessibility and conservation requirements necessary for a binding site to be recognized by Hfq-dependent sRNA. To answer this question, I performed a computational analysis to determine the accessibility and conservation profiles that distinguish targets from non-targets of several trans-encoded sRNAs.

In Chapter 4, I experimentally tested MicC regulation on several mRNA target sequences which contained base-pairing sites for MicC. Therefore, for the MicC sRNA, the collection of positive and negative examples were obtained based on the results of the western blots. For other analyzed sRNAs, I collected information of targets and non-targets from published experiments.

First, the analysis indicated that MicC is capable of repressing mRNA targets in two different contexts, dependent and independently of Hfq binding. In both cases, the fact that base-pairing sites as well as Hfq sites are predominantly composed of conserved and accessible nucleotides emerged as the explaining factor for MicC recognition. As this analysis is extended to other *trans*-encoded sRNAs for which experimental data is available, I demonstrate that other sRNAs are also able to recognize targets in two different contexts, but accessibility and conservation of binding sites may not be essentially required for target selection for all of them.

5.1 Computing accessibility and conservation profiles of putative targets and non-targets

In order to verify whether accessibility and conservation properties related to base-pairing and Hfq sites were able to distinguish putative targets from non-targets, I computed the accessibility and conservation (AC) profiles for the list of mRNA sequences tested for MicC regulation with the GFP reporter assay.

The AC profile of a sequence contains information regarding (*i*) the accessibility value of each nt of the folded mRNA, which refers to the probability of being unpaired in the ensemble of RNA structures found by RNAfold; (*ii*) the conservation score of the sequence in Enterobacteria, as obtained from the PhasCons track of the Microbes Genome Browser, and (*iii*) the location of base-pairing and Hfq sites. The sequences were scanned for Hfq sites using triplets of the motif A-R-N, implicated in binding to mRNAs (106). Figure 23 depicts examples of AC profiles for a few mRNA sequences, including putative MicC targets *valS*, *ydgA* and *tolC*.

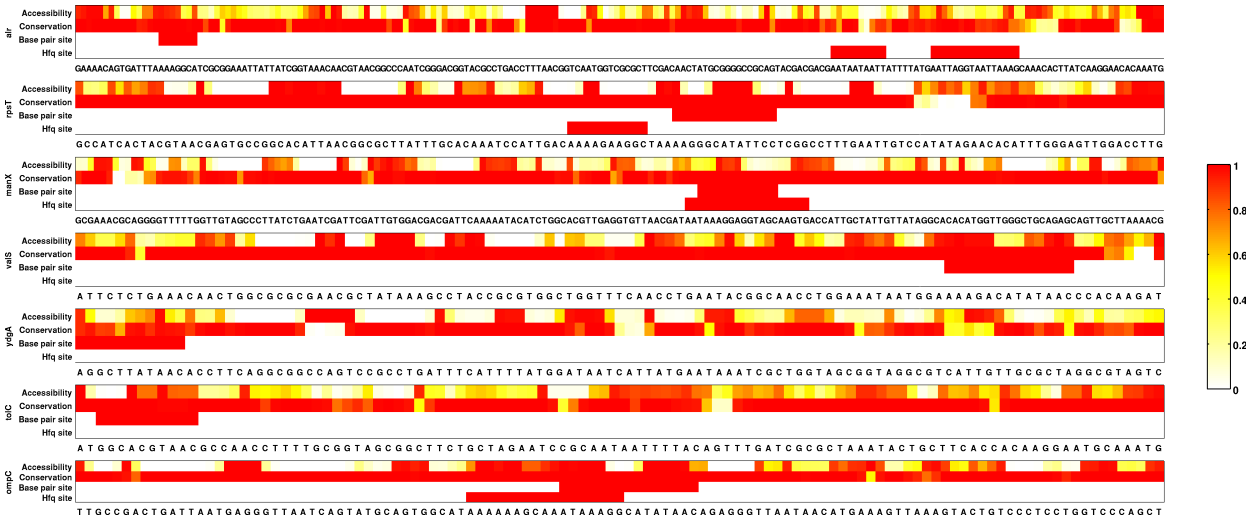

Figure 23: Accessibility and conservation profiles along with the location of predicted Hfq and base-pairing sites in tested mRNA sequences. The color indicates the level of conservation and accessibility at each nt in the sequence.

Based on the AC profiles, a series of features concerning the composition and relationship of base-pairing and Hfq sites were computed (described in Table Error: Reference source not found). Examples of composition features include the mean accessibility and mean conservation of binding sites. To help define other meaningful features that combine accessibility, conservation and the relative position of binding sites, two definitions were introduced. First, the term “seed nt” was used to designate those nt that are both conserved and unpaired in the secondary structure (AC threshold > 0.75, see Table Error: Reference source not found). Second, a Hfq site was deemed proximal when lying within a 35 nt window around the base-pairing site, and distant otherwise. This last definition was used to create features such as whether base-pairing sites were flanked by Hfq sites, and to compute the accessibility Hfq sites neighboring the base-pairing region.

Feature name	Description
Bp_acc	Mean accessibility of base-pairing site
Bp_cons	Mean conservation of base-pairing site
Bp_flanked_Hfq	If base-pairing site contains a proximal and non-overlapping Hfq site
Bp_#proximalHfq	Number of proximal non-overlapping Hfq sites[a]
Acc_proximalHfq	Mean accessibility of proximal Hfq site[b]
Cons_proximalHfq	Mean conservation of proximal Hfq site
Bp_unpairedNT	Proportion of unpaired nt within the base-pairing site
Bp_seedNT	Proportion of nt that are both conserved and unpaired within the base-pairing site[c]
ProximalHfq_seedNT	Proportion of nt that are both conserved and unpaired within the proximal Hfq site

[a]A Hfq site is considered proximal when is located within a window of 35 nt around the base-pairing site. The tolerance for overlapping is 1 nt.
[b]If more than one proximal sites exist in the 35 nt window, the site with highest accessibility is chosen
[c]A seed nt contains a conservation score > 0.75 and accessibility (unpaired probability) > 0.75.

Table 10: Accessibility and conservation features used for mRNA sequence analysis.

5.3 Accessibility of base-pairing sites is critical for MicC recognition

The ability of a given feature in discriminating between positive and negative examples of MicC regulation was tested using a Wilcoxon rank-sum test (Figure 24A). The test compares the feature distribution between genes considered targets (6 examples) and non-targets (14 examples), and the corresponding p-value indicates the separability power of the feature. According to the rank-sum test, features related to the accessibility of the base-pairing site were able to achieve separation. The most significant separation is achieved using the proportion of seed nt (*BP_seedNT*, P=0.02), followed by proportion of unpaired nt (*Bp_unpairedNT*, P=0.18), and mean accessibility of (*Bp_acc*, P=0.2).

In general, base-pairing sites recognized by MicC contain a proportion of seed nt significantly higher than expected by chance (Figure 24B). The only exception was *ydgA*, for

which the seed nt composition was comparable to random. Any flanking Hfq sites, when present (*infC* and *ompC*), were found to be highly conserved and accessible.

Next, I tested whether the conservation and accessibility profiles of adjacent base-pairing and Hfq sites could achieve a better separation between putative targets and non-targets. Specifically, a decision tree classifier with two features – *BP_seedNT* and *ProximalHfq_seedNT* –, was applied to define whether the seed nt composition of binding sites could optimally separate positive and negative examples.

The resulting decision tree found that MicC putative targets can be optimally separated from non-targets based on those two features (Figure 24C). The decision tree partitioned the input dataset into two major sub-types of MicC putative targets. Accordingly, sub-type I targets harbor base-pairing sites heavily composed by seed nt (at least 76%), which are not flanked by Hfq sites. The lack of flanking Hfq sites suggests that sub-type I genes (*valS*, *tolC*, *xerD)* do not depend directly on Hfq binding for translational repression. On the other hand, sub-type II targets carry a base-pairing site that is adjacent to a highly conserved and accessible Hfq site. The seed nt proportion in base-pairing sites for sub-type II is slightly lower than for sub-type I, but still high when compared to a randomized value (Figure 24B). Interestingly, complete knock-down of protein translation was observed for sub-type II genes *ompC* and *infC*. This implies that an accessible and conserved Hfq site may act in synergy with the base-pairing site to strengthen the regulatory effect.

YdgA was the only putative target that falls out of these sub-groups, since it contains a base-pairing site with a low frequency of seed nt and no predicted Hfq sites. The existence of YdgA suggests that MicC recognition is not strictly dependent on accessibility of binding sites.

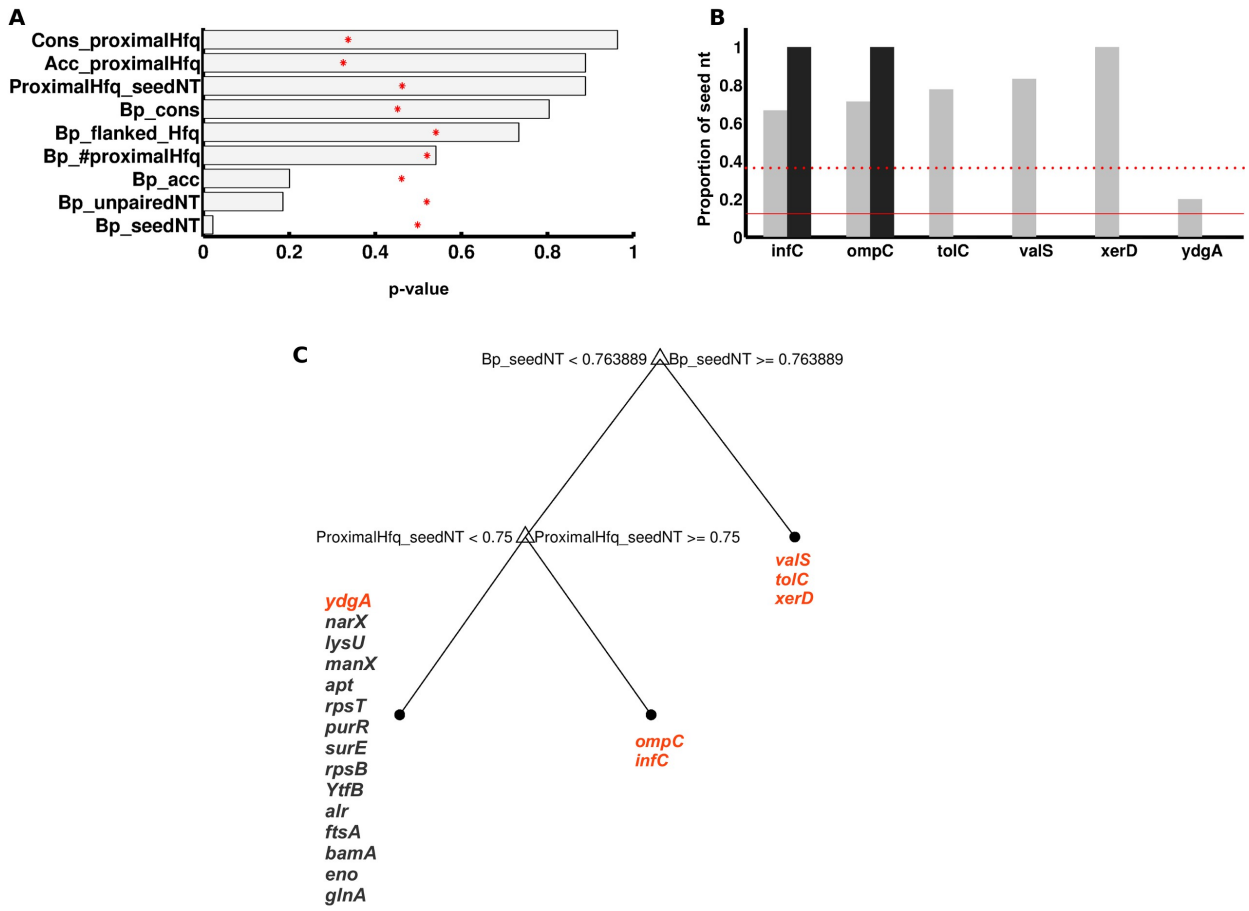

Figure 24: Discriminating MicC targets from non-targets based on accessibility and conservation of binding sites. (A) Separability power of each individual feature was evaluated using a Wilcoxon rank-sum test. The p-value (gray bar) indicates whether the feature distribution have different medians between targets and non-targets. The red star denotes an expected random p-value resulting from performing 1,000 rank-sum tests on randomized features values (obtained by shuffling conservation and accessibility profiles of sequences). **(B)** Proportion of seed nucleotides found in base-pairing sites (gray bars) and proximal Hfq sites if existent (black bars) of MicC targets. The solid red line indicates the proportion of seed nt expected by chance (μ_r) and the dotted red line indicates two standard deviations from the random mean ($\mu_r+2\sigma$). The null distribution of seed nt was generated using 1,000 randomized AC profiles, as described above. **(C)** Separation of MicC targets and non-targets using a decision tree classifier revealed two sub-types of mRNA sequences repressed by MicC.

5.4 Accessibility and conservation of binding sites in other sRNAs

Next, I tested whether features related to the accessibility of base-pairing sites are predictors of

regulation in other *trans*-encoded sRNAs. Specifically, positive and negative examples of genes that are regulated by GcvB, MicA, RyhB, OmrA and Spot42 (Spf) were collected from the literature, and the target separability was evaluated using the rank-sum test.

The results showed that the accessibility features identified for MicC were not able to achieve a significant separation of targets from non-targets when applied to these additional sRNAs (Figure 25A). Also, by analyzing the distribution of seed nt in base-pairing sites recognized by such sRNAs (Figure 25B), I observed that they tend to interact with a broad spectrum of base-pairing sites, that range from low to high seed nt content. For instance, the median proportion seed nt for GcvB, MicA, and OmrA is significantly higher than random, but they interact with targets through less accessible/conserved regions as well. Interestingly, most of Spot42 base-pairings occur in relatively poorly accessible/conserved regions.

Likewise MicC, a better separability of targets and non-targets of other sRNAs is achieved by using a decision tree with two features, namely the *BP_seedNT* and *ProximalHfq_seedNT.* As shown in Figure 26A, GcvB regulates both sub-type I targets (2 genes) as well as sub-type II targets (3 genes). However, for GcvB, the feature *ProximalHfq_seedNT* sites is a better predictor of regulation, since it give a better likelihood of finding a true target (3:1 likelihood, *ProximalHfq_seedNT* > 0.58).

I also observed that Spf targets do not fall into the sub-type I and II classification, as depicted in the decision tree of Figure 26B. The majority of binding sites recognized by Spf are found in low accessibility/conservation regions (*BP_seedNT* < 0.37), and are not flanked by Hfq sites.

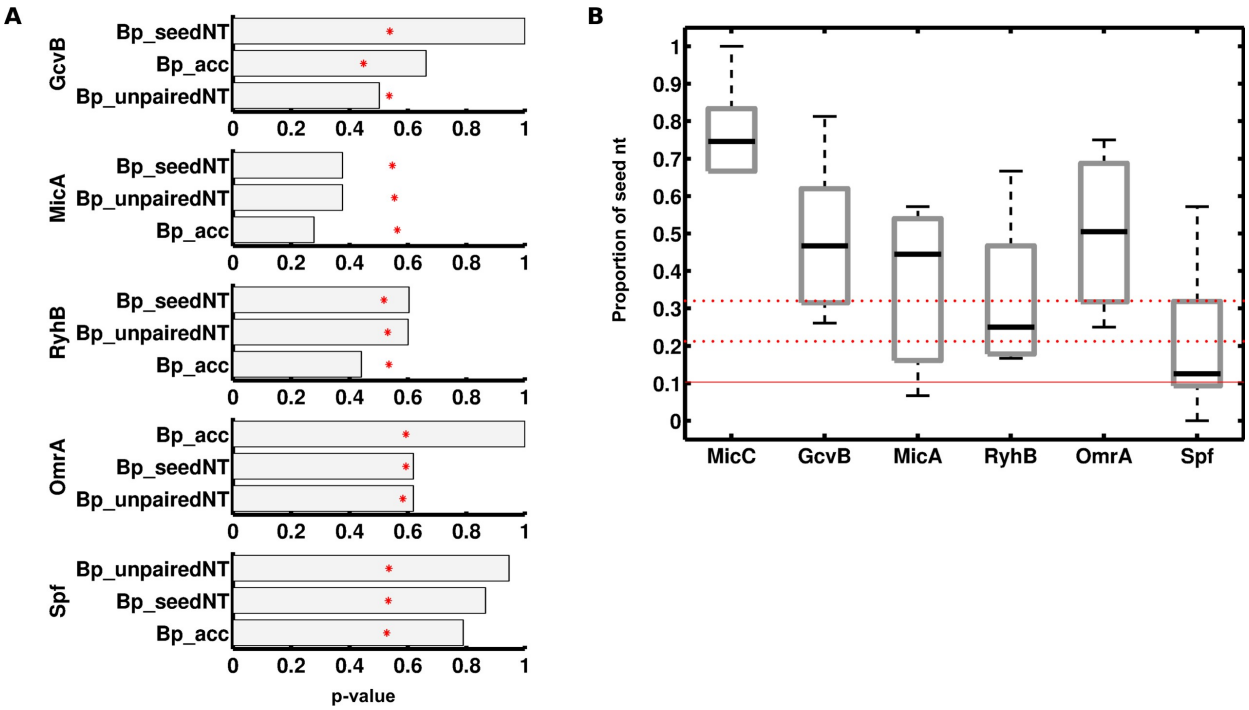

Figure 25: Accessibility-related features do not discriminate targets from non-targets in other trans-encoded sRNAs. (A) P-values of the rank-sum test using the corresponding feature on the right for separation. The red star denotes the random p-value computed as described in Figure 24A. Positive examples were taken from a list of sRNA targets compiled in Peer *et al* (86) and Beisel *et al* for Spf (93). Negative data was retrieved from the sRNAtar database (112) and Beisel *et al* for Spf, and included only genes tested using a small-scale experiment such as northern blots and reporter assays. **(B)** Distribution of the proportion of seed nt in the base-pairing sites of known targets of several sRNAs. The solid red line indicates the expected random proportion of seed nucleotides, while dotted lines denote one and two standard deviations from the random mean.

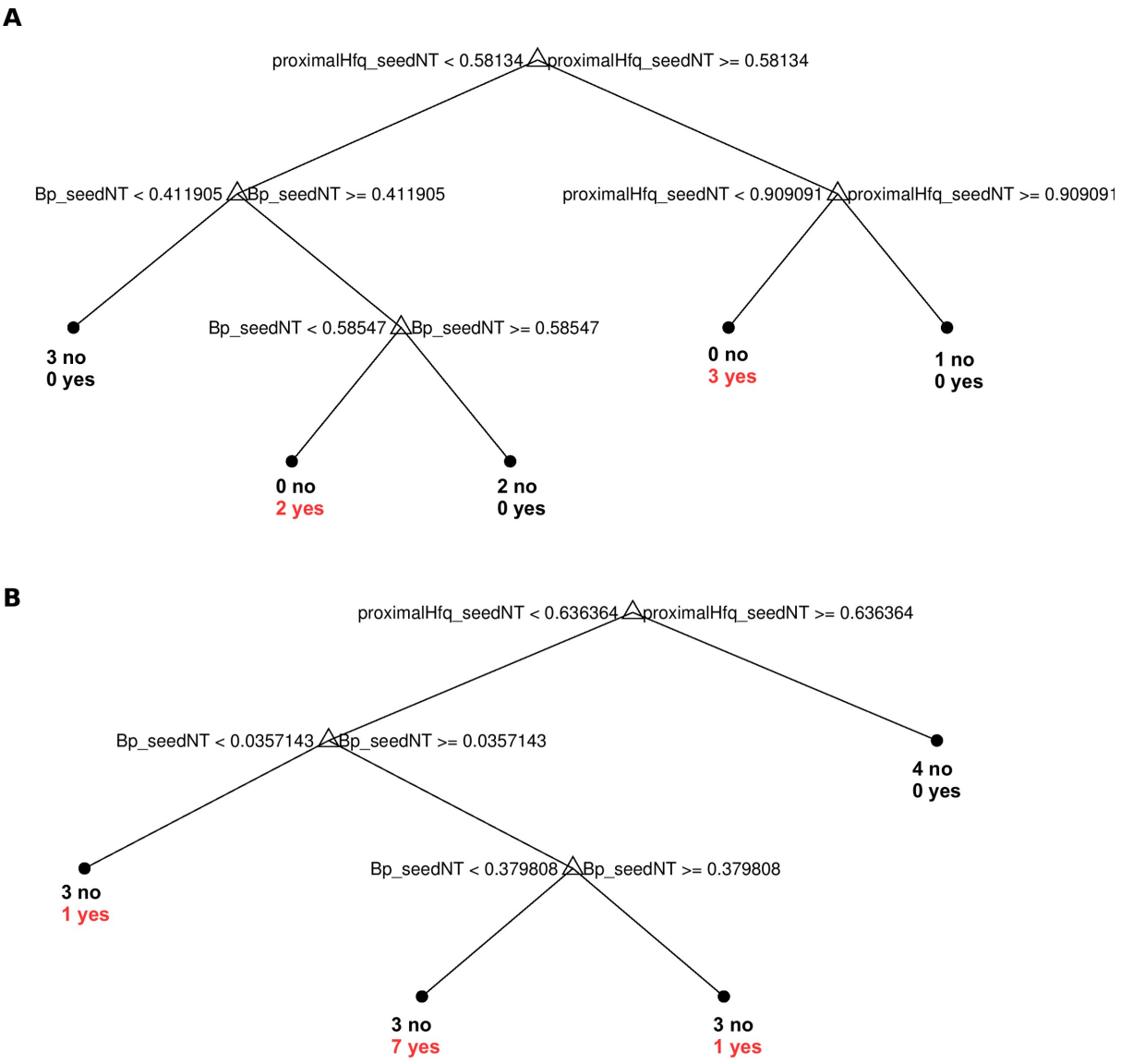

Figure 26: Decision tree analysis using the seed composition of base-pairing and Hfq sites applied to other *trans*-encoded sRNAs. (A) GcvB decision tree. (B) Spf decision tree.

5.4 Discussion

The importance of target site accessibility and conservation for non-coding RNA regulation is well-known in eukaryotes. Herein, using a systematic approach to evaluate base-pairing properties, I showed that site accessibility and conservation may also be a major factor for target

selection by bacterial sRNAs. As demonstrated by the decision tree analysis, all genes considered as MicC targets either contained a high-proportion of seed nt within the base-pairing site (sub-type I), or in the Hfq site adjacent to the base-pairing region (sub-type II).

The hypothesis of the existence of sub-types of targets was observed not only for MicC, but for other *trans*-encoded sRNAs for which experimental data was publicly available. Accordingly, extending the separability analysis to other sRNAs, I found that the separability of targets and non-targets in sub-types I and II also can be observed in other analyzed sRNAs, such as GcvB. However, Spf seems to challenge this view since its targets could not be classified into these sub-types, suggesting that target selection by sRNAs may occur by other mechanisms yet to be discovered, independently of site conservation and accessibility.

Here it was shown that only when considering a combination of features (rather than individual features alone), the analysis was effective in separating targets from non-targets, either for MicC as well as for other sRNAs. This indicates that target selection by sRNAs is dictated by multiple factors, and the seed composition of base-pairing sites and Hfq sites may be included among them.

Altogether, these findings suggest that accessibility and conservation of base-pairing and Hfq sites should be explicitly considered in sRNA target prediction, and computational methods should be adapted to search different sub-types of targets. In the case of the rank-conciliation discussed in Chapter 4, the method can be extended to include two additional rankings related to accessibility and conservation of binding sites. That could be done, for instance, by including rankings for the frequency of seed nt within base-pairing and adjacent Hfq sites.

Another outcome of this analysis is that adjacent Hfq sites may act as enhancer elements working to increase the regulatory strength of the sRNA. Hfq is a RNA-binding protein that has a dual role in cells: (*i*) protecting RNA from degradation, and (*ii*) catalyzing the formation of the

sRNA-mRNA duplex, by increasing the association rate of the duplex formation. Several sRNAs dependend on Hfq for exerting their regulatory effects on mRNA targets. For instance, MicC-*ompC* regulation is absent in strains lacking Hfq (94). Here, I observed that when accessible and conserved Hfq sites are present in the mRNA (*ompC* and *infC* genes), protein translation was completely abolished. Therefore, Hfq-binding mRNAs may constitute ideal sRNA targets, and an experimentally defined catalog of such mRNAs would be helpful to narrow the search for targets. In Salmonella, co-immunoprecipitation (Co-IP) using a Hfq antibody combined with high-throughput sequencing identified that one fifth of genes (~700 mRNAs) are bound by Hfq (64). Restricting the search for targets within this catalog of Hfq-binding mRNAs may prove to be useful for prediction of sub-type II targets.

In addition, the catalog of Hfq-binding mRNAs should be obtained from physiologically relevant conditions, where the sRNA under study is intrinsically up-regulated in the cell. For MicC, low temperatures and minimum medium are ideal conditions for inducing expression (94). Searching for sRNA-regulated genes among Hfq-binding mRNAs identified in such conditions could substantially improve the predictability of sub-type II targets.

Chapter 6

General Conclusion and Outlook

Mechanisms that allow pathogens to colonize the host are not the product of isolated genes, but instead emerge from the operation of biological networks. Therefore, identifying components and the systemic behavior of networks is necessary to a better understanding of gene regulation and pathogenesis. The methodologies developed in this dissertation were applied (*i*) to identify novel network interactions at post-transcriptional level and hence sRNAs that might be related to virulence in *Mtb*, and (*ii*) to obtain a systemic view of *Mtb* survival strategies in the host at a transcriptional level.

Additionally, I presented (*iii*) a methodology to infer sRNA-mRNA interactions based on multiple types of data, that was applied to prioritize putative targets for validation and expedite functional characterization of sRNAs. Accordingly, the putative MicC targets found by this approach link this sRNA to translation arrest and starvation response. This methodology can be ported to *Mtb*, as long as sRNA expression data becomes available.

Systems biology advocates the rationale construction of hypotheses by analysis of large-scale models, followed by experimental testing. This workflow needs to be further optimized on both computational and experimental aspects. For instance, transforming computational workflows into automated units of code would facilitate the generation of *in silico* predictions. Some tools developed to this end allow network inference to be coupled to functional

enrichment analysis (110). Besides making the data analysis reproducible, such automated units can be re-utilized by other investigators analyzing similar data.

Experimental testing represents another bottleneck in the workflow. Confirmatory experiments employed nowadays such as Western blot analysis and RT-qPCR operate in a gene-by-gene basis. Then, by standard approaches, the functional characterization of a single sRNA molecule, which comprises validation of interaction partners and expressed conditions, may take months to years to be completed. Novel assays that increase the output of gold-standard experiments, such as the high-throughput multiplex PCR that allows thousands of PCR reactions in a single microfluidic chip (111) are addressing this issue. Such parallel versions of high sensitivity assays will be increasingly used to perform in batch validation of *in silico* generated hypotheses, thereby optimizing the discovery process.

In summary, I believe that the application of the systems biology workflow – which entails the construction and testing of large-scaled models of biological systems based on integrated datasets - will be pervasive in modern biology, and will be an invaluable tool for aiding biological discovery.

References

1. Zhang W, Li F, Nie L. Integrating multiple “omics” analysis for microbial biology: application and methodologies. Microbiology 2010 Feb;156(2):287 –301.

2. Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chaba R, Lee S, Kazmierczak KM, Lee KJ, Wong A, Shales M, Lovett S, Winkler ME, Krogan NJ, Typas A, Gross CA. Phenotypic Landscape of a Bacterial Cell. Cell 2011 Jan;144(1):143–156.

3. Veiga DFT, Dutta B, Balázsi G. Network inference and network response identification: moving genome-scale data to the next level of biological discovery. Mol Biosyst 2010 Mar;6(3):469–480.

4. Carter SL, Brechbühler CM, Griffin M, Bond AT. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 2004 Sep;20(14):2242–2250.

5. de la Fuente A, Bing N, Hoeschele I, Mendes P. Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 2004 Dec;20(18):3565–3574.

6. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 2006;7 Suppl 1:S7.

7. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-Scale Mapping and Validation of Escherichia coli Transcriptional

Regulation from a Compendium of Expression Profiles. PLoS Biol 2007 Jan;5(1):e8.

8. Modi SR, Camacho DM, Kohanski MA, Walker GC, Collins JJ. Functional characterization of bacterial sRNAs using a network biology approach. Proc. Natl. Acad. Sci. U.S.A. 2011 Sep;108(37):15522–15527.

9. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat. Genet. 2005 Apr;37(4):382–390.

10. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, Lasorella A, Aldape K, Califano A, Iavarone A. The transcriptional network for mesenchymal transformation of brain tumours. Nature 2010 Jan;463(7279):318–325.

11. Cooper AM. Cell-mediated immune responses in tuberculosis. Annu. Rev. Immunol 2009;27:393–422.

12. Monack DM, Mueller A, Falkow S. Persistent bacterial infections: the interface of the pathogen and the host immune system. Nat. Rev. Microbiol 2004 Sep;2(9):747–765.

13. Ghebreyesus TA, Kazatchkine M, Sidibé M, Nakatani H. Tuberculosis and HIV: time for an intensified response. The Lancet 2010 May;375(9728):1757–1758.

14. Charles A Janeway Jr, Paul Travers, Mark Walport, Mark J Shlomchik. Immunobiology [Internet]. 5th ed. New York: Garland Science; 2001.Available from: http://www.ncbi.nlm.nih.gov/books/NBK10757/

15. Ehrt S, Schnappinger D. Mycobacterial survival strategies in the phagosome: defence against host stresses. Cell. Microbiol 2009 Aug;11(8):1170–1178.

16. Flannagan RS, Cosío G, Grinstein S. Antimicrobial mechanisms of phagocytes and bacterial evasion strategies. Nat. Rev. Microbiol 2009 May;7(5):355–366.

17. Luo M, Fadeev EA, Groves JT. Mycobactin-mediated iron acquisition within macrophages. Nat. Chem. Biol 2005 Aug;1(3):149–153.

18. Shi L, Sohaskey CD, Kana BD, Dawes S, North RJ, Mizrahi V, Gennaro ML. Changes in energy metabolism of Mycobacterium tuberculosis in mouse lung and under in vitro conditions affecting aerobic respiration. Proc. Natl. Acad. Sci. U.S.A 2005 Oct;102(43):15629–15634.

19. Meena LS, Rajni. Survival mechanisms of pathogenic Mycobacterium tuberculosis H37Rv. FEBS J 2010 Jun;277(11):2416–2427.

20. Russell DG, VanderVen BC, Lee W, Abramovitch RB, Kim M, Homolka S, Niemann S, Rohde KH. Mycobacterium tuberculosis wears what it eats. Cell Host Microbe 2010 Jul;8(1):68–76.

21. Schnappinger D, Ehrt S, Voskuil MI, Liu Y, Mangan JA, Monahan IM, Dolganov G, Efron B, Butcher PD, Nathan C, Schoolnik GK. Transcriptional Adaptation of Mycobacterium tuberculosis within Macrophages: Insights into the Phagosomal Environment. J. Exp. Med 2003 Sep;198(5):693–704.

22. Fontan P, Aris V, Ghanny S, Soteropoulos P, Smith I. Global Transcriptional Profile of

Mycobacterium tuberculosis during THP-1 Human Macrophage Infection. Infect. Immun. 2008 Feb;76(2):717–725.

23. Tailleux L, Waddell SJ, Pelizzola M, Mortellaro A, Withers M, Tanne A, Castagnoli PR, Gicquel B, Stoker NG, Butcher PD, Foti M, Neyrolles O. Probing Host Pathogen Cross-Talk by Transcriptional Profiling of Both Mycobacterium tuberculosis and Infected Human Dendritic Cells and Macrophages. PLoS ONE ;3(1)

24. Rohde KH, Abramovitch RB, Russell DG. Mycobacterium tuberculosis Invasion of Macrophages: Linking Bacterial Gene Expression to Environmental Cues. Cell Host & Microbe 2007 Nov;2(5):352–364.

25. Rohde KH, Veiga DFT, Caldwell S, Balazsi G, Russel DG. Linking the Transcriptional Profiles and the Physiological States of Mycobacterium tuberculosis During an Extended Intracellular Infection. PLoS Pathogens (*in press*).

26. Balázsi G, Heath AP, Shi L, Gennaro ML. The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest. Mol. Syst. Biol 2008;4:225.

27. Jacques P-E, Gervais AL, Cantin M, Lucier J-F, Dallaire G, Drouin G, Gaudreau L, Goulet J, Brzezinski R. MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis. Bioinformatics 2005 May;21(10):2563–2565.

28. Madan Babu M, Teichmann SA, Aravind L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J. Mol. Biol 2006 Apr;358(2):614–633.

29. Krawczyk J, Kohl TA, Goesmann A, Kalinowski J, Baumbach J. From Corynebacterium glutamicum to Mycobacterium tuberculosis--towards transfers of gene regulatory networks and integrated data analyses with MycoRegNet. Nucleic Acids Res 2009 Aug;37(14):e97.

30. Guo M, Feng H, Zhang J, Wang W, Wang Y, Li Y, Gao C, Chen H, Feng Y, He Z-G. Dissecting transcription regulatory pathways through a new bacterial one-hybrid reporter system. Genome Res 2009 Jul;19(7):1301–1308.

31. Balázsi G, Barabási A-L, Oltvai ZN. Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. Proc. Natl. Acad. Sci. U.S.A 2005 May;102(22):7841–7846.

32. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell 2000 Dec;11(12):4241–4257.

33. Hecker M, Völker U. General stress response of Bacillus subtilis and other bacteria. Adv. Microb. Physiol 2001;44:35–91.

34. Huet G, Daffe M, Saves I. Identification of the Mycobacterium tuberculosis SUF Machinery as the Exclusive Mycobacterial System of [Fe-S] Cluster Assembly: Evidence for Its Implication in the Pathogen's Survival. J. Bacteriol. 2005 Sep;187(17):6137–6146.

35. Manganelli R, Voskuil MI, Schoolnik GK, Dubnau E, Gomez M, Smith I. Role of the extracytoplasmic-function sigma factor sigma(H) in Mycobacterium tuberculosis global gene expression. Mol. Microbiol 2002 Jul;45(2):365–374.

36. Kaushal D, Schroeder BG, Tyagi S, Yoshimatsu T, Scott C, Ko C, Carpenter L, Mehrotra J, Manabe YC, Fleischmann RD, Bishai WR. Reduced immunopathology and mortality despite tissue persistence in a Mycobacterium tuberculosis mutant lacking alternative sigma factor, SigH. Proc. Natl. Acad. Sci. U.S.A 2002 Jun;99(12):8330–8335

37. Millington KA, Fortune SM, Low J, Garces A, Hingley-Wilson SM, Wickremasinghe M, Kon OM, Lalvani A. Rv3615c is a highly immunodominant RD1 (Region of Difference 1)-dependent secreted antigen specific for Mycobacterium tuberculosis infection. Proc. Natl. Acad. Sci. U.S.A 2011 Apr;108(14):5730–5735.

38. Rustad TR, Sherrid AM, Minch KJ, Sherman DR. Hypoxia: a window into Mycobacterium tuberculosis latency. Cell. Microbiol 2009 Aug;11(8):1151–1159.

39. Stewart GR, Wernisch L, Stabler R, Mangan JA, Hinds J, Laing KG, Young DB, Butcher PD. Dissection of the heat-shock response in Mycobacterium tuberculosis using mutants and microarrays. Microbiology (Reading, Engl.) 2002 Oct;148(Pt 10):3129–3138.

40. Stewart GR, Snewin VA, Walzl G, Hussell T, Tormay P, O'Gaora P, Goyal M, Betts J, Brown IN, Young DB. Overexpression of heat-shock proteins reduces survival of Mycobacterium tuberculosis in the chronic phase of infection. Nat. Med 2001 Jun;7(6):732–737.

41. Gao C-H, Yang M, He Z-G. An ArsR-like transcriptional factor recognizes a conserved sequence motif and positively regulates the expression of phoP in mycobacteria. Biochem. Biophys. Res. Commun 2011 Aug;411(4):726–731.

42. Py B, Barras F. Building Fe–S proteins: bacterial strategies. Nat Rev Micro 2010

Jun;8(6):436–446.

43. van Beilen JB, Neuenschwander M, Smits THM, Roth C, Balada SB, Witholt B. Rubredoxins involved in alkane oxidation. J. Bacteriol 2002 Mar;184(6):1722–1732.

44. Lawn SD, Zumla AI. Tuberculosis. The Lancet 2011 Jul;378(9785):57–72.

45. Russell DG, Barry CE, Flynn JL. Tuberculosis: What We Don't Know Can, and Does, Hurt Us. Science 2010 May;328(5980):852–856.

46. Kim M-J, Wainwright HC, Locketz M, Bekker L-G, Walther GB, Dittrich C, Visser A, Wang W, Hsu F-F, Wiehart U, Tsenova L, Kaplan G, Russell DG. Caseation of human tuberculosis granulomas correlates with elevated host lipid metabolism. EMBO Mol Med 2010 Jul;2(7):258–274.

47. Jamshidi N, Palsson BØ. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol 2007;1:26.

48. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011 Jan;39(Database issue):D561–568.

49. Beisel CL, Storz G. Base pairing small RNAs and their roles in global regulatory networks. FEMS Microbiol. Rev 2010 Sep;34(5):866–882.

50. Tu KC, Long T, Svenningsen SL, Wingreen NS, Bassler BL. Negative feedback loops

involving small regulatory RNAs precisely control the Vibrio harveyi quorum-sensing response. Mol. Cell 2010 Feb;37(4):567–579.

51. Yu Y-TN, Yuan X, Velicer GJ. Adaptive evolution of an sRNA that controls Myxococcus development. Science 2010 May;328(5981):993.

52. Massé E, Gottesman S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. Proceedings of the National Academy of Sciences 2002 Apr;99(7):4620 –4625.

53. Durand S, Storz G. Reprogramming of anaerobic metabolism by the FnrS small RNA. Molecular Microbiology 2010 Mar;75(5):1215–1231.

54. Vogel J, Papenfort K. Small non-coding RNAs and the bacterial outer membrane. Current Opinion in Microbiology 2006 Dec;9(6):605–611.

55. Waters LS, Storz G. Regulatory RNAs in bacteria. Cell 2009 Feb;136(4):615–628.

56. Thomason MK, Storz G. Bacterial Antisense RNAs: How Many Are There, and What Are They Doing? Annu. Rev. Genet. 2010 Dec;44(1):167–188.

57. Brantl S. Regulatory mechanisms employed by cis-encoded antisense RNAs. Curr. Opin. Microbiol 2007 Apr;10(2):102–109.

58. Storz G, Vogel J, Wassarman KM. Regulation by Small RNAs in Bacteria: Expanding Frontiers. Mol. Cell 2011 Sep;43(6):880–891.

59. Pfeiffer V, Papenfort K, Lucchini S, Hinton JCD, Vogel J. Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. Nat

Struct Mol Biol 2009;16(8):840–846.

60. Massé E, Gottesman S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. Proceedings of the National Academy of Sciences 2002 Apr;99(7):4620 –4625.

61. Sharma CM, Papenfort K, Pernitzsch SR, Mollenkopf H-J, Hinton JCD, Vogel J. Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA. Mol. Microbiol. 2011 Sep;81(5):1144–1165.

62. Arnvig KB, Young DB. Identification of small RNAs in Mycobacterium tuberculosis. Mol. Microbiol 2009 Aug;73(3):397–408.

63. Dichiara JM, Contreras-Martinez LM, Livny J, Smith D, McDonough KA, Belfort M. Multiple small RNAs identified in Mycobacterium bovis BCG are also expressed in Mycobacterium tuberculosis and Mycobacterium smegmatis. Nucleic Acids Res. 2010 Jul;38(12):4067-78

64. Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, Hinton JCD, Vogel J. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. PLoS Genet 2008;4(8):e1000163.

65. Schlüter J-P, Reinkensmeier J, Daschkey S, Evguenieva-Hackenberg E, Janssen S, Jänicke S, Becker JD, Giegerich R, Becker A. A genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium Sinorhizobium meliloti. BMC Genomics 2010;11:245.

66. Bohn C, Rigoulay C, Chabelskaya S, Sharma CM, Marchais A, Skorski P, Borezée-Durant E, Barbet R, Jacquet E, Jacq A, Gautheret D, Felden B, Vogel J, Bouloc P. Experimental discovery of small RNAs in Staphylococcus aureus reveals a riboregulator of central metabolism. Nucleic Acids Research 2010 Oct;38(19):6620 –6636.

67. Irnov I, Sharma CM, Vogel J, Winkler WC. Identification of regulatory RNAs in Bacillus subtilis. Nucleic Acids Research 2010 Oct;38(19):6637 –6651.

68. Miotto P, Forti F, Pellin D, Ambrosi A, Veiga DFT, Balazsi G, Gennaro ML, Di Serio C, Ghisotti D, Cirillo DM. Identification and Validation of Novel Small RNAs in Mycobacterium tuberculosis. *Submitted*

69. REHMSMEIER M, STEFFEN P, HÖCHSMANN M, GIEGERICH R. Fast and effective prediction of microRNA/target duplexes. RNA 2004 Oct;10(10):1507 –1517.

70. Tjaden B, Goodwin SS, Opdyke JA, Guillier M, Fu DX, Gottesman S, Storz G. Target prediction for small, noncoding RNAs in bacteria. Nucleic Acids Res 2006;34(9):2791–2802.

71. Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL. Thermodynamics of RNA-RNA binding. Bioinformatics 2006 May;22(10):1177–1182.

72. Tafer H, Hofacker IL. RNAplex: a fast tool for RNA-RNA interaction search. Bioinformatics 2008 Nov;24(22):2657–2663.

73. Busch A, Richter AS, Backofen R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics 2008

Dec;24(24):2849 –2856.

74. Schubert S, Grünweller A, Erdmann VA, Kurreck J. Local RNA target structure influences siRNA efficacy: systematic analysis of intentionally designed binding regions. J. Mol. Biol 2005 May;348(4):883–893.

75. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 1995;57(1):289–300.

76. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4(1):44–57.

77. Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol 2003;4(10):R70.

78. Burrus V, Waldor MK. Shaping bacterial genomes with integrative and conjugative elements. Res. Microbiol 2004 Jun;155(5):376–386.

79. Tsolaki AG, Gagneux S, Pym AS, Goguet de la Salmoniere Y-OL, Kreiswirth BN, Van Soolingen D, Small PM. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis. J. Clin. Microbiol 2005 Jul;43(7):3185–3191.

80. Shi L, Jung Y-J, Tyagi S, Gennaro ML, North RJ. Expression of Th1-mediated immunity in mouse lungs induces a Mycobacterium tuberculosis transcription pattern characteristic of nonreplicating persistence. Proc. Natl. Acad. Sci. U.S.A. 2003 Jan;100(1):241–246.

81. Olivarius S, Plessy C, Carninci P. High-throughput verification of transcriptional starting sites by Deep-RACE. BioTechniques 2009 Feb;46(2):130–132.

82. Boysen A, Møller-Jensen J, Kallipolitis B, Valentin-Hansen P, Overgaard M. Translational regulation of gene expression by an anaerobically induced small non-coding RNA in Escherichia coli. J. Biol. Chem 2010 Apr;285(14):10690–10702.

83. Mitarai N, Benjamin J-AM, Krishna S, Semsey S, Csiszovszki Z, Massé E, Sneppen K. Dynamic features of gene expression control by small regulatory RNAs. Proc. Natl. Acad. Sci. U.S.A 2009 Jun;106(26):10655–10659.

84. Caron M-P, Lafontaine DA, Massé E. Small RNA-mediated regulation at the level of transcript stability. RNA Biol 2010 Apr;7(2):140–144.

85. Papenfort K, Bouvier M, Mika F, Sharma CM, Vogel J. Evidence for an autonomous 5′ target recognition domain in an Hfq-associated small RNA. Proceedings of the National Academy of Sciences 2010 Nov;107(47):20435 –20440.

86. Peer A, Margalit H. Accessibility and Evolutionary Conservation Mark Bacterial Small-RNA Target-Binding Regions. J. Bacteriol. 2011 Apr;193(7):1690–1701.

87. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA websuite. Nucleic Acids Res 2008 Jul;36(Web Server issue):W70–74.

88. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Research 2007

Dec;36(Database):D866–D870.

89. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Letters 2004 Aug;573(1-3):83–92.

90. Zopounidis C, Pardalos P. Handbook of Multicriteria Analysis. Berlin Heidelberg: Springer; 2010.

91. Nedjah N, Mourelle L de M. Swarm intelligent systems. Berlin Heidelberg: Springer; 2006.

92. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H, Bonavides-Martinez C, Abreu-Goodger C, Rodriguez-Penagos C, Miranda-Rios J, Morett E, Merino E, Huerta AM, Trevino-Quintanilla L, Collado-Vides J. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Research 2007 Dec;36(Database):D120–D124.

93. Beisel CL, Updegrove TB, Janson BJ, Storz G. Multiple factors dictate target selection by Hfq-binding small RNAs. EMBO J. 2012 Apr;31(8):1961–1974.

94. Chen S, Zhang A, Blyn LB, Storz G. MicC, a Second Small-RNA Regulator of Omp Protein Expression in Escherichia coli. J. Bacteriol. 2004 Oct;186(20):6689–6697.

95. Urban JH, Vogel J. Translational control and target recognition by Escherichia coli small

RNAs in vivo. Nucl. Acids Res. 2007 Feb;35(3):1018–1037.

96. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell 2010 Apr;141(2):344–354.

97. Koebnik R, Locher KP, Van Gelder P. Structure and function of bacterial outer membrane proteins: barrels in a nutshell. Molecular Microbiology 2000 Jul;37(2):239–253.

98. Pos KM. Drug transport mechanism of the AcrB efflux pump. Biochim. Biophys. Acta 2009 May;1794(5):782–793.

99. Dhamdhere G, Zgurskaya HI. Metabolic shutdown in Escherichia coli cells lacking the outer membrane channel TolC. Mol. Microbiol 2010 Aug;77(3):743–754.

100. Stenberg F, Chovanec P, Maslen SL, Robinson CV, Ilag LL, von Heijne G, Daley DO. Protein complexes of the Escherichia coli cell envelope. J. Biol. Chem 2005 Oct;280(41):34409–34419.

101. Inoue T, Shingaki R, Hirose S, Waki K, Mori H, Fukui K. Genome-wide screening of genes required for swarming motility in Escherichia coli K-12. J. Bacteriol 2007 Feb;189(3):950–957.

102. Ibba M, Soll D. Aminoacyl-tRNA synthesis. Annu. Rev. Biochem 2000;69:617–650.

103. Laursen BS, Sørensen HP, Mortensen KK, Sperling-Petersen HU. Initiation of protein synthesis in bacteria. Microbiol. Mol. Biol. Rev. 2005 Mar;69(1):101–123.

104. Salmon K, Hung S, Mekjian K, Baldi P, Hatfield GW, Gunsalus RP. Global gene

expression profiling in Escherichia coli K12. The effects of oxygen availability and FNR. J. Biol. Chem 2003 Aug;278(32):29837–29855.

105. Chitsaz H, Salari R, Sahinalp SC, Backofen R. A partition function algorithm for interacting nucleic acid strands. Bioinformatics 2009 Jun;25(12):i365–373.

106. Vogel J, Luisi BF. Hfq and its constellation of RNA. Nature Reviews Microbiology 2011 Aug;9(8):578–589.

107. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most Mammalian mRNAs Are Conserved Targets of microRNAs. Genome Res. 2009 Jan;19(1):92–105.

108. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat. Genet. 2007 Oct;39(10):1278–1284.

109. Tafer H, Ameres SL, Obernosterer G, Gebeshuber CA, Schroeder R, Martinez J, Hofacker IL. The impact of target site accessibility on the design of effective siRNAs. Nat Biotech 2008 May;26(5):578–583.

110. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet 2006 May;38(5):500–501.

111. Spurgeon SL, Jones RC, Ramakrishnan R. High Throughput Gene Expression Measurement with Real Time PCR in a Microfluidic Dynamic Array. PLoS ONE 2008 Feb;3(2):e1662.

112. Cao Y, Wu J, Liu Q, Zhao Y, Ying X, Cha L, Wang L, Li W. sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. RNA 2010

Nov;16(11):2051–2057.

Vita

Diogo F. T. Veiga was born in the city of Chapecó, southern Brasil. He received his bachelor's degree in Computer Science from the Federal University of Santa Catarina, and later a master's degree from the University of Pernambuco. He also worked at the National Laboratory of Scientific Computing. He was awarded a CAPES/Fulbright doctoral scholarship in 2007 to study Biology at the University of Texas Health Science Center in Houston.