Experimental Characterization of the *Mycobacterium tuberculosis* Gene Regulatory Network

Kyle J. Minch

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

David R. Sherman, Chair

Jay Shendure

Colin Manoil

Program Authorized to Offer Degree

Molecular and Cellular Biology

University of Washington

Abstract

Experimental Characterization of the *Mycobacterium tuberculosis* Gene Regulatory Network

Kyle J. Minch

Chair of the supervisory committee:

Affiliate Professor David R. Sherman

Department of Global Health

Tuberculosis is a massive public health problem on a global scale and the success of *Mycobacterium tuberculosis* is linked to its ability to persist within humans for long periods without causing overt disease symptoms. Hypoxia is predicted to be a key host-induced stress limiting growth of the pathogen *in vivo*. However, studies indicate that *M. tuberculosis* coordinates a complex transcriptional program in response to long-term changes in oxygen tension *in vitro*, with the differential regulation of >20% of all transcriptional regulatory proteins. The results presented in this dissertation describe our efforts to create an experimental framework to define the control logic underpinning transcript regulation in *M. tuberculosis*. Creating a multi-platform experimental foundation for gene regulatory network construction is a critical step in generating predictive models of complex responses in prokaryotes. We describe a workflow that couples a defined perturbation to a regulatory response using chromatin immunoprecipitation followed by high throughput sequencing and transcriptional profiling by tiling microarray. We implement this experimental platform to reconstruct the transcriptional regulatory network of *M. tuberculosis* with particular attention to oxygen-responsive DNA binding proteins. This network allows us to generate predictive models of gene expression during an *in vitro* time course of hypoxia and reaeration. In this context, we describe the physiological consequences of aerobic induction of the early hypoxia-responsive regulator DosR. We find that *M. tuberculosis* growth is

unaffected upon the upregulation of *dosR* and the DosR regulon – in support of the hypothesis that growth inhibition as a result of oxygen limitation is mediated by a complex regulatory response. We also report studies that define the degradation rate of the mRNA pool in *M. tuberculosis* under different experimental conditions. We find that *M. tuberculosis* contains an unusually stable pool of mRNA and that this pool can be further stabilized by physiologically relevant alterations to the bacterial environment. There are obvious and pressing needs for greater understanding of the basic biology of *M. tuberculosis*, and this dissertation describes advancements we have made toward achieving this goal.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Dedication

To my family, to my friends.

Thank you.

Chapter 1 - Introduction

The following text has been modified slightly from a review article appearing in the journal Cellular Microbiology: Rustad, T. R., Sherrid, A. M., Minch, K. J. and Sherman, D. R. (2009), Hypoxia: a window into *Mycobacterium tuberculosis* latency. **Cellular Microbiology**, 11: 1151–1159. doi: 10.1111/j.1462-5822.2009.01325.x. Disease statistics have been updated to reflect the most current estimates, and additional information regarding the nature, and experimental characterization, of gene regulatory networks has been added.

Abstract

Tuberculosis is a massive public health problem on a global scale and the success of *Mycobacterium tuberculosis* is linked to its ability to persist within humans for long periods without causing any overt disease symptoms. Hypoxia is predicted to be a key host-induced stress limiting growth of the pathogen *in vivo*. However, multiple studies *in vitro* and *in vivo* indicate that *M. tuberculosis* adapts to oxygen limitation by entering in to a metabolically altered state, while awaiting the opportunity to reactivate. Molecular signatures of bacteria adapted to hypoxia *in vitro* are accumulating, though correlations to human disease are only now being established. Similarly, defining the mechanisms that control this adaptation is an active area of research. In this review we discuss the historical precedents linking hypoxia and latency, and the gathering knowledge of *M. tuberculosis* hypoxic responses. We also examine the role of these responses in tuberculosis latency, and identify promising avenues for future studies.

Introduction

Few infectious diseases in history have had the intimate, long-term association with humans and the profound global impact of tuberculosis (TB). Paleological evidence suggesting TB infections dates

back to the Neolithic Age, and reports of a disease very much like TB are found in the writings of the ancient Hindus and Greeks [1]. Hippocrates apparently went so far as to warn other physicians against treating advanced cases, lest the ensuing death of the patient mar the physician's reputation [1]. And even a cursory glance at global health today makes clear that TB remains one of humankind's most pernicious afflictions. At present, roughly four people in the world die of TB every minute. With about 9 million new cases of active disease and 1..45 million deaths annually (http://www.who.int/tb/publications/global_report/en/), TB is a global health emergency of massive proportions. The great brunt of the disease falls on the developing world. However, synergy with the AIDS epidemic [2, 3], emigration from TB-endemic countries [4], and the recent surge of multi-drug-resistant (MDR) cases [3] insure that TB remains a top public health priority in the developed world as well. The need for improved understanding of TB and the factors that promote its survival and spread has never been more urgent.

Mycobacterium tuberculosis (MTB) usually spreads via aerosols produced when someone with active disease coughs. When aerosol droplets harboring infectious bacilli lodge in the lung of a new host, several outcomes are possible [5]. About two-thirds of people exposed to TB (close contacts of active cases) never show any sign of infection [6]. The basis for this innate protection is entirely unknown. In other people, the bacilli survive and multiply enough to elicit an adaptive immune response, generally manifest as skin-test sensitivity to the MTB antigen mix known as PPD. Those who convert to PPD+ in the absence of vaccination with the attenuated *M. bovis* strain, BCG, are said to be infected with TB, even in the absence of clinical symptoms. Nearly half of the people who convert to PPD+ develop active disease within one year [7]. The rest are generally thought to harbor a latent infection, although it is not obvious how many persons clear the bacilli while remaining PPD+. Still, in a real sense latency shapes the current TB pandemic. The WHO estimates that 1.86 billion people (32% of

the world population) are PPD+ [2], and this vast reservoir of latency is a constant source of reactivation disease (**Figure 1-1**).

The situation is especially grave when MTB and HIV co-infect. A person with latent tuberculosis has about a 10% lifetime chance of developing active disease [2]. When such a person contracts HIV, the risk of developing reactivation TB increases to 8 – 10% per year [2, 8]. Since at least 15 million persons are co-infected at present, roughly 10% of all active TB or more than one million cases per year are directly attributable to complications of HIV [8]. The prognosis for these patients is extremely poor. Worldwide, TB kills more AIDS patients than any other opportunistic infection, accounting for about 30% of AIDS deaths [8].

Latency and hypoxia

From a variety of human, animal and *in vitro* studies, it is apparent that oxygen tension is intimately associated with the outcome of MTB infection. Numerous examples, outlined below, demonstrate that MTB growth, metabolic activity, and transcriptional profile are acutely sensitive to varying levels of oxygen *in vitro* and *in vivo*. These observations have led to a model of TB latency and reactivation that postulates a prominent role for oxygen status [9-12]. In brief, this model holds that MTB deposited in the lung grows unabated for days or weeks within alveolar macrophages until an appropriate adaptive immune response is mounted. Then activated macrophages and other host components surround the infected cells in an organized display, a granuloma, creating conditions that are no longer permissive for MTB replication. In humans, this stage can lead to a variety of outcomes, both in terms of disease and pathology. Granulomas can evolve morphologically, producing areas of caseous necrosis in some cases and fibrous, calcified deposits in others. Morphologically distinct lesions can co-exist within a single patient's lungs. Granulomas are thought to limit MTB growth by restricting bacterial access to oxygen and nutrients and exposing the bacilli to acidic pH and immune effectors such as nitric oxide. However, the bacilli are not necessarily eradicated. Instead, they may adapt to a state of

bacteriostasis or very slow replication that can persist for many years. These bacteria within the granuloma are viewed as the seeds of reactivation TB, waiting until HIV or some other factor restores conditions permissive for active disease. This model forms the framework for most recent discussions of TB latency, although many key features including the nature of changes in oxygen tension and their relevance to TB latency and reactivation have never been adequately tested, and recent evidence indicates that the granuloma may actually facilitate early growth and dissemination of the bacilli [13]. Here, we will review the origins of the hypothesis that oxygen levels are linked to the transition between active and latent disease, and then describe recent advances and interesting new directions in exploring this model.

Hypoxia in human and animal studies

Several lines of tantalizing but indirect evidence link changes in oxygen tension with varying TB disease. TB infections are preferentially associated with the most oxygen-rich sites within the body [14], suggesting that O_2 availability is one factor that constrains MTB growth *in vivo*. Within the lungs of patients failing TB chemotherapy, high bacterial numbers are only found in lesions directly connected with open airways whereas lesions lacking direct contact with air are paucibacillary [15]. In addition, reactivation from latency in humans occurs most frequently in the upper lobes of the lung [14], the single most-oxygenated region of the body. In contrast, the most oxygen-rich body site of four-legged creatures like rabbits and cows is the dorsal portion of the lungs, the very area where TB localizes in these animals. In an extreme test of the link between TB localization and oxygenation, Medlar maintained TB-infected rabbits in an upright position for 11 hours a day [16]. Strikingly, the TB lesions in these animals were predominantly in the apical lung zones.

While reactivation occurs most in highly oxygenated regions, reduced oxygen tension is associated with inhibition of MTB growth and TB latency. Maintaining infected animals at reduced O_2 tension slows disease progression, with less pathology [17]. Human epidemiology suggests the same

trend, with rates of TB transmission and/or disease decreasing at higher altitudes where oxygen tension is lower [18]. More pronounced oxygen deprivation has a more dramatic effect. In dogs, collapse of an infected lung resulted in marked inhibition of disease progression relative to the normal lung in the same animal [19]. Similar results are evident in humans, where artificial lung collapse became the pinnacle of TB therapy in the pre-antibiotic era. This treatment, performed on many thousands of patients, appears to have been effective at arresting TB, albeit with horrendous and often fatal complications [20].

The role of hypoxia in current animal models

Various animal models of TB disease are used currently [5], though their relevance to TB latency and reactivation in humans remains unsettled (**Table 1-1**). Mice are most commonly used for reasons of cost, convenience, the availability of inbred strains, and immunological reagents [21]. However, unlike latent TB in humans, mice maintain a high bacillary burden throughout infection. Drug interventions can generate a latent-like state in mice, the results of which have proven challenging to reproduce [22]. In addition, the mouse may be particularly unsuited for studying a link between hypoxia and MTB persistence. Granulomas in mice are usually small and poorly differentiated, and recent evidence indicates that TB lesions in mouse lungs are not especially hypoxic [10, 23]. Caseation and cavitation, two hallmarks of human TB associated with dramatic changes in oxygen tension, are rarely or never seen in infected mice. Finally, MTB continues to replicate during chronic infection of mice, indicating that the bacilli are never truly dormant in this system [24].

Though mice are the most common animal model of TB, other model systems may be more useful for exploring aspects of latency. Guinea pigs and rabbits have been used to model TB for at least a century, in part because their lesions show caseous necrosis similar to that seen in humans, and recent evidence indicates that mature lesions in guinea pigs and rabbits are hypoxic [10]. However, guinea pigs are exquisitely sensitive to TB, with no evidence for a latent phase, while rabbits must be infected with

M. bovis because they are resistant to MTB [21]. Zebrafish are sensitive to infection with *Mycobacterium marinum*, and this fascinating system is increasingly used to model TB [21]. However, the oxygen tension in zebrafish lesions has not been reported, and the role of this model for latency studies must still be determined. Primates produce TB disease that is strikingly like that of humans, including hypoxic, caseating lesions and what appears to be a latent state in about 40% of infected animals [21, 25]. Still, use of this model will always be limited. Of course, experiments in any animal system can require a year or more to complete, and because few validated molecular correlates of the latent state in humans are known, the value of each model remains difficult to assess. As a result, generating molecular correlates of latency is an important objective of current TB research.

***In vitro* systems: The Wayne model**

To characterize better the state of MTB in reduced oxygen, Larry Wayne and colleagues developed an *in vitro* system that relies on a self-generated O_2 gradient. In the Wayne model, a sealed, standing culture is allowed to incubate over a period of days while the bacteria deplete the available oxygen. The culture becomes progressively more hypoxic with a concomitant shift in MTB physiology [26, 27]. Gentle stirring and a defined culture-to-headspace ratio improves reproducibility [28]. Two distinct states of non-replicating persistence (NRP stage 1, stage 2) are evident. These stages reflect discrete metabolic and drug-susceptibility states compared to log phase growth.

Attracted by the relative ease of the Wayne model, several researchers have used this approach to characterize non-replicating, microaerophilic bacilli [29]. As might be expected, translation is severely reduced in these cells [30]. However, against this background of declining protein synthesis, expression of certain genes is enhanced. The alternative sigma factors *sigB, sigE and sigF* are all up-regulated [31]. Alternative sigma factors promote broad changes in transcription, so expression of these genes may help drive the adaptation to persistence. Recently, Wayne model studies have identified NAD synthesis,

maintenance of the protonmotive force, and ATP synthesis as essential processes in non-replicating bacteria, and has aided in the development of compounds effective against hypoxic MTB [11, 32-34].

In vitro systems: The defined hypoxia model

Despite improvements, the Wayne model is still dogged by issues of reproducibility [28]. In response, some researchers expose bacilli to defined hypoxic atmospheres that remain constant throughout the experiment. In this approach, oxygen tensions of 1% or less halt replication but bacteria remain viable [35-37].

Although their value in generating testable hypotheses is clear, the *in vitro* models of hypoxia-induced bacteriostasis must still demonstrate their relevance to TB latency *in vivo*. How precisely do the bacteria adapt to initial stages of hypoxia? What are the transcriptional, proteomic and metabolic changes that characterize hypoxic bacteria? Can these approaches be used to model reactivation? And most important, how might the molecular signatures of hypoxic adaptation defined *in vitro* inform our understanding of human TB disease? Recent work using *in vitro* hypoxic models is shedding some light on these questions.

The Initial Hypoxic Response and DosR

With the advent of the MTB genome sequence it became possible to characterize the whole transciptomic response to hypoxia using microarrays. In a series of experiments utilizing the defined hypoxia model, our lab described the initial response to hypoxia, which consisted of ~100 genes with altered expression [38]. The repression of well-characterized genes involved in protein synthesis, DNA synthesis/cell division, lipid or amino acid synthesis and aerobic metabolism indicated a defined shift to reduced metabolic activity. The 47 induced genes are mostly uncharacterized, but among those of known or predicted function are several that may play a role in adaptation to hypoxic stress: *acr* (stabilizing partially denatured proteins); *narX*, *nark2*, and *fdxA* (nitrate accumulation and alternative

electron transport); *nrdZ* (dNTP synthesis under microaerophilic conditions); *tgs1* (triglyceride synthase) and six MTB orthologues of the universal stress protein (Usp) family (resistance to DNA damage).

The co-regulation of this gene set in response to hypoxia suggested a shared mechanism of transcriptional control. One of the induced loci contained both parts of the two component transcriptional regulatory system DosR/DosS (sometimes called DevR/DevS [39]). To assess the role of this regulator in the initial hypoxic response, we performed targeted disruption of this locus and compared the transcriptional profile of wild type and the Δ*dosR* mutant strain. Nearly all the genes initially induced by hypoxia require DosR for their induction [38, 40]. Further work by Voskuil *et al*. showed that the DosR-dependent regulon is also induced in response to nitric oxide, another stress that induces bacteriostasis *in vitro* [41]. Later studies showed that the DosR regulon is also induced in macrophages [42], in both early [43] and late mouse infections [44], and in response to carbon monoxide [45, 46], SDS [47], and low pH [48].

The DosR transcriptional response has become one of the best characterized signal transduction systems in mycobacteria. *In silico* analysis of the DosR-dependent hypoxia induced genes identified a 20-mer degenerate palindromic motif associated with genes that respond to the initial hypoxic stress [40, 49]. A variant of this consensus sequence is located upstream of nearly all MTB operons rapidly induced by hypoxia. Mutations within this site abolish both DosR binding and hypoxic induction of a downstream reporter gene. Additional mutation experiments confirmed sequence-based predictions that the DosR C-terminus is responsible for DNA binding and that the conserved aspartate at position 54 is essential for function [40]. Crystallography experiments demonstrated that DosR forms a tetramer that interacts with DNA at Lys179, Lys182, and Asn183 [50]. DosR is phosphorylated to its active state by two histidine kinases, DosS and DosT [51]. Both kinases bind heme as a prosthetic group [52]; DosS binds to ferrous iron and senses a shift to ferric iron as an indicator of the net redox state of the cell, while DosT binds to heme bound to O_2 and is activated by the disassociation of O_2, directly sensing the

oxygen tension. Altogether these results confirm that DosR is a transcription factor of the two-component response regulator class that functions as initial mediator of the MTB responses to hypoxia, nitric oxide, and carbon monoxide.

The role of the DosR response

DosR is widely considered essential for TB latency. This idea is grounded in the arguments summarized above associating hypoxia with a latent disease state, and because hypoxia and the other DosR triggers can all induce a non-replicating (dormant) phenotype. Its very name is an abbreviation for **do**rmancy **s**urvival [53]. However several lines of evidence suggest that the link between the DosR regulon and latency is tenuous at best.

The initial hypoxic response controlled by DosR features powerful induction of many genes, which led us to predict a strong phenotype for the MTB Δ*dosR* mutant. However, *in vitro* the mutant was able to enter bacteriostasis with the same dynamics as wild type and showed no survival defect until three weeks in the Wayne model, when the mutant showed a 1-2 log survival defect. This is weeks after the time the DosR regulon is induced in wild-type bacilli [54]. The BCG Δ*dosR* mutant is reported to have a stronger survival defect in the Wayne model [53], but we have been unable to confirm that observation (unpublished).

Experiments to test the link between the DosR regulon and non-replicating persistence *in vivo* have been equivocal. Parish and Stoker reported hypervirulence of the Δ*dosR* mutant in SCID and immunocompetent mice [55]. In contrast, we found that loss of *dosR* produced no discernable phenotype in three separate mouse strains, even though expression of DosR regulon genes was perturbed [54]. Others have found very modest phenotypes of the Δ*dosR* mutant in guinea pigs, mice, and rabbits [56, 57]. The alpha-crystallin gene (acr) is highly induced and tightly regulated by DosR, making it a good marker for induction of the DosR regulon [37, 38]. More than three-quarters of TB patients with active disease express antibodies directed against Acr [58], and abundant *acr* transcript

has been detected in infected mice [44]. However, *acr* mRNA appears several weeks prior to the chronic phase of bacterial growth in the mouse model [54]. Acr is also powerfully induced when MTB is growing within macrophages [37] and in interferon-γ-deficient mice [44] in which there is no latent phase. These experiments indicate that, while the DosR regulon may be expressed at some points during latency, these genes are also expressed at times of active growth. Altogether the *in vivo* data suggest that the DosR regulon plays some role in dormancy, but is not the key to establishing or maintaining a non-replicating state, and may also play another role during infection.

Variation in DosR regulation by clinically distinct strains of MTB has shed more light on the role of the initial hypoxic response. The W-Beijing lineages of MTB, associated with epidemic spread and increased drug resistance worldwide, constitutively overexpress the DosR regulon [59]. However, these strains grow normally in culture and the limited evidence available so far does not suggest that these strains are more likely to enter into a latent state in the host [60]. Constitutive over-expression of the DosR regulon in these strains leads to the accumulation of triglycerides, which may confer a slight advantage in microaerophillic conditions and in response to NO stress. This observation reinforces the interesting and potentially significant role that the DosR regulon plays in TB infection, but in a context removed from dormancy.

In summary, expression of DosR regulon genes is not synonymous with 'the dormancy response' of MTB. Evidence from several groups shows that the DosR regulon is induced during stages of active growth and that loss of DosR has only mild phenotype in many models of dormancy. Rather than the dormancy response, the DosR regulon may play a much more focused role in survival of extended respiratory, nitrosative or redox stress. These observations led us to investigate hypoxic responses of MTB that are downstream of DosR.

The Rest of the Iceberg: The Enduring Hypoxic Response

If the DosR regulon is not essential for promoting bacteriostasis, the genes responsible for induction or maintenance of latency may remain uncharacterized. To identify these genes we extended the transcriptional analysis of the hypoxic response to later time points. As mentioned above, the DosR response consists of ~50 rapidly induced genes. However, by eight hours of oxygen limitation, the total number of induced genes had doubled to more than 100. That total doubled again by 24 hours and again by 4 days of hypoxia to ~400 genes induced relative to log phase (**Figure 1-2**). Though some of the genes induced at each point were unique to that time, almost half of the genes induced at any time point endured for the remainder of the experiment. The intersection of genes induced between four and seven days of hypoxia is a set of 230 genes that we named the Enduring Hypoxic Response (EHR). Thirty of these genes are transcriptional regulators, suggesting a highly complex and possibly redundant web of regulation. The extended induction of the EHR genes suggests they may have important roles in non-replicating persistence.

One of the roles posited for the DosR regulon was the establishment of a basal hypoxic response that would lay the groundwork for subsequent adaptations [38]. To test this idea we examined how loss of DosR affects downstream gene expression. Surprisingly, all but a handful of the hundreds of genes induced after the initial hypoxic response are DosR independent. Some of the genes of the DosR regulon remain induced at later time points, though not to the levels of the initial response. The DosR-independent EHR is larger and more powerfully induced than the DosR response at all points after 2 hours of hypoxia (**Figure 1-2**). This provides further indication the DosR response is not essential to establish a non-replicating persistent state.

Alternative sigma factors can function as master regulators of stress responses by controlling broad transcriptional responses. Sigma factors interact with the -10 and -35 regions of the promoter region and are essential parts of RNA polymerase. The EHR includes two such genes: *sigE* and *sigH*.

These genes have been previously characterized as controlling responses important to the survival of heat shock, SDS detergent cell surface stress and oxidative stress, as well as during the growth of MTB in human macrophages [31]. We have begun dissecting the EHR by analyzing mutants of these two sigma factors and preliminary results suggest that up to a quarter of the EHR may be disregulated in a mutant lacking either *sigE* and *sigH* (Rustad and Sherman, unpublished). Like the DosR mutant, the *ΔsigEΔsigH* mutant is able to enter bacteriostasis in response to hypoxia and shows a survival defect in response to hypoxic stress. These genes are obviously important to hypoxic survival, but we predict disruption of genes controlling more of the EHR will yield a stronger phenotype.

Hypoxic MTB: Today and Tomorrow

Current research involving hypoxia and MTB has the potential to greatly enhance our understanding of human TB disease. The detailed analysis of the MTB transcriptional response to hypoxia, coupled with the recent confirmation that MTB granulomas are hypoxic in the primate model, opens up many potential lines of inquiry. New methods and advances in the field give us improved tools to answer critical questions. Do the adaptations of MTB to hypoxia *in vitro* correlate to what occurs in hypoxic granulomas? What are the core regulators of the MTB hypoxic response, and are they necessary for establishment of latent disease? What are the roles of the DosR regulon and the EHR *in vivo*? Can we determine which transcriptional profile and metabolic state MTB adopts in various lesion types? Furthermore, can we pinpoint mechanisms to kill bacteria more effectively in these states, or identify predictors of reactivation from latency?

The first vital question to explore is how closely the hypoxic models described above parallel the environment of MTB and its response to those conditions during latent infection. Analyzing the MTB transcriptome during latency in animal models and humans has been challenging because MTB transcripts are not abundant in these lesions and the host mRNA is much more abundant. Sensitive quantitative real-time RT-PCR and high-throughput expression sequencing may be suitable tools to

characterize the transcriptome of MTB in these lesions. We anticipate that there will be substantial overlap between MTB genes expressed in hypoxic granulomas from latent infection and the *in vitro* hypoxia models, though signals other than hypoxia are doubtless also relevant. Similarly, we expect that the MTB in morphologically distinct lesions will probably express distinct transcriptional profiles. These differences are likely to be biologically relevant, but to explore them we will need more information about how granulomas and the bacteria they contain evolve over time.

However, gene expression analyses will not tell the whole story of the MTB response to hypoxia. Systems biology approaches are just beginning to be applied to the analysis of non-replicating MTB [61], and we expect further work to contribute materially to studies of latent TB. The dynamic nature of the niche in which MTB resides in the host requires that viable bacteria be able to adapt to multiple, temporally-overlapping, stimuli. Analyses like those described for the EHR above suggest that underlying the bacterial response to such a mutable environment is a complex regulatory network; however, a directed approach to experimentally characterize a complex regulatory response has not been done in *M. tuberculosis*. As such, our current understanding of these systems in MTB relies heavily on general principles of transcription regulation networks characterized in model organisms.

Transcriptional regulatory networks in different organisms appear to be populated by recurring motifs and themes [62]. Promoter regions in regulatory networks of *Escherichia coli* tend to be occupied by a limited number transcription factors [63], and a similar pattern is observed in *Saccharomyces cerevisiae* [64]. Conversely, a transcription factor can theoretically interact with any number of promoters/genes in the network, but in actuality most transcription factors functionally interact with a limited number of targets. Increasingly smaller numbers of transcription factors interact with greater and greater numbers of promoters/genes. This has been observed in *E. coli, S. cerevisiae,* and in an *in silico* analysis of *M. tuberculosis* [61, 63, 64]

Experimental approaches to determine genome-wide transcription factor association include gene knock-outs and promoter occupancy studies. The first method involves creating individual transcription factor knock-out strains with subsequent measurement of the transcriptional response to a certain experimental condition as compared to the wild-type response. This approach will provide a list of differentially or disregulated genes/targets, but does not capture whether the interaction between transcription factor and target is direct or through an intermediate regulatory protein [65]. The second approach makes use of chromatin immunoprecipitation (ChIP) to isolate a specific transcription factor with associated covalently cross-linked DNA via antibody pull-down. The DNA is subsequently purified and hybridized to a microarray (ChIP-chip) or sequenced on a next generation platform (ChIP-seq). ChIP-chip requires microarrays with dense tiling of the genome in order to obtain resolution of a binding event, and true signal identification on the array platform is subject to limitations based on background fluorescence effects, whereas the dynamic range of high-throughput sequencing platforms is far greater [66]. Defining interactions by ChIP alleviates the confounding issue of indirect interactions that are faced in knock-out studies as described above; however, the limitation exists that binding at a given genomic site does not necessarily imply that expression levels of a target gene change. To distinguish between "active" or "silent" transcription factor binding it is necessary to measure RNA levels in response to a perturbation, e.g., induction of a transcription factor (the perturbation) with concomitant ChIP and RNA sampling.

Data from these experiments should provide insight in to both the breadth and connectivity of the MTB transcription regulatory network as it adapts to multiple, potentially overlapping, stimuli. We predict that several basic features will emerge. It is likely that few regulators functionally interact with many promoters, and many regulators functionally interact with few promoters; furthermore, some promoters are likely regulated by multiple transcriptional regulators, though the number of regulators affecting a particular gene or operon will fluctuate over a very narrow range. Given the scope of

differential gene regulation upon long-term exposure to hypoxia (>20% of all genes) it is extremely likely that in leveraging this stimulon as our baseline experimental model, we will uncover regulatory interactions far beyond the limits of the EHR (those genes induced at 4 and 7 days hypoxia).

In addition to the regulatory network studies described above, future work studying the changing abundance of proteins, lipids, or metabolites should be informative. For example, proteomic analysis could be applied to these *in vitro* models to determine how the cellular protein pool changes during hypoxic conditions. With a complex, lipid-rich cell wall and a large portion of the genome devoted to genes involved in lipid synthesis and metabolism, evaluating the changes in lipid moieties could provide insight on the altered metabolism and cell wall modifications made in hypoxia-adapted MTB. Finally, while some metabolic pathways are known to have a role during infection, the specific metabolic adaptations MTB makes during latent infection are not clear. Recent work is beginning to bring these adaptations to light [11, 32], promoting a broader view of how MTB adapts to a hypoxic environment and allowing more informed design of drugs to kill metabolically-altered MTB.

Another important area is the progression of granuloma morphology over the course of infection. Which lesions become hypoxic, and which will provide favorable conditions for reactivation at a later date? It has been suggested that the metabolic state of MTB within hypoxic granulomas may render them drug tolerant [9], but neither the specific type of lesion nor the transcriptional profile associated with this state is understood. With the recent advances in the primate model and its striking similarity to human latency and reactivation, these hypotheses may now be testable. Imaging technologies such as positron emission tomography (PET) and computed tomography (CT) could be valuable tools to investigate the progression of a granuloma as it establishes a hypoxic environment. These imaging tools could be used in combination with quantitative real-time PCR or high-throughput sequencing to determine how MTB gene expression differs among lesion types and different stages of

infection. It may even be possible to identify attributes of the granuloma that predict the establishment of bacterial non-replication or re-initiation of replication.

Finally, the major impact of latent TB on the tuberculosis epidemic is its ability to reactivate to fully transmissible and frequently deadly active disease at an unpredictable future time point. Little is known about the mechanism by which MTB reactivates from latent infection. The tools now being applied to study hypoxia-adapted MTB could also be utilized to determine how MTB adapts during reactivation to active disease. Reaeration from hypoxia may provide a useful model to study MTB reactivation *in vitro*.

Recent efforts have set the stage for important future advances in understanding latent MTB. *In vitro* hypoxic models have been developed and are providing insight on how MTB survives in a non-replicating state. The hypothesis that MTB in granulomas are exposed to a hypoxic environment has recently been confirmed [10]. Studies on the progression of granulomas over the course of infection have great potential to reveal the bacterial environment *in vivo* and how MTB responds to challenges posed by the host. It must be determined how closely the *in vitro* models correlate to the state of MTB in latent infection. However, if these models are predictive of human disease, the information they provide in combination with advances in animal models, imaging, and analysis will substantially aid in the development of drugs capable of killing MTB in altered metabolic states, and possibly shortening the course of TB therapy.

The results presented in this dissertation represent the products of investigation based on the intellectual framework described above. Chapter 2 describes the first integrated approach to interrogate genome-wide transcription factor-DNA associations incorporating the transcriptional impact of that perturbation in a prokaryote, and provides a description of the constructs and methods employed in chapters 3 and 4. Chapter 3 describes efforts to experimentally develop a condition-agnostic predictive gene regulatory network for *M. tuberculosis* that is subsequently tested under

microaerophilic conditions. Chapter 4 details the aerobic expression of the transcriptional regulator DosR and its associated regulon, with a discussion of the physiological implications of this induction. Chapter 5 details work to characterize the degradation rate, and mRNA half-life, of the transcript pool of *M. tuberculosis* under multiple experimental conditions. The final chapter of this dissertation is a brief synthesis of the preceding sections, detailing efforts to integrate multiple threads of gene regulation in tuberculosis, and concludes with a consideration of certain outstanding questions surrounding tuberculosis gene regulatory networks and systems biology.

Table 1. Animal Models of Tuberculosis : Relevance to Latency		
Model	**Advantages**	**Disadvantages**
Mouse	Plentiful immunologic reagents Transgenic and knockout strains Low cost and space req.	No true latent stage Non-caseating granulomas Granulomas not hypoxic
Zebrafish	Transparent embryo stage Native host/pathogen pair BL2 Level Pathogen Low cost and space req. Caseating granulomas (adult)	Potential differences between *M.marinum* and MTB Limited immunological reagents
Guinea Pig	Caseating granulomas Granulomas are hypoxic	No true latent stage Limited immunological reagents Moderate cost and space req.
Rabbit	Caseating granulomas, cavitation Granulomas are hypoxic Paucibacillary stage (MTB)	*M.bovis* virulent, MTB less so Limited immunologic reagents Higher cost and space req.
Non-Human Primate	Caeseating granulomas, cavitation Granulomas are hypoxic Paucibacillary latency Reactivation of latent infection Plentiful immunologic reagents	Very high cost and space req.

Table 1-1: Animal models of tuberculosis.

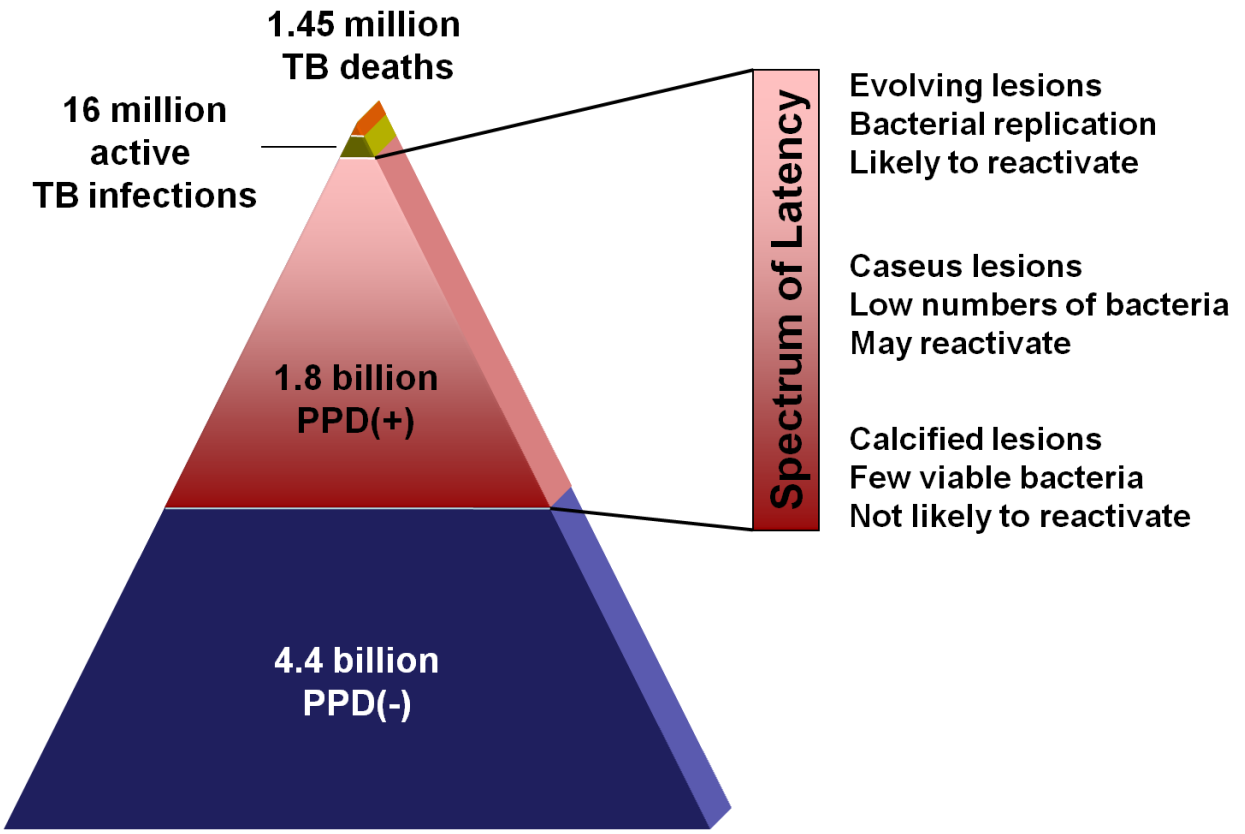


Figure 1-1: WHO estimates of the global burden of tuberculosis. Each section of the pyramid drawn roughly to scale. The bar to the right represents the spectrum of different lesions types seen in latent infections.

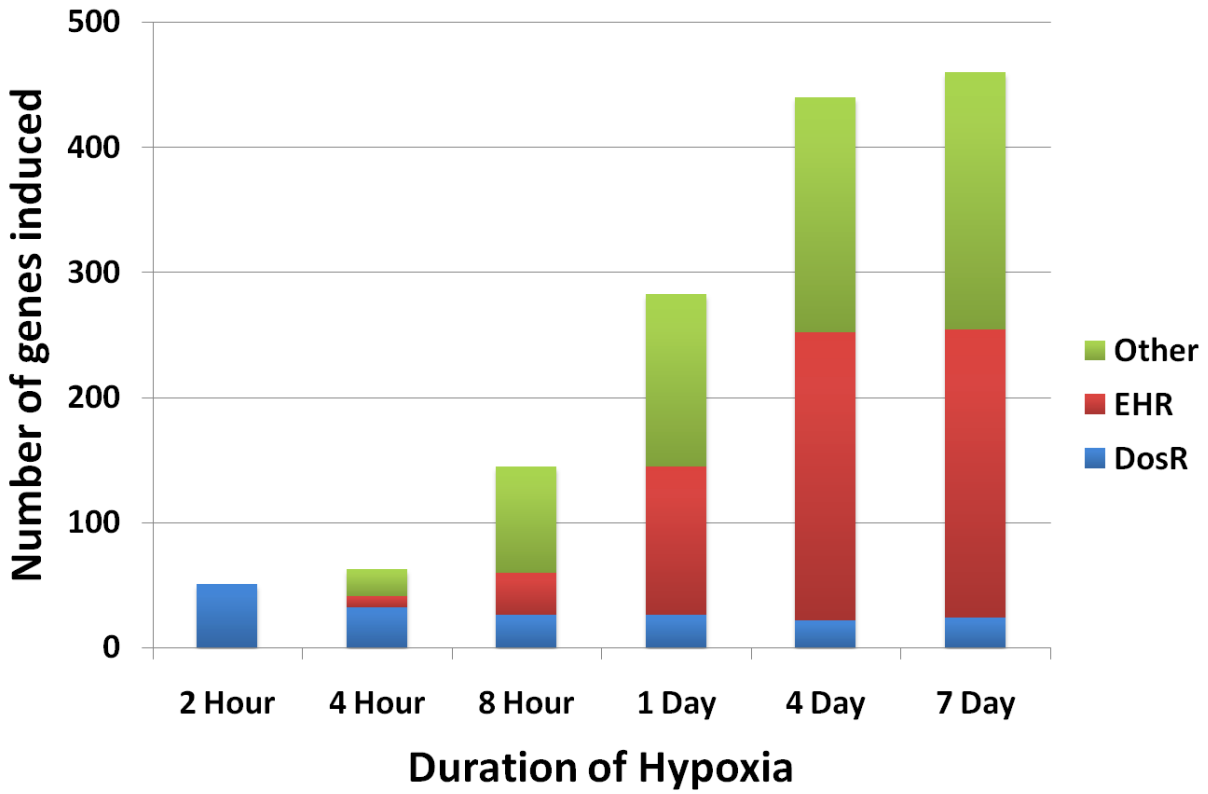


Figure 1-2: *M. tuberculosis* gene expression in hypoxia. Hypoxic gene expression is rapidly dominated by the Enduring Hypoxic Response (EHR). Each bar represents the total number of genes induced at that hypoxic time point. The bars are divided into the genes of the DosR response (blue), the EHR (red), and other induced genes (green).

Chapter 2 – An Experimental Workflow to Characterize DNA Binding Proteins in Prokaryotes

The methods described in this chapter feature prominently in work described in chapters 3 and 4 of this dissertation, and contribute to a lesser extent to the content of chapter 5. Supplementary figures are indicated by the form "Figure S2-x," where "x" denotes the order of the figure in the text.

Abstract:

Generating a multi-platform experimental foundation to gene regulatory network construction is a critical step in generating predictive models of complex behaviors in prokaryotes. To date, transcriptional profiling under diverse environmental conditions provides the bulk of the data used in these efforts; however, introducing interaction constraints, such as those offered by associating DNA binding proteins with their targets, has the potential to improve the predictive power of regulatory models. To this end, we describe a workflow to capture the genome-wide DNA binding patterns of prokaryotic transcription factors, and the associated transcriptional change attendant with that binding. We uniformly induce the expression of epitope-tagged transcription factors from multiple transcription factor families and two different bacterial species: *Mycobacterium tuberculosis* and *M. smegmatis*, and using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) we investigate the genome-wide binding properties of proteins from nearly 15 different protein families.

Introduction

Gene regulatory networks underlie the complex responses of biological systems to dynamic environments [62]. Direct experimentation to understand the transcriptional control mechanisms of such networks can be accomplished in a number of ways, including genome-wide measurements of transcript levels (e.g., by microarray or RNA-sequencing) and protein-DNA interactions through the

application of chromatin immunoprecipitation followed by hybridization to a microarray (ChIP-chip) or ChIP followed by high-throughput DNA sequencing (ChIP-seq). There are notable examples of large-scale attempts to comprehensively and uniformly interrogate DNA binding patterns of transcription factors in eukaryotic model organisms [64, 67, 68]; however, characterization of prokaryotic gene regulatory networks has primarily been achieved through meta-analyses incorporating decades of genetic studies [69-71], or transcriptional profiling under multiple environmental conditions to generate clusters of co-regulated genes and DNA-binding proteins [72, 73]. These approaches offer the potential to uncover previously obscure regulatory connections within an organism; however, because they are based on retrospective analyses or inference-based modeling, they may overlook *in-trans* or subtle interactions. The combination of co-regulatory transcriptional patterns with implementation of physical constraints afforded by protein-DNA binding studies offers an attractive method for generating regulatory models with interaction boundaries imposed by multiple data types. Here, we describe a generalizable method for experimentally deriving genome-wide DNA-binding patterns of proteins, as well as the transcriptional impact of that defined perturbation using two mycobacterial species: *M. smegmatis*, and *M. tuberculosis*.

M. smegmatis is an environmental Mycobacterium with a high-GC content genome that is often used as a model organism to study certain characteristics of *M. tuberculosis* due to its relatively fast replication rate and minimal biosafety concerns [74]. *M. tuberculosis* is a slow-growing human pathogen of tremendous global health importance (http://www.who.int/tb/publications/global_report/en/). While primarily a pulmonary pathogen infecting alveolar macrophages, the bacterium is able to survive in multiple diverse cell types and tissue niches within the host [75, 76]. These dissimilar environments impose varying adaptive demands on the bacterium, and the first whole-genome sequencing effort of a commonly studied strain, H37Rv, indicated that there were ~180 putative accessory DNA-binding proteins and 13 sigma factors in the genetic repertoire of *M. tuberculosis* [77].

In the present work, chapter 2 of this dissertation, we describe an inducible expression system with uniformly epitope-tagged DNA-binding proteins. We detail an integrated method to characterize binding sites upon gene induction using ChIP-seq, and the corresponding transcriptional impact of that perturbation using genome-wide tiling microarrays. Furthermore, we provide evidence for the applicability of this method for prokaryotic accessory transcription factors from diverse (>15) protein families. Studying DNA-binding patterns in both *M. tuberculosis* and *M. smegmatis* we propose that this method may be broadly applicable to genetically tractable high-GC content prokaryotes. This experimental platform forms the basis for data integration and modeling described in chapter 3 of this dissertation.

Materials & Methods

Construction of Expression Vectors and Strains:

Creation of pDestination Tet N-/C-terminal FLAG Tag

pDEST-Tet, an episomally-replicating vector containing an anhydrotetracycline (ATc)-inducible promoter [78] upstream of a Gateway recombination cassette (Invitrogen), was modified to encode for an N- or C-terminal FLAG epitope tag. Utilizing a stitching PCR strategy the FLAG sequence (5' - gactacaaggacgacgacgacaag) was inserted in-frame either upstream or downstream of the pDest-Tet AttR sites. In the case of the N-terminal FLAG tag, a start codon was added to the 5' terminus of the FLAG sequence. In the case of the C-terminal FLAG tag, a stop codon was added to the 3' terminus of the FLAG sequence. Described is the process for creating the C-terminal FLAG vector, though the strategy employed for creating the N-terminal FLAG vector was conceptually identical. Utilizing primers KJM10F/R and KJM11F/R PCR products were created using pDEST-Tet as a template, and adding the "sense" FLAG sequence to the 3' end of the KJM10 PCR product and the "antisense" FLAG sequence to the 5' end of the KJM11 PCR product. In the third PCR reaction, the products of PCRs KJM10 and KJM11

were used as template, and for 5 cycles no exogenous oligos were added to prime the reaction, using only the complementary FLAG sequences to prime synthesis of the up- and downstream fragments. Subsequently, KJM10F and KJM11R were added to the reaction and, after 25 cycles of amplification, the resulting 2.2 kb PCR product contained the FLAG epitope tag, as well as ScaI and BglII restriction endonuclease recognition sites at opposing termini. Digestion of PCR product with ScaI/BglII and parent vector with ScaI/BglII/PstI (PstI cut at a unique restriction site within the liberated insert sequence, and used to reduce background from partial digestion products) and subsequent ligation using a 1:3 vector:insert ratio yielded pDTCF (plasmid Destination Tetracycline-inducible C-terminal FLAG Tag). In the case of inserting the tag at the opposing end of the expression cassette, the final product was named pDTNF (plasmid Destination Tetracycline-inducible N-terminal FLAG Tag). In both cases the resulting vector conferred chloramphenicol- and hygromycin B-resistance as well as encoded the CcdB toxin.

Creation of Entry Clones

We had at our disposal an entry clone library containing ~2600 *M. tuberculosis* ORFs in the backbone of pDONR221 (PFGRC/Colorado State University under NIAID contract HHSN266200400091c). In cases where the gene of interest was not a part of the library we sub-cloned the desired product by PCR amplification of the target DNA using purified H37Rv genomic DNA as the template. In accordance with Gateway entry clone-construction specifications from Invitrogen, the 5' termini of primers used for amplification were augmented in the following manner: The 5'/upstream primer was 5' - ggggACAAGTTTGTACAAAAAAGCAGGCTCTATG + gene specific sequence, while the 3'/downstream primer was 5' - ggggACCACTTTGTACAAGAAAGCTGAGTC + gene specific sequence. Site-specific recombination between the AttB-containing PCR product and the AttP-containing pDONR221 vector using the BP Clonase (Invitrogen) enzyme cocktail was carried out according to manufacturer recommendations. In brief, 100 ng of pDONR221 and 150 ng of gene-specific PCR product were

combined with 1 µL of BP Clonase II. Volume was adjusted to 5 µL by the addition of TE, pH 8.0, and the reactions were incubated at room temperature for 4 hours. To halt recombination, 1 µL (2 µg) of Proteinase K was added to the reaction and incubated at 37°C for 10 minutes. 1.25 µL of this product was transformed in to sub-cloning efficiency DH5α *E. coli* and plated on LB + 50 µg/mL Kanamycin. Isolated colonies were selected, expanded, and the plasmid was sequenced to confirm fidelity (data not shown). The resulting kanamycin-resistant, AttL-containing, vectors were termed pENTR-xxxx, where "xxxx" denotes the Rv gene number of the clone.

Creation of epitope-tagged expression vectors

In vitro recombination of the AttL-containing entry clones, and the AttR-containing destination vectors was carried out by site-specific recombination using LR Clonase (Invitrogen) as described by the manufacturer. Briefly, 100 ng of pDTCF or pDTNF and 80 ng of pENTR-xxxx were combined with 1 µL of LR Clonase II. Volume was adjusted to 5 µL by the addition of TE, pH 8.0, and the reactions were incubated at room temperature overnight. To halt recombination, 1 µL (2 µg) of Proteinase K was added to the reaction and incubated at 37°C for 20 minutes. 1.25 µL of this product was transformed in to sub-cloning efficiency DH5α *E. coli* and plated on LB + 200 µg/mL hygromycin B. Isolated colonies were selected, expanded, and the plasmid was sequenced to confirm fidelity (data not shown). The resulting product was named pEXCF-xxxx (plasmid Expression Tetracycline-inducible C-terminal FLAG Tag, where "xxxx" denotes the Rv gene number of the clone). In the case of epitope-tagging the N-terminus of the gene, the vector was named pEXNF-xxxx (plasmid Expression Tetracycline-inducible N-terminal FLAG Tag, where "xxxx" denotes the Rv gene number of the clone). These vectors are kanamycin- and chloramphenicol-sensitive, hygromycin B-resistant, episomally-replicating, and ATc-inducible.

Culturing Conditions:

M. tuberculosis strain H37Rv and *M. smegmatis* strain mc^2155 were cultured in Middlebrook 7H9 with the ADC supplement (Difco), 0.05% Tween80 at 37° C with constant agitation. For

transformation with ATc-inducible expression vectors and subsequent expansion/experimentation, cultures were grown with the addition of 50 μg/mL hygromycin B. All experiments were performed under aerobic conditions and growth was monitored by OD600. At an OD600 of 0.35, expression of a gene of interest was induced for the approximate duration of one cell doubling (18 hours and 4 hours for *M. tuberculosis* and *M. smegmatis*, respectively) using an ATc concentration 100ng/mL culture.

Chromatin immunoprecipitation:

DNA-protein interactions were characterized by cross-linking 50 mL of culture with 1% formaldehyde while agitating cultures at room temperature for 30 minutes. Cross-linking was quenched by the addition of glycine to a final concentration of 250 mM. Cells were pelleted, washed in 1x PBS + 1x protease inhibitor cocktail (Sigma), and resuspended in ChIP Buffer 1 (20 mM KHEPES – pH 7.9, 50 mM KCl, 0.5 mM DTT, and 10% glycerol) + 1x protease inhibitor cocktail. Due to the thick cell wall of *M. tuberculosis*, samples were mechanically lysed using Lysing Matrix B tubes and 3 rounds of bead beating at max speed for 30 seconds, with cooling on ice between treatments. Samples were centrifuged for 1 minute at 13.2 xg to pellet beads. Supernatants were collected and sample volumes were normalized to 500 μL in ChIP Buffer 1. We then utilized a Covaris S2 ultrasonicator at settings: amplitude = 20%, power = 5, cycles/burst = 200, for 16 minutes to shear chromatin to a uniform size centered around 200bp (**Supplementary Figure S2-1**). Following shearing, the sample was adjusted to buffer IPP150 (10 mM Tris-HCl – pH 8.0, 150 mM NaCl, and 0.1% NP40) and immunoprecipitation of FLAG-tagged proteins was initiated by incubating samples overnight rotating at 4°C with 10 μg M2 anti-FLAG antibody (Sigma). The following day, samples were incubated with protein G-coupled agarose beads (Pierce) rotating for 30 minutes at 4°C and 90 minutes at room temperature. Agarose bead-protein complexes were pelleted by centrifugation for 2 minutes at 2,000xg at which point the supernatant was discarded, and the samples were subjected to five rounds of washing in IPP150 buffer (rotate for 2 minutes, pellet bead-protein complex, discard supernatant). Increasing the stringency, the final two washes were carried out with TE,

pH 8.0. Protein complexes were eluted off the beads in two steps. In the first step, protein-bead complexes were incubated in elution buffer 1 (50 mM Tris-HCl – pH 8.0, 10 mM EDTA, and 1% SDS) for 15 minutes at 65°C. After pelleting and saving the supernatant, protein-bead complexes were treated with TE – pH 8.0 and 1% SDS for 5 minutes at 65°C. Elution supernatants were pooled and the proteins were digested/cross-links were reversed by incubation with 1 mg/mL Pronase for 2 hours at 42°C followed by 9 hours at 65°C. Immunoprecipitated DNA was subsequently column purified using QiaQuick PCR purification columns (Qiagen) and eluted twice with 20 µL 10 mM Tris-HCl, pH 8.5.

Illumina Sequencing Library Prep

All libraries were prepared according to standard Illumina protocols. Samples were sequenced on the Illumina GAIIx sequencer, generating 30-50 million 40bp reads.

Illumina Sequencing, Data Processing, Peak Calling, and Center-point Determination

A custom, freely-available, analysis pipeline was generated for this work (see chapter 3 of this dissertation and Peterson M.W., et al., manuscript-in-preparation). Briefly, reads were aligned to the appropriate reference genome using MAQ. A model was fit to the sequence coverage, and significantly enriched regions of 100 contiguous nucleotides or longer were included in downstream analysis. Because ChIP-seq enriched regions possess a characteristic bi-modal strand distribution [79], a cross-correlation filter requiring a stranded peakshift of >60 nucleotides was imposed. Centerpoints of enriched regions, the de facto transcription factor-bound sequences, were calculated via the implementation of the CSDeconv blind deconvolution algorithm ([80] Gomez et al. manuscript-in-preparation).

“Universal” Electrophoretic Mobility Shift Assays (EMSAs)

For “universal” electrophoretic mobility shift assays (EMSAs), three oligos (Integrated DNA Technologies) were resuspended to 50 µM in dsDNA annealing buffer (10mM Tris-HCl – pH 7.5, 100 mM NaCl, 1 mM EDTA). In this scheme, oligo 1 consisted of 30 nucleotides taken directly from a binding site

of interest in the *M. tuberculosis* genome. Oligo 2 consisted of 42 nucleotides: the reverse complement of the oligo 1 30-mer, as well as a 12 nucleotide "scaffold" sequence at the 3' end to which oligo 3 is the reverse complement. Oligo 3 consisted of a 12-mer with a Cy5.5 or IR800 fluorophore covalently coupled to the 5' end. The Cy5.5 12-mer scaffold/universal sequences were different than the IR800 12-mer. The three ssDNA oligos were combined to a final concentration of 50 µM, vigorously agitated, and heated to 95°C for 10 minutes on a benchtop heat block. The entire metal block was subsequently removed from heat and allowed to cool to room temperature over a period of ~3 hours protected from light. The resulting dsDNA product became the substrate for subsequent EMSA experiments.

Purified KstR protein (obtained from Los Alamos Systems Proteomics) was removed from storage buffer (10 mM Tris-HCl – pH 8.0, 150 mM NaCl, 1 mM DTT, 5% glycerol) and exchanged to reaction buffer (20 mM HEPES – pH 8.0, 75 mM NaCl, 10 mM $MgCl_2$, 5 ng/µL BSA, 1 ng/µL poly(dI:dC)) using a 10kDa-cutoff spin column (Amicon). 25 nM specific, Cy5.5-labeled, dsDNA target was used in all reactions. For specific and non-specific competition experiments, 20x molar excess IR800-labeled dsDNA was added to the reaction mixture (final concentration = 500 nM). All components of a reaction were combined, mixed, and incubated protected from light for 30 minutes at room temperature. 15 µL of reaction product was loaded on to 10% polyacrylamide TBE gel and run at a constant 150V for 75 minutes, protected from light. Owing to the lower melting temperature of the universal 12-mer used in these experiments (~65°C), the gel box was contained in an ice bath for the duration of electrophoresis. The gel was washed once in PBS prior to visualization on a Li-cor Odyssey scanner.

RNA isolation

RNA was isolated as described previously [54]. Briefly, cell pellets in Trizol were transferred to a tube containing Lysing Matrix B (QBiogene, Inc.), and vigorously shaken at max speed for 30 seconds in a FastPrep 120 homogenizer (Qbiogene) three times, with cooling on ice between steps. This mixture was centrifuged at max speed for 1 minute and the supernatant was transferred to a tube containing 300 µL

chloroform and Heavy Phase Lock Gel (Eppendorf North America, Inc.), inverted for two minutes, and centrifuged at max speed for five minutes. RNA in the aqueous phase was then precipitated with 300 µL isopropanol and 300 µL high salt solution (0.8M Na citrate, 1.2M NaCl). RNA was purified using an RNeasy kit following manufacturer's recommendations (Qiagen) with one on-column DNase treatment (Qiagen). Total RNA yield was quantified using a Nanodrop (Thermo Scientific).

Microarray analysis

RNA was converted to Cy dye-labeled cDNA probes as described previously [54]. For all microarrays described here, 3 µg of total RNA was used to generate probes. Sets of fluorescent probes were then hybridized to custom NimbleGen tiling arrays consisting of 135,000 probes spaced at ~100 bp intervals around the *M. tuberculosis* H37Rv genome (NCBI Geo Accession #: GPL14896). Arrays were scanned and spots were quantified using Genepix 4000B scanner with GenePix 6.0 software. These data were exported to NimbleScan for mask alignment and robust multichip average (RMA) normalization [81]. Subsequent statistical analysis and data visualization was carried out using Arraystar software. To compare against a standard, baseline, expression set, median expression values were calculated for all genes across all input microarrays (N=110). Altered gene expression was considered significant if it produced a moderated t-test $P < 0.01$ after Benjamini Hochberg multiple testing correction.

Results and Discussion

We sought to interrogate genome-wide protein-DNA interactions in systematic fashion using *M. tuberculosis* and *M. smegmatis* as the primary organisms of study. The original *M. tuberculosis* genome sequencing and annotation described ~180 putative transcription factors and DNA binding proteins [77]. As subsequent studies have shown that some sequences annotated as "conserved hypothetical" genes encoded for functional DNA-binding proteins (for example see [82, 83]) we performed manual curation of *M. tuberculosis* genes in order to create a more inclusive test set of putative DNA-binding proteins.

Through a combination of COG category and literature annotations we arrived at 215 putative DNA-binding proteins.

To interrogate DNA-protein association under uniform conditions, and independent of native regulation, we adapted an anhydrotetracycline (ATc)-inducible expression vector [78] containing a Gateway Recombination cassette (kind gift of Eric Rubin) with a FLAG epitope tag that would be translated at the N- or C-terminus of the expressed protein-of-interest. This Gateway destination vector was termed "plasmid destination N- or C-terminal FLAG tag" (pDTNF or pDTCF, **Figure 2-1A**). At the outset of the project we had available a Gateway entry clone library consisting of ~2600 *M. tuberculosis* ORFs cloned in to the vector pDONR221 (PFGRC, contracted by the NIAID). This library contained 145 of the predicted DNA binding proteins of *M. tuberculosis*. We subsequently sub-cloned 64 of the remaining 70 putative DNA binding-binding proteins encoded in the *M. tuberculosis* genome. Identity of all clones was confirmed by DNA sequencing (data not shown). Entry clones carrying the ORF-of-interest were subsequently recombined in to pDTCF or pDTNF to create "plasmid expression N- or C-terminal FLAG tag (pEXNF-xxxx or pEXCF-xxxx, where xxxx denotes H37Rv gene number), and subsequently transformed in to *M. tuberculosis* H37Rv. At the time of writing this collection consists of 208 strains, each merodiploid for a different ATc-inducible, FLAG-tagged, transcription factor. Strains of *M. tuberculosis* H37Rv carrying expression vectors are being made available from BEI Resources.

We performed proof-of-principle experiments with the transcriptional activator Rv3133c/DosR/DevR with an N-terminal FLAG tag (data not shown), as well as the transcriptional repressor Rv3574/KstR using a C-terminal FLAG tag (see below). To assess protein-DNA interactions and transcriptional impact of transcription factor induction in a uniform growth state all strains containing an expression vector were cultured under identical conditions. In trial experiments it was found that inducing gene expression with 100 ng ATc per mL of culture for 18 hours, the duration of approximately 1 generation for *M. tuberculosis*, yielded sufficient expression levels for chromatin immunoprecipitation.

ChIP and RNA samples were collected as described in detail in the methods section of this chapter. Following Illumina single-end read sequencing library preparation, samples were sequenced on the Illumina GAIIx. Sequencing data were processed using a custom pipeline written for this application (See chapter 3 of this dissertation, and Peterson et al., manuscript-in-preparation) and aligned to the appropriate source organism genome (e.g., *M. tuberculosis* H37Rv). A schematic overview of the entire protocol is depicted in **Figure 2-1B**.

Genome-wide binding of KstR in *M. tuberculosis* and *M. smegmatis*

To explore the applicability of this ChIP-seq work-flow across bacterial species, we chose to investigate the genome-wide binding characteristics of the tetR family transcription factor KstR. KstR is a highly conserved regulator in actinomycetes [84], and the genomes of both *M. tuberculosis* and *M. smegmatis* encode orthologs of this gene. In *M. smegmatis* a regulon for KstR was defined by gene knockout and measuring differential expression of putatively regulated genes using microarrays [85]. As the dis-regulated genes in the knockout were transcriptionally induced relative to wildtype *M. smegmatis*, the authors concluded $KstR_{Msmeg}$ was a transcriptional repressor controlling the expression of several genes in the cholesterol catabolism pathway. These authors also defined a KstR regulon in *M. tuberculosis* based on gene orthology. Finally, in the same work, a DNA consensus motif was defined for both $KstR_{Msmeg}$ and $KstR_{MTB}$ and binding of the protein was confirmed using electrophoretic mobility shift assay (EMSA). Thus, with a defined regulon in different bacterial species, and targeted DNA-binding data, but no comprehensive ChIP-chip or ChIP-seq data available, KstR made a suitable test case.

Using the method described in **Figure 2-1B**, we interrogated the genome-wide binding patterns for $KstR_{MTB}$ with ChIP-seq after 18 hours of ectopic transcription factor induction. Using single-end Illumina sequencing and a custom analysis pipeline (mansuscript in preparation), 28,387,294 reads were aligned to the ~4.4 megabase *M. tuberculosis* H37Rv genome (**Figure 2-2A**). As has been described elsewhere [79], we note the characteristic bi-modal read distribution corresponding to sequencing the

Watson and Crick strands of DNA from a center-point indicating the TF binding site (see **Figure 2-2A, inset**). In control experiments we created a background model in which wildtype samples were treated with ATc and processed for immunoprecipitation as described in the methods. Using this null binding model as a comparator we identify >200 significantly enriched regions bound by $KstR_{MTB}$. Importantly, we identify all previously identified KstR binding sites, and several of the most-enriched regions in this experiment correspond to the regulatory regions of those genes identified by Kendall et al. (indicated by stars in **Figure 2-2A**). By dividing the maximum peak height of an enriched region by the mode of genome-wide sequencing coverage (read depth of 83) we were able to calculate a total reads-corrected quality score for previously validated binding sites (**Table 2-1**).

In an effort to validate this method across bacterial species, we investigated the genome-wide binding characteristics of $KstR_{Msmeg}$ using a slightly modified ectopic expression system as was described for $KstR_{MTB}$. Specifically, the ribosomal binding sequence was modified for expression in *M. smegmatis* by altering the ribosomal binding site to more closely approximate the 5' UTR regions of highly-expressed *M. smegmatis* genes. In addition, to account for the more rapid doubling time of *M. smegmatis* compared to *M. tuberculosis*, the period of gene induction was correspondingly shortened to 4 hours. As demonstrated in **Figure 2-2B**, we find that the ChIP-seq protocol we describe was suitable for use in two different bacterial species. 40,503,760 single-end Illumina reads were mapped to the ~6.9 megabase *M. smegmatis* mc^2155 genome. Analysis of the bound regions indicates a substantial overlap with those genes originally described as members of the $KstR_{Msmeg}$ regulon – of 31 regions defined in the original work that describes $KstR_{Msmeg}$ regulation [85], 27 are bound in this analysis. Importantly, all bound regions that have been validated by EMSA were also bound in the present study (indicated by open circles in **Figure 2-2B**).

As indicated, ectopic induction of KstR from an episomal plasmid offers strong signal-to-noise ratios, though we observe considerably more binding than was anticipated based on the published data

for KstR [85]. This may be due to several factors. ChIP-seq as a method consistently reveals more widespread protein-DNA interactions than are detected by alternate means such as ChIP-qPCR or ChIP-chip [86, 87]. This is in part due to the agnostic nature of high-throughput sequencing as a detection method: one does not need *a priori* knowledge of bound regions against which to design primers, and thus the entire sequence universe of the organism under study is detectable. In addition, the increased dynamic range of high-throughput sequencing as compared to fluorescence-detection microarray strategies results in detection of signals that would previously have been technically indistinguishable from background [66]. Finally, while the use of an inducible/non-native promoter does allow for precise documentation of the impact of a known perturbation, it correspondingly increases the likelihood of detecting cellular interactions that would be too transient or low-affinity in the ordinary cellular milieu. However, as described below, there is no linear correlation between the number of regions bound by a transcription factor and the abundance of the expressed gene in the cell.

EMSA validation of binding

To test the reproducibility of $KstR_{MTB}$ binding events across the affinity spectrum we interrogated a subset of known and novel KstR-bound sites by implementing a fluorescent "universal EMSA" method (**Figure 2-3A**). This experimental strategy circumvents the use of radioactive or chemiluminescent substrates, and offers the possibility to simultaneously visualize target and competitor DNA species labeled with different fluorophores. In this approach a Cy5.5 or IR800 fluorophore is covalently coupled to a "universal 12-mer" single-stranded DNA sequence, and then annealed to a complementary scaffold sequence at the 3' end of a 30-mer ssDNA containing the binding target of interest. The final ssDNA is complementary to the target DNA of the scaffold strand. When annealed, the resulting 3-part, fluorescently-labeled, dsDNA can be used as a substrate in EMSA reactions and visualized on the Li-Cor Odyssey Scanner. For validation purposes, we chose to interrogate 3 KstR-DNA interactions. The first two targets are illustrative of intergenic regions that have

been previously identified to be bound by KstR. To interrogate this type of interaction we performed EMSA experiments employing KstR target sequence from the Rv3515c-Rv3516 and Rv3573c-Rv3574 intergenic regions. Rv3515c-Rv3516 was moderately bound by KstR and Rv3573c-Rv3574 is strongly bound by KstR, with enrichment scores of 188 and 535, respectively (**Table 2-1**). The final DNA target that we interrogated is illustrative of enriched intergenic regions newly-identified as being KstR-bound in this study. This class is represented by sequence from Rv2798c-Rv2799 with moderate binding enrichment and a score of 169. In each case the binding identified by ChIP-seq and confirmed by EMSA is specific in the face of molar excess non-specific competitor DNA, though binding can be abrogated by the addition of molar excess specific, alternately labeled, competitor DNA (**Figure 2-3B**).

The Scope and Numbers of Transcription Factor Families Analyzed To Date:

As demonstrated, the protocol for ChIP-seq described above was amenable for use across bacterial species, and novel bound regions can be validated by alternate methods. We also sought to determine if the method would work between transcription factor families. In this scenario one family of transcription factors within an organism may represent tractable ChIP-seq targets, while a second family might remain refractory to the epitope-tagging/induction strategy we have outlined. The utility of a ChIP-seq method then corresponds to its application across species *and* transcription factor families. To date we have performed ChIP-seq on 51 different transcription factors from 21 different transcription factor families in *M. tuberculosis* (**Table 2-2,** all data publicly available on TBDB.org). The bacterial transcriptional regulation paradigm is weighted towards repression as a primary means gene expression control [71, 88]. Not surprisingly, the majority of proteins interrogated in this set are predicted to be transcriptional repressors, with >35% of transcription factors assayed belonging to three families: TetR, ArsR, or GntR. This has significant implications for the successful interrogation of the DNA-binding patterns for a transcription factor. As dissociation from DNA targets, de-repression, is often mediated by allosteric interaction with small molecules [89, 90], proteins that are “primed” to

bind DNA in their native state should be amenable to binding analyses under a standard condition – mid-log phase growth in rich medium.

In addition to the canonical repressors described above, we observed reproducible DNA binding from canonical transcriptional activators as well. Included in this set are two noteworthy examples of transcriptional activators that require covalent modification for activity: Rv3133c/DosR/DevR, and Rv0757/PhoP. Both are response regulators of the two-component response regulator class, and both have been studied extensively (reviewed in [91] & [92]). In the case of DosR, we observe binding to every known target site in the genome, and closely recapitulate the known consensus motif to these targets (**Supplementary Figure S2-2A**) [40, 93]. There have been fewer studies of the DNA binding characteristics of PhoP on the genome-wide scale, though the consensus motif we identify suggests a head-to-head binding conformation (**Supplementary Figure S2-2B**), and is strongly reminiscent of published results [94, 95].

Transcriptional profiling

Protein-DNA interaction maps can be readily defined by methods such as ChIP-seq; however, the proximal functional implications of a DNA binding event are necessary in order to create dynamic gene regulatory networks. Because we describe a method in which all measurements are made following a specific and defined perturbation – e.g., chemical induction of a transcription factor – this protocol can be used to understand the transcriptional impact of TF induction. Immediately prior to cross-linking cultures for ChIP sample processing we dedicated a fraction of the sample for total RNA isolation. Following cDNA-synthesis and Cy dye coupling we hybridized samples from transcription factor inductions to a custom-designed NimbleGen tiling array platform (NCBI GEO Accession #: GPL14896) To date, we have analyzed 110 RNA samples, with up to 4 biological replicates per DNA-binding protein, from 46 of the 51 transcription factors for which we have ChIP-seq data. **Figure 2-4A** demonstrates a hierarchical clustering of all array results. In most cases the induction of a transcription

factor results in relatively modest or localized transcriptional changes – calculating the correlation of any induction vs. OD-matched wildtype yields R^2 values 0.822 – 0.983; however, even at a macroscopic level it is possible to associate certain inductions with more profound changes to the transcriptional architecture of the cell (**Figure 2-4B**).

Correlating TF overexpression with binding

With both the genome-wide binding patterns and transcriptional data from the same induction perturbation, analyses of transcription factor abundance vs. binding become tractable. As described above, a concern with over-expression of a protein is the appearance of "non-physiological" interactions that would not typically be observed under native or steady-state levels of abundance. In particular, with ChIP-seq as a read-out, it is possible to "drive" binding of a protein to disparate, non-specific, sites on the chromosome. In our hands, the most widespread DNA-binding protein analyzed to date associates with ~550 sites on the chromosome, whereas several bind fewer than 5 sites. This suggests that simply over-expressing a transcription factor does not force equivalent "blanket" binding effects. To address the possibility that different steady state levels (e.g. uninduced) of transcription factor were predictive of the number of sites bound by the protein, we determined the median expression value of each transcription factor based on absolute pixel intensity of probe from each microarray we have analyzed. In **Figure 2-5A** we note a slight negative correlation (Spearman's $\rho = -0.11$) between steady state transcript levels and number of sites bound. Similarly, when we analyze the transcript abundance of the ectopically induced gene, we observe that the absolute level of transcript tends to approach a maximum, and regardless of the starting abundance of the gene, there is ~2-fold difference between the least- and most-induced genes (**Figure 2-5B**). Finally, while the relative induction (induced transcript abundance – median transcript abundance) ranged from ~1.5-fold induction to >150-fold induction, we again do not see a significant correlation between induction level and number of sites bound by the protein (Spearman's $\rho = 0.17$, **Figure 2-5C**). Thus, while the physiological significance of low-level

binding events afforded by the increased sensitivity of ectopic gene induction and ChIP-seq remains an open question, the lack of correlation between absolute expression levels and number of binding events suggests that even in this non-native system the intrinsic DNA-binding proclivity of the protein plays a strong role in dictating what, and where, DNA sequences are bound.

Integrating binding and transcription, generation of regulatory networks:

Having described the approach for interrogating both the DNA-binding characteristics and transcriptional impact from the same transcription factor induction, this method allows for straightforward assembly of gene regulatory network maps on the single- or multiple-DNA-binding protein scale. In **Figure 2-6** we show the integration of DNA binding and transcriptional impact of that binding for the two-component response regulator, DosR. In this depiction we use DNA-binding as the seed criteria for inclusion in the network and plot the relative transcriptional impact of that binding event for all bound genes, relative to the median expression value of a gene-of-interest across all TF induction microarrays. Using ChIP-seq and microarray data from an ectopically induced transcription factor we capture all expected transcriptional changes attributable to proximal DosR binding in addition to ~35 silent binding events. An expanded discussion of data integration and regulatory implications are provided in chapter 3 of this dissertation.

Summary

In this work, comprising chapter 2 of this dissertation, we describe methods for the uniform interrogation of the DNA-binding properties across DNA binding protein families and two bacterial species. Using a chemically-inducible, epitope-tagged, expression system we demonstrated specific protein-DNA associations for TetR family transcription factors in both *M. tuberculosis* and *M. smegmatis*: Rv3574/$KstR_{MTB}$ and MSmeg6042/$KstR_{MSMEG}$. In addition, using a novel EMSA method, we tested and confirmed binding of $KstR_{MTB}$ to three different DNA targets – two known and one novel. Furthermore, because we induce expression of transcription factors from a uniform state, we are able to

capture both the DNA-binding patterns of each protein, but also the transcriptional impact associated with that perturbation using tiling DNA microarrays. The integration of these data allowed us to reconstruct a subnetwork seeded by the DNA-binding of the two component response regulator DosR, and elaborate on this network by painting on the transcriptional impact of that binding. These methods form the experimental basis of the data described in chapter 3 of this dissertation.

Acknowledgments

The raw data presented in this chapter are the result of a collaboration with the laboratory of Dr. James Galagan at Boston University. We wish to acknowledge particular members of this group including Matt Peterson, Sang Tae Park, and Chris Mawhinney.

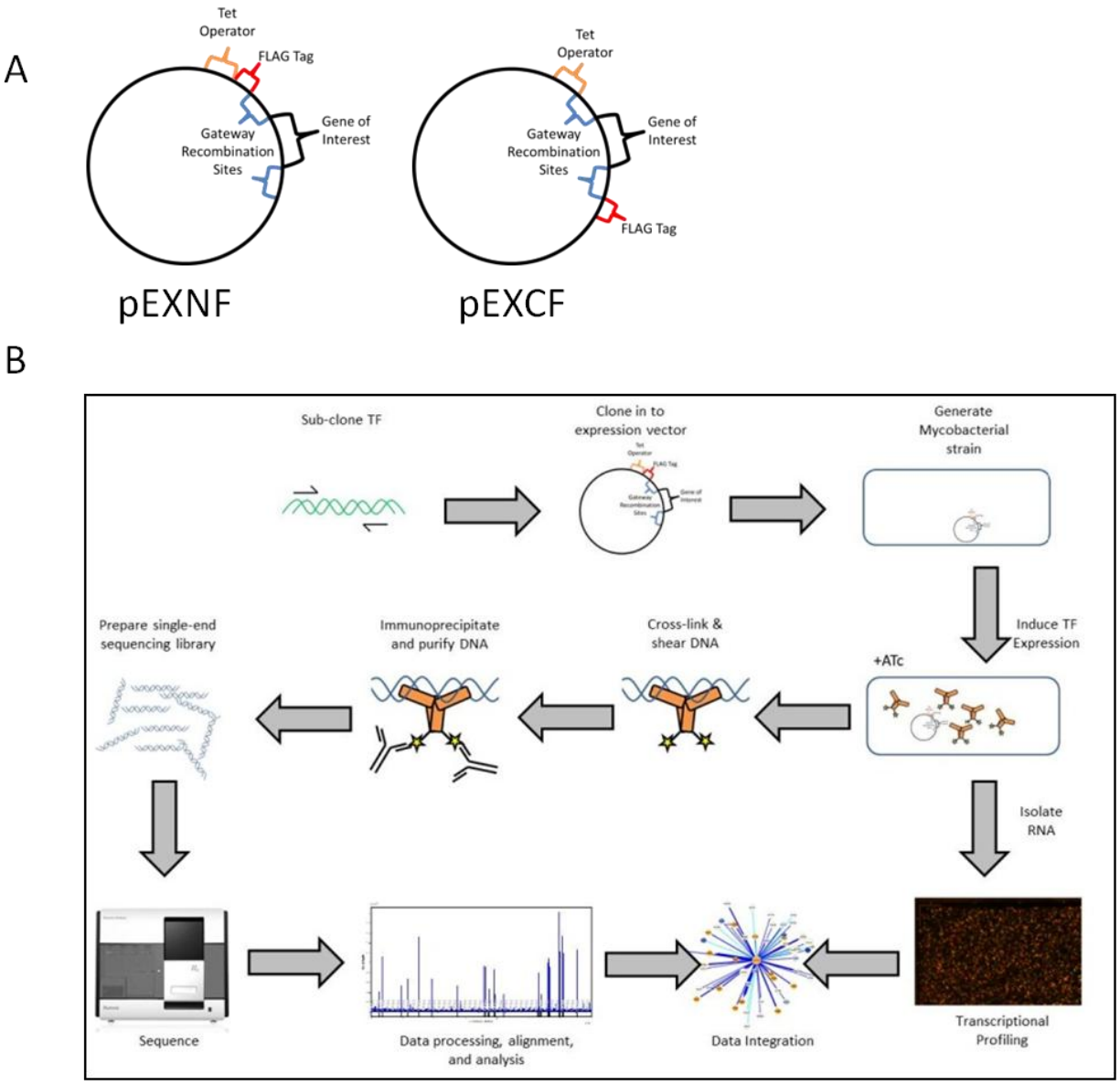


Figure 2-1: Overview of Reagents and Experimental Approach. **A**) Two anhydrotetracycline-inducible Gateway-compatible destination vectors, bearing a FLAG (DYKDDDDK) epitope tag at either the N- or C-terminus were generated for this work. As described in **B**) the work-flow for a standard experiment begins with sub-cloning the transcription factor-of-interest from the H37Rv genome and recombining in to the desired backbone to create an ATc-inducible expression vector. Upon transformation in to *M. tuberculosis* expression of the transcription factor is induced from the merodiploid copy of the gene. After approximately one cell generation (18 hours in *M. tuberculosis*) samples are split and processed for RNA and expression profiling, or for ChIP-seq. In the latter case, samples are cross-linked, quenched, and sheared using a Covaris S2 instrument. Following decontamination and removal from the BL3, samples are immunoprecipitated with monoclonal α-FLAG antibody. DNA is purified and a library is prepared for Illumina single-end sequencing. Reads are subsequently processed and aligned to the appropriate genome for enrichment analysis. Data integration and network analysis are described in more detail in chapter 3 of this dissertation.

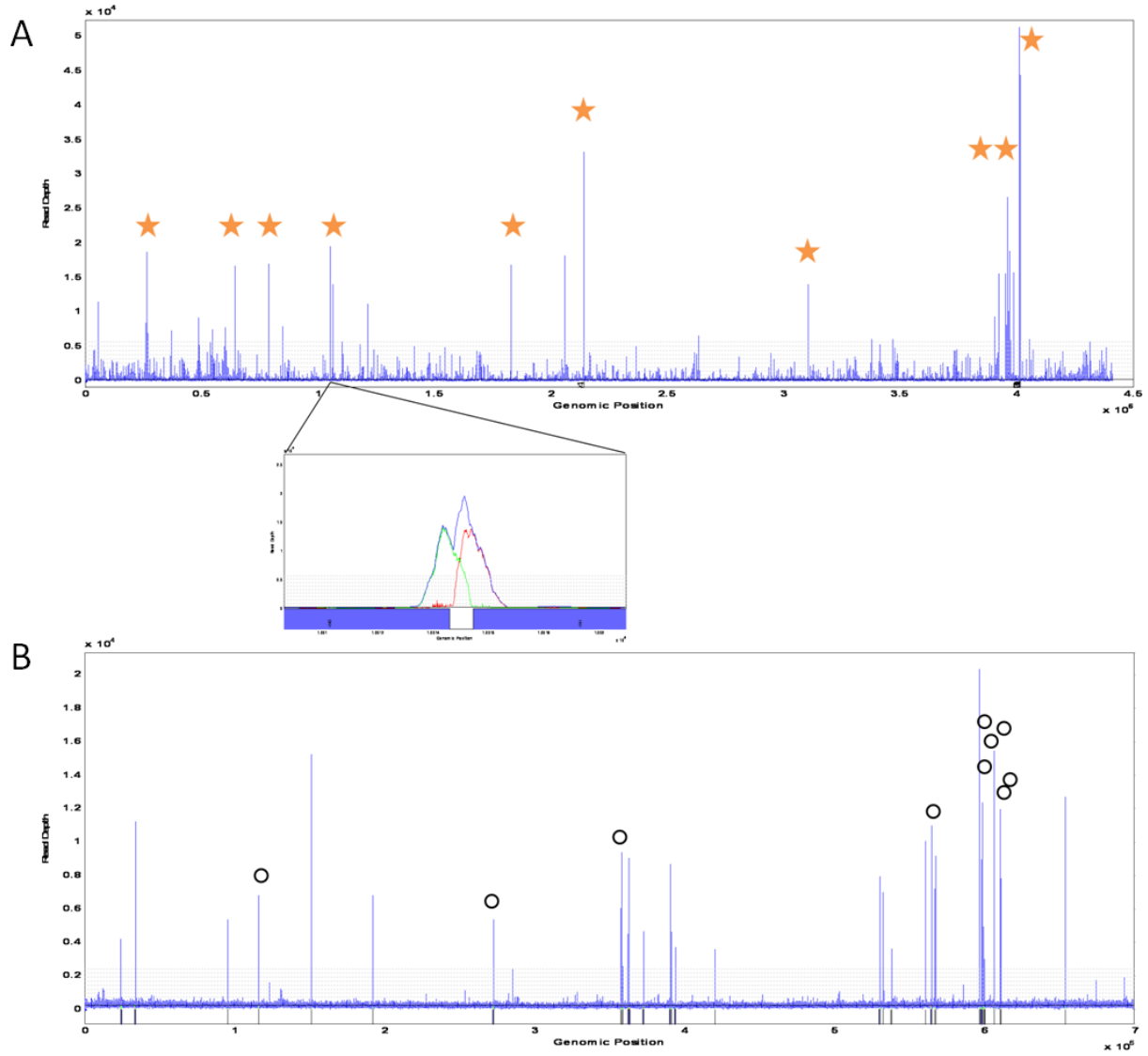


Figure 2-2: Genome-wide binding plots of $KstR_{MTB}$ and $KstRM_{SMEG}$. **A**) The genome-wide binding patterns for Rv3574/$KstR_{MTB}$. The x-axis is a linear representation of the 4.4 Mb *M. tuberculosis* genome. The Y-axis is read-depth from baseline. Horizontal grey lines indicate standard deviations of the mean coverage, plotting to the 10^{th} standard deviation. Blue bars represent KstR-bound/enriched regions, and the inset shows a representative peak with the characteristic ChIP-seq bimodal strand distribution. Red and green lines correspond to the read alignment profile for forward and reverse strands, respectively. Maximum peak height is a read-depth of 40,000. Enriched regions indicated with a star are those that were previously described as being regulated by KstR in [85] **B**) The genome-wide binding patterns for Msmeg6042/$KstR_{MSMEG}$. The x-axis is a linear representation of the 6.9 Mb *M. smegmatis* genome. As in panel A, the Y-axis represents read depth, and the horizontal grey lines represent multiples of the standard deviation of the mean. Maximum peak height is a read-depth of 20,000. EMSA-validated KstR-bound regions (as described in [85]) are indicated with open circles over the peak.

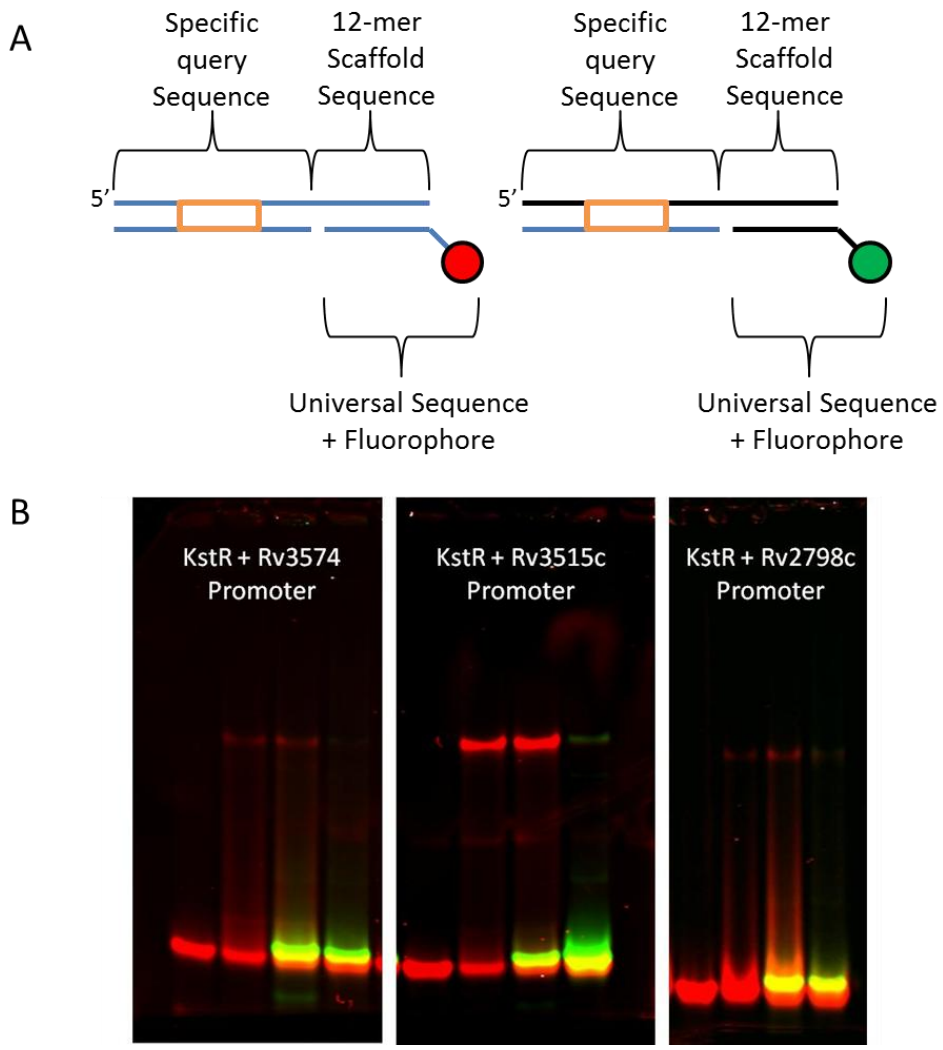
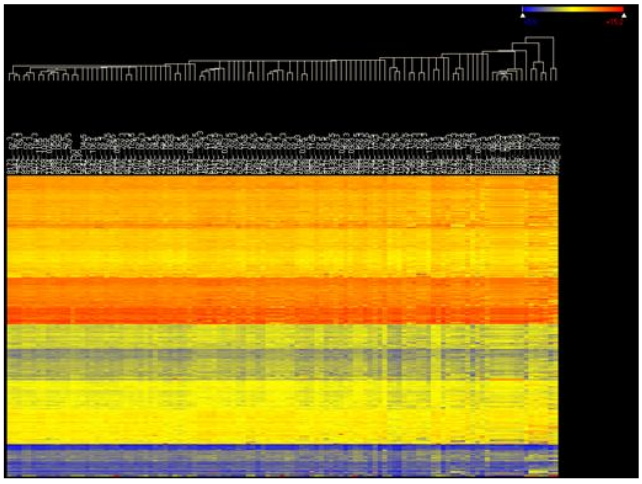


Figure 2-3: EMSA validation of select $KstR_{MTB}$ binding events. Enriched regions from the $KstR_{MTB}$ ChIP-seq results were interrogated for *in vitro* association with purified, recombinant, $KstR_{MTB}$. **A**) Schematic of universal fluorophore dsDNA targets. A single 42-mer contains the specific query sequence (30 nts), as well as universal scaffold sequence (12 nts). The scaffold sequence can be made unique to hybridize to a universal 12-mer sequence covalently attached to a Cy5.5 (left) or IR800 (right) fluorophore. The final ssDNA is a 30-mer containing the specific target and, if necessary, flanking nucleotides. All species are annealed by complementary basepairing and used in subsequent universal EMSA experiments. **B**) Three unique DNA targets, Rv3574 (left panel), Rv3515c (middle panel), and Rv2798c (right panel) kept fixed at 25nM were interrogated with 1.5, 1.5, and 4.5 µM KstR, respectively. In all panels, the leftmost lane contains Cy5.5-labeled DNA alone. The second lane contains Cy5.5-labeled DNA *and* KstR protein. The third lane contains the same Cy5.5-labeled specific DNA target, KstR protein, and 20x molar excess (500nM) IR800-labeled non-specific competitor DNA. The fourth, rightmost, lane contains the same Cy5.5-labeled specific DNA target, KstR protein, and 20x molar excess (500nM) IR800-labeled specific competitor DNA. In each panel, it is apparent that there is a protein-dependent shift of the DNA, that this interaction is specific in the face of molar excess non-specific competitor, but can be out-competed by molar excess specific competitor.

A



B

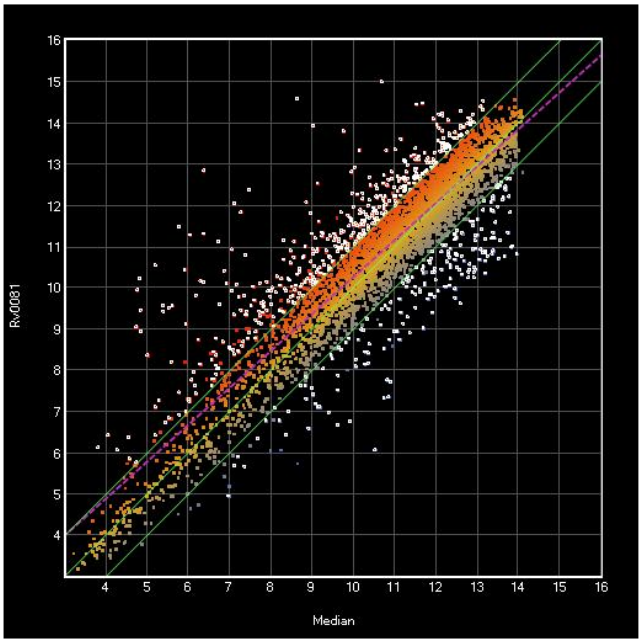


Figure 2-4: Transcriptional profiling of induced transcription factors. RNA samples taken at the time of ChIP-seq harvest were processed and hybridized to NimbleGen tiling microarrays where they were scanned and RMA-normalized prior to analysis. **A**) K-means clustering of all arrays germane to this dissertation chapter. On the y-axis are all genes of the H37Rv genome. The x-axis is the clustering configuration of all arrays. Macroscopically, ectopic transcription factor induction does not massively perturb the transcriptional landscape of the cell. **B**) Despite generally "local" perturbations to the transcriptional landscape of the cell, more profound changes are occasionally manifest via the induction of specific transcription factors. Shown is the transcriptional impact of Rv0081 induction. Shown on the x-axis are median pixel-intensity values for all genes (median value derived from all analyzed arrays). The y-axis is the pixel intensity of 4 biological replicates of Rv0081 induction collapsed in to a single (median) value. Each point represents the expression level of a gene. All points highlighted in white show a significant >2-fold change above or below median (moderated T-test $p<0.01$ after Benjamini-Hochberg FDR correction). 610 genes are observed to be differentially regulated.

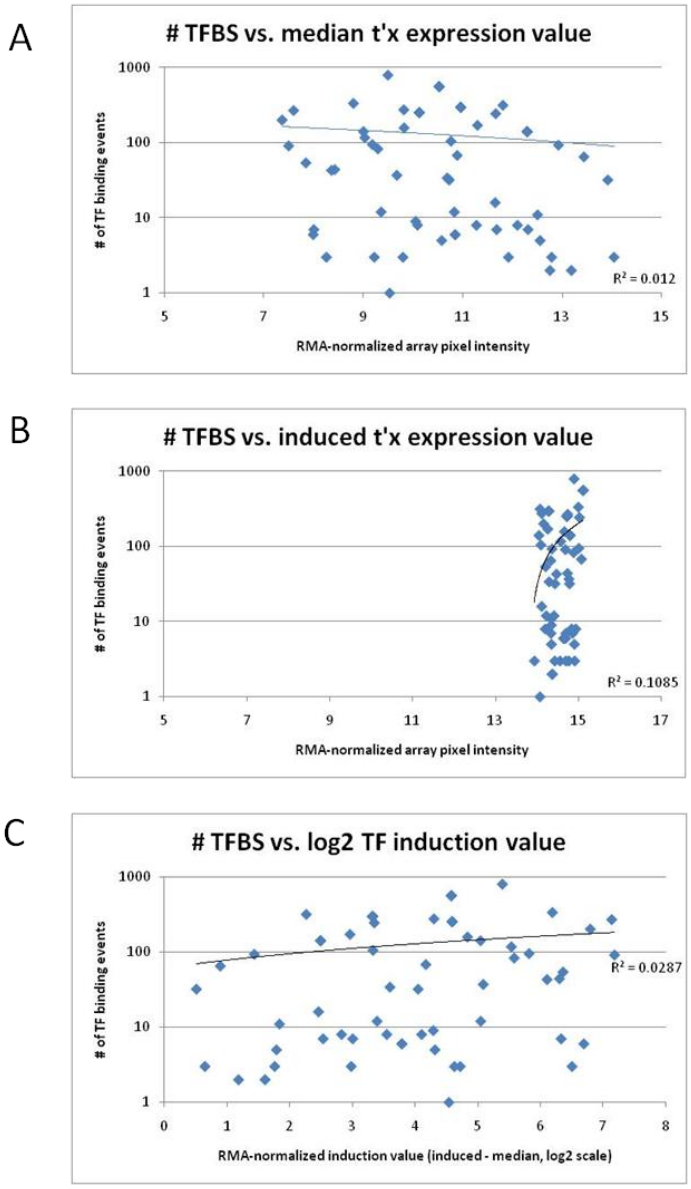


Figure 2-5: TF induction levels do not correlate with number of regions bound. We find no correlation between the number of regions a TF is observed to bind by ChIP-seq and its **A**) median/baseline expression level, **B**) its induced expression level, or **C**) the induction level upon addition of ATc (RMA-corrected absolute pixel intensity – RMA-corrected median pixel intensity).

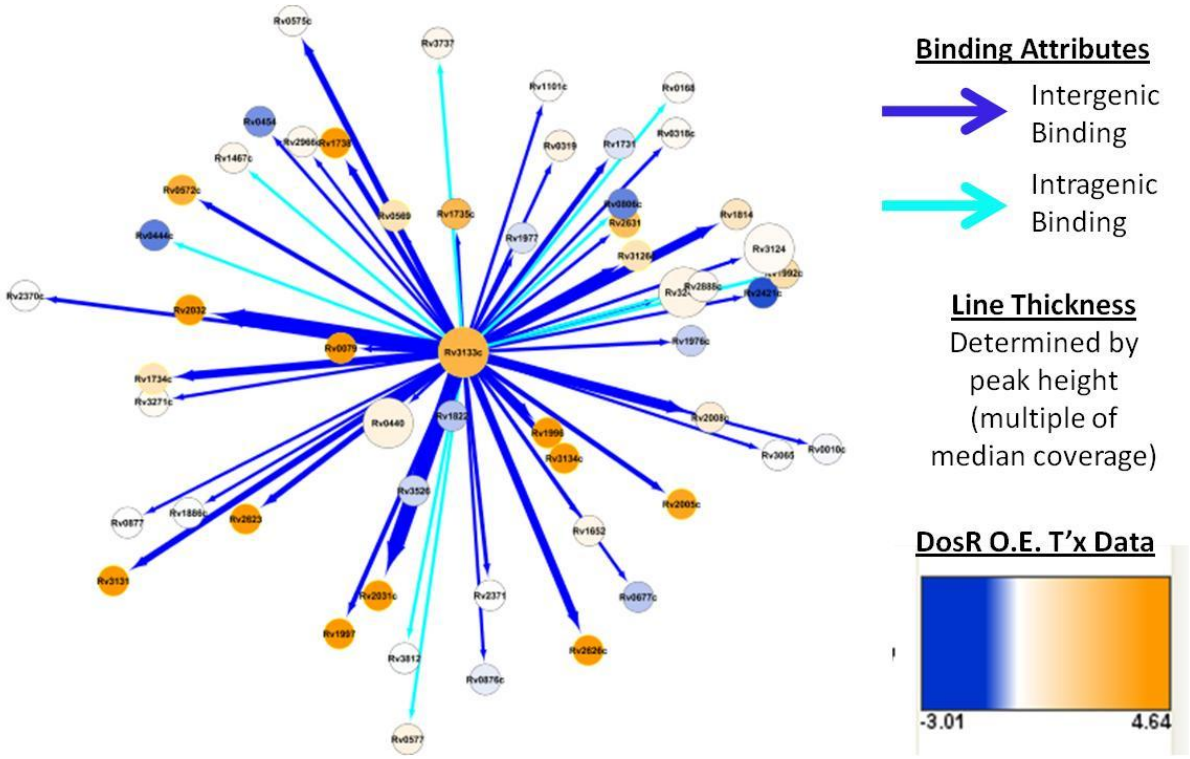


Figure 2-6: Union of binding and transcriptional data for Rv3133c/DosR. A binding height-weighted Cytoscape [96] plot of DosR binding and the transcriptional impact associated with that interaction. The network was seeded based on DosR-DNA binding. All edges are directional, radiating out from DosR to the target nodes. Dark blue edges are indicative of intergenic binding. Light blue edges are indicative of intragenic binding. Only the gene proximal to the binding site is shown (regulation through operon structure is not displayed). Edge thickness was determined by ChIP-seq peak height, calculated as a multiple of the median coverage, where larger peaks are depicted by thicker edges. Node color was determined by transcript level at the time of harvest (18 hours post-ATc addition). Log_2 fold-change was calculated by comparing Rv3133c-induced to median transcript levels.

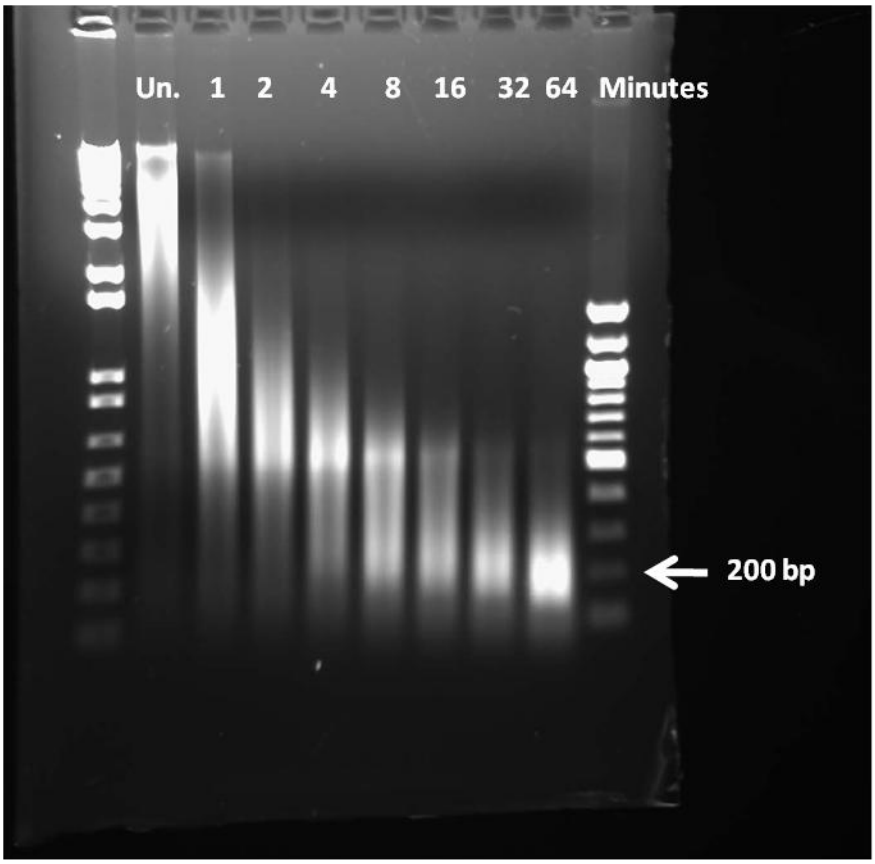


Figure S2-1: Agarose gel demonstrating chromatin shearing patterns with increasing processing time. Formaldehyde-crosslinked and glycine-quenched cell lysates were processed using the Covaris S2 ultrasonicator at settings: amplitude = 20%, power = 5, cycles/burst = 200 for the time duration indicated above each lane. To avoid the introduction of sample volume-based artifacts, each lane represents an individual treatment from the same unsheared stock sample (lane 2). 16 minutes of shearing provided DNA fragments of acceptable size without imposing restrictive time requirements.

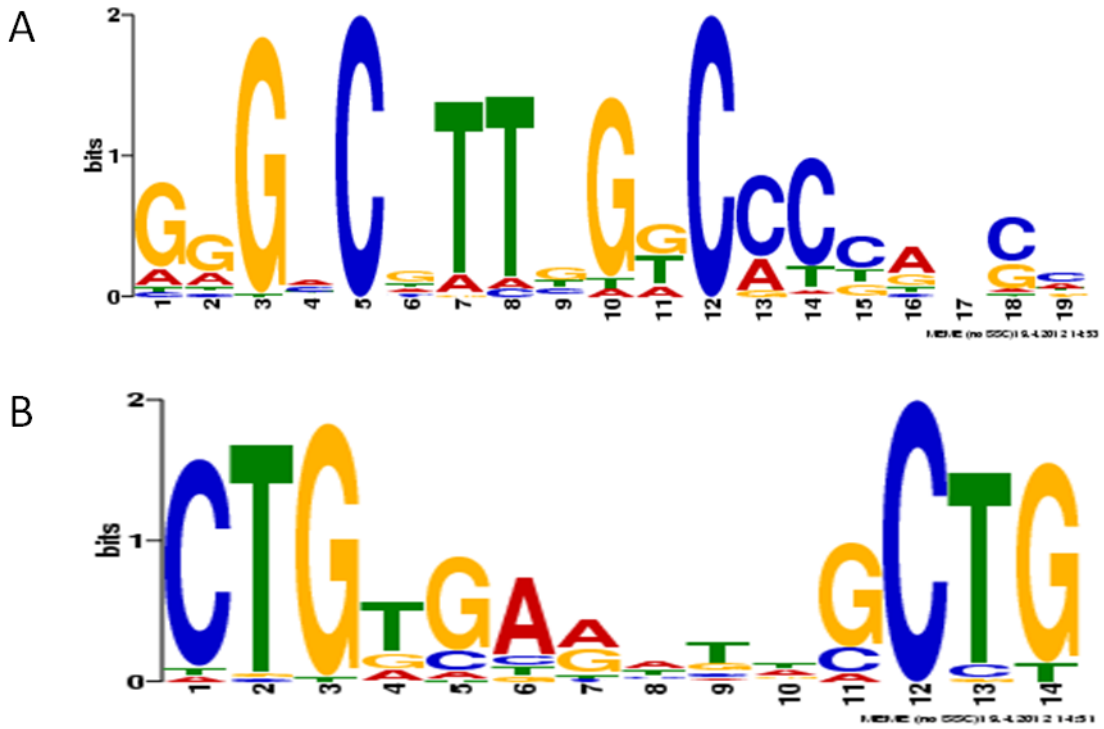


Figure S2-2: Consensus motifs identified for *M. tuberculosis* two-component response regulators DosR and PhoP. **A**) An unrestricted consensus motif search interrogating the 100 nucleotides flanking the centerpoint of binding for all enriched sites of DosR binding >1 standard deviation from mean coverage (read depth > 1613). **B**) An unrestricted consensus motif search interrogating the 100 nucleotides flanking the centerpoint of binding for the all enriched sites of PhoP binding >1 standard deviation from mean coverage (read depth > 1688). Consensus motifs were identified and generated using the MEME algorithm [97]. E-values for the motifs depicted are $3.2e^{-35}$ and $4.9e^{-20}$ for DosR and PhoP, respectively. Interrogating more regions of lower enrichment preserves major features of the above motifs, but dilutes low-information nucleotide content (see chapter 3 of this dissertation).

Gene	Score
Rv0551c	201.4
Rv0940c	236.3
Rv0953c	169.9
Rv1894c	401.7
Rv3503c	189.0
Rv3515c	188.3
Rv3520c	94.7
Rv3531c	227.9
Rv3545c	189.6
Rv3570c	620.0
Rv3573c	535.7

Table 2-1: ChIP-seq enrichment scores for previously-validated $KstR_{MTB}$-bound regions. Read depth of maximum peak heights for previously-validated $KstR_{MTB}$-bound regions were divided by the mode of genome-wide coverage (read depth = 83) to arrive at enrichment scores.

Family	Represented X times
2crr	4
AraC	1
ArgR	1
ArsR	5
AsnC	2
CarD	1
DeoR	1
FuR	1
GntR	4
HigA	1
IclR	2
LysR	1
MarR	2
MerR	1
MoaR	1
PadR	1
PhoU	1
TetR	10
WhiB	3
Xre	4
H-T-H	4

Table 2-2: Overview of *M. tuberculosis* transcription factor families interrogated in this study.

Chapter 3 - Reconstruction of the *Mycobacterium tuberculosis* Regulatory Network and Deconstruction of the Hypoxic Response

The following text is modified from a manuscript submitted to the journal **Nature**: Galagan, J.E., Minch, K.J., Peterson, M.W., et al, (2011) Reconstruction of the *Mycobacterium tuberculosis* Regulatory Network and Deconstruction of the Hypoxic Response. At the time of this dissertation submission, requests for revisions from reviewers have been addressed. This manuscript represents a significant collaboration with the Galagan lab at Boston University, who were instrumental in the computational aspects of this work. In addition to playing a significant role in writing and editing the manuscript, my primary contributions to this work reside in the conception, development, and execution of experiments with *M. tuberculosis*, as well as interpretation and validation of the described regulatory network. Text has been removed from the complete manuscript to more accurately reflect my efforts. To conform to the formatting of this dissertation and conserve the relative appearance of all figures within the submitted manuscript, all supplementary figures and tables have been re-named to appear as "Figure S3-x," where "x" denotes the order of the figure in the text.

Abstract

We have generated the first genome-scale model of the *M. tuberculosis* (MTB) regulatory network based on ChIP-Seq and combined this network with system-wide profiling of mRNA in MTB during hypoxia and re-aeration. Adaptations to hypoxia are thought to play a prominent role in MTB pathogenesis. We have developed a high-throughput system based on ChIP-Seq for comprehensively mapping regulatory binding, and integrated this with expression data from the induction of the same factors. Using this method we have reconstructed a regulatory network model based on 51 transcription factors (TFs). The network doubles the number of regulators whose interactions have

been studied in MTB, identifies thousands of new interactions and assigns functions to a substantial number and suggests many more functional interactions for even well-studied regulators. The network model also reveals an interconnection between the hypoxic response, lipid catabolism, lipid anabolism and the production of cell wall lipids. The regulatory network reveals TFs underlying these changes, and suggests Rv0081 plays a central role as a regulatory hub. The network further allows us to predict the pattern of expression of a significant number of genes during hypoxia and re-aeration as a function of the expression of their predicted regulators.

Introduction

Every minute, another three people in the world die of tuberculosis (TB). With more than 8 million new cases of active disease and nearly 1.5 million deaths annually, TB is a global health emergency of overwhelming proportions [98]. *Mycobacterium tuberculosis* (MTB) has been associated with human disease for thousands of years [99], and its success is fueled in part by the organism's ability to survive within the host for months to decades in an asymptomatic state. The mechanisms underlying this ability to persist in the host are poorly understood, though adaptations to hypoxia are thought to play a prominent role in MTB pathogenesis and latency [5, 6, 9, 29, 100-103]. Hypoxia is known to produce wide-spread changes in the bacterium, in part controlled by the regulator DosR [38, 40], and induces the pathogen to enter into a non-replicating state characterized by phenotypic drug tolerance. Within the host, MTB also shifts to lipids as a primary nutrient [42, 104]. In particular, it has been argued that cholesterol is a required lipid source during all stages of infection [105, 106], the degradation of which is controlled by the highly conserved regulator KstR [85, 107-109]. Lipid catabolism is, in turn, linked to the biosynthesis of a wide range of lipids that serve as energy stores [110, 111], factors associated with virulence and immunomodulation [112], and components of the unique and complex cell wall of MTB [113].

The regulatory mechanisms underlying these and other adaptations are largely unknown, as functions for only a small fraction of the 180+ MTB regulators are known, direct DNA binding data exists for only a handful of sites, and the interactions between regulators necessary for complex behavior have not been studied [114]. We also lack a comprehensive understanding of the cellular changes underlying pathogenesis, with existing studies typically focused on specific molecular components that can be difficult to integrate with independent studies of other components. To address these challenges, we have performed a systems analysis of the MTB regulatory networks, with an emphasis on conditions that contribute to MTB persistence in the host. Specifically we have performed chromatin immunoprecipitation followed by sequencing (ChIP-Seq) on 51 regulators, and integrated these data with expression data from the systematic over-expression of the same TFs to develop a predictive regulatory network model. We have also performed profiling of mRNAs in a systems-level and comparable fashion during a hypoxic time course and integrated these data in to a predictive network. We have demonstrated the ability of the regulatory network model to predict the pattern of gene expression during the hypoxic time-course and to predict key regulators of the hypoxic response of MTB. The integration of this regulatory network model with transcriptional profiling of responses during hypoxia, in turn, reveals new insight into the cellular adaptations that occur in a condition thought to mimic aspects of TB latency. We report an initial analysis, and we anticipate additional analyses based on these data. Toward this end, all raw data are publically available without restriction on TBDB.org.

Mapping and Functional Validation of MTB Regulatory Interactions

Systematic mapping of transcription factor-DNA binding

To systematically map transcription factor binding sites, we performed chromosome immunoprecipitation followed by sequencing (ChIP-Seq) [86, 115, 116] using FLAG-tagged transcription factors episomally expressed under control of a mycobacterial tetracycline-inducible promoter [78, 117-119](see chapter 2 of this dissertation), a method that has been validated for ChIP-Seq in other systems

[120, 121]. The use of an inducible promoter system allows us to study all the regulators of MTB in a standard and reproducible reference state without *a priori* knowledge of the conditions that normally induce their expression. Using a custom pipeline (**Figure S3-1** and **Table S3-1**) we identified binding locations in regions of enrichment with single nucleotide resolution (**Figure S3-2**). Using this method, we mapped 51 TFs, the largest single set of consistent ChIP-Seq data for any organism. We compared the results with previous reports for two well-studied regulators for which strong evidence for direct binding exists: the activator DosR (Rv3133c) [40, 80, 122-126] and the repressor KstR (Rv3574) [85, 107-109].

Our method displays high sensitivity and reproducibility. We identified all previously reported direct binding sites for KstR (27 sites) and DosR (30 sites) (**Figure 3-1A** and **Figure S3-3**, respectively), and recover the known motifs for these factors (Table S3-2). Biological replicates for 8 regulators shows that coverage for enriched sites is highly correlated between replicates ($R^2 > 0.83$ for all TFs –**Figure3-1B** and **Figure S3-4**). Furthermore, there is high reproducibility in binding location, with distances between replicate binding less than the length of predicted binding site motifs for the vast majority of sites (**Figure S3-1B**, **Figure S3-4**).

The number of binding sites detected is influenced by the concentration of the TF. ChIP enrichment is a function of the number of cells in which a site is bound [127, 128] which in turn is governed by the affinity of the site and the concentration of the factor. Thus, over-expressing a factor was predicted to increase the occupancy of strong sites up to a saturation limit while also occupying weaker affinity sites. This is confirmed by comparing ChIP-Seq experiments after inducing factors to different expression abundances (**Figure 3-1C**). In addition, although DosR requires phosphorylation by DosS/T during hypoxia for maximal binding [51], we recover all previously reported binding sites at higher levels of expression even in normoxia (**Figure S3-3**). Thus, in at least some circumstances, the use

of an inducible promoter system allows mapping of factors that normally require post-translational modification for strongest binding.

In addition, we identify more binding sites than were previously reported for KstR and DosR (**Figure 3-1A** and **Figure S3-3**). Most new sites have lower ChIP-Seq coverage than the majority of previously identified sites, although many display enrichment equivalent to some known peaks. Abundant binding of transcription factors, particularly to low affinity sites [87, 128], has been reported in *Saccharomyces cerevisiae* [129], *Caenorhabditis elegans* [130], *Drosophila melanogaster* [130-132], and mammalian cells [133, 134], but, to our knowledge, these data represent the first large-scale observation of this phenomenon in a prokaryote.

Our data suggest that non-sequence-specific binding is limited in our system. First, as noted above, binding is reproducible even for weak peaks. Second, a binding motif can be discerned for nearly all binding sites, and motif strength appears correlated with peak height in many instances (**Figure 3-1D**, **Figure S3-5**). Finally, we have confirmed direct binding to selected novel sites using EMSA (manuscript in preparation, described in chapter 2 of this dissertation). Thus, binding appears to be driven by sequence specific affinity and not by random association with DNA. Conversely, our data are not explained by binding of factors to all available sequence motifs. For all factors, less than 45% of the instances in the genome of predicted (or known) motifs are bound based on ChIP-Seq. Therefore only a fraction of possible sites are occupied, but reproducibly so in replicates. Binding thus displays genome context specificity, as previously noted, though the determinants of this specificity are not fully understood [128].

Associating binding with regulation

To assess the degree to which binding is associated with transcriptional regulation, we performed transcriptional analysis on RNA from the same cultures in which regulators were induced for ChIP-Seq (described in chapter 2 of this dissertation). Using these data we developed a procedure for

determining the possible regulatory roles of the binding sites identified by ChIP-Seq (**Figure 3-2A** and Methods). For each bound site, we examine all genes in a window of 1Kb around the site to determine if induced over-expression of the corresponding TF significantly alters expression of these genes. Peaks are considered validated if any gene in the window displays an expression level greater than a threshold value after correction for multiple testing (see Methods). This method identified a regulatory effect for 91% and 75% of previously-identified DosR and KstR sites, and associated regulation with 37% and 29% of new DosR and KstR binding sites revealed using ChIP-Seq. Many, but not all, new sites show weaker ChIP-Seq enrichment indicating evidence for regulatory effects of weak binding even for well-studied regulators [124, 126, 135]. Applying our method to all peaks from all analyzed TFs, we could assign a potential regulatory role to 25% of binding peaks at a window size of 1000 bp (**Figure 3-2A**). Moreover, stronger binding sites are more often associated with regulation than weaker sites, suggesting a possible correlation between binding strength and regulatory impact (**Figure S3-6**). Extending the window to 4000 bp, the distance between binding sites and associated target genes displays a pattern consistent with expectation: binding sites are typically located within 500-1000bp of the start codon of the gene they are predicted to regulate (**Figure 3-2B**), with the majority of peaks assigned regulation located in the upstream intergenic region. However, 76% of binding sites fall into annotated coding regions (**Table S3-2**) and a significant proportion can be assigned regulation (**Figure 3-2A**). Extensive genic binding has been previously reported, particularly in eukaryotic systems [116, 134, 136], and there remains no consensus on its functional significance. Prokaryotic binding sites have been largely mapped using lower resolution ChIP-Chip data that frequently display regions of enrichment broadly overlapping both genic and intergenic regions [137]. Our method detects binding sites at single nucleotide resolution, and suggests that, in at least some cases, genic binding may reflect the extension of promoter regions into upstream genes, alternative promoter regions within genes, or errors in the current annotation (**Figure 3-2B**).

We also tested the degree to which observed binding could be used to develop models predictive of gene expression (**Figure 3-2C**). We developed computational models relating the expression of target genes to the expression of TFs predicted to bind the target (Methods). The relationship between TFs and target genes were parameterized based on subsets of the overexpression data and tested on the remaining using 10-fold cross-validation. Only models showing a statistically significant fit were considered. Performance was assessed relative to random TF assignments with the same model structure. We could generate models that predict more accurately than random TFs for 47% of genes with binding (1411 genes). Although not a test of causality, these results confirm that binding improves predictive power for gene expression. More importantly, as described below, these models allowed us to begin to predict expression for genes in an independent data set.

As with previous reports [87], we cannot assign regulatory roles to all detected binding sites (**Figure S3-6** and **Figure S3-7**). However, our method may underestimate the proportion of binding with regulatory effects for several reasons. First, our validation is designed to identify strong regulation and is limited by the sensitivity of the microarray platform. It would thus be less likely to identify weak regulation which is likely present. Consistent with this, we do not validate all known regulation for DosR and KstR. Second, repression is more difficult to detect through over-expression of TFs due to diminished dynamic range for down-regulated genes using microarrays. This is consistent with the higher validation rate for the activator DosR (91%) relative to the repressor KstR (75%). Third, regulation might also act at a distance from the binding site, e.g. through DNA-looping or long range interactions between factors [138], which would not be detected by our method. Fourth, regulation might be obscured by combinatorial interactions or the need to recruit additional factors to the binding site that are not present in our perturbations. Finally, binding may serve other functional roles including the recruitment of other factors and modulation of chromatin structure [139], or alterations in the affinity of nearby sites as has been shown for DosR [135].

An MTB Regulatory Network Model

Using the combination of binding site mapping and functional validation, we analyzed the regulatory interactions of 51 TFs (26% of predicted MTB TFs). Our selection was weighted towards TFs that respond to hypoxia [40, 54, 125] or have a predicted role in lipid metabolism. By linking TFs with genes based on binding proximity and potential regulation, we constructed the regulatory network model shown in **Figure 3-3A**. A TF-target gene link was included irrespective of predicted regulatory role if the TF has a binding peak in the upstream intergenic region for the target gene or in the target gene itself. Links were also included for peaks in upstream genes, but only if within 500bp of the predicted start codon for the target gene and if a regulatory role was predicted (networks based on alternative criteria are shown in **Figure S3-8**). Consistent with known examples in MTB [140, 141] and *Corynebacterium glutamicum* [142], most factors are predicted to have both inducing and repressing roles. Although based on a subset of TFs, our network – constructed through a systematic approach – recovers biological relationships previously identified through reductionist methods, provides a scaffold for re-interpreting and integrating previously published results, and reveals new relationships that would not have been readily discovered with traditional methods.

The TB regulatory network model displays topological features seen in regulatory networks for other organisms [64, 67, 143]. Specifically, the number of genes with which individual TFs interact (out-degree) can be roughly fit to a power law distribution ($p(k) \sim k^{-2}$) where a few genes, or "hubs", interact with many genes, while most genes interact with fewer. Conversely, the number of TFs that interact with a given gene (in-degree) can be fit to an exponential ($p(k) \sim e^{-0.38}$) such that most genes interact with only one TF, while a few interact with more (**Figure 3-3B**). Moreover, 28 of the mapped regulators display putative autoregulation either through direct auto-binding (21 TFs) or by binding upstream of a polycistronic transcript encoding for multiple products including the induced TF (7 TFs). We also identify other motifs commonly found in biological networks, including over 900 putative feed-forward loops

(FFLs). FFLs are network motifs that mediate a range of functional dynamics including low-pass filters, sign-sensitive delays, and pulse generators [143-147].

Rv0081 and Rv3597c (Lsr2) are interacting hubs

Surprisingly Rv0081 forms the largest hub identified to date. Rv0081 is part of the initial hypoxic response [40, 125], but has been little studied. Rv0081 binds at 1120 sites, and its overexpression differentially regulates over 600 genes equally split between activation and repression. Rv0081 also displays a statistically significant overlap ($p<1e-12$) with sites from another hub, Lsr2 (Rv3597c), binds in the Lsr2 promoter, and is predicted to repress Lsr2. Lsr2 is an MTB analog of the H-NS nucleoid binding protein [148-150] (also a hub in *E. coli*) and binds and represses a wide range of targets with diverse physiological roles. Our data confirms widespread binding of Lsr2 (24% of all genes as compared to 21% from Chip-chip data [148]) associated with high AT regions. Lsr2 alters chromatin structure through DNA looping and thus likely modulates binding of other factors [151]. Noteworthy in this regard, Rv0081 and Lsr2 are the top two factors with significant overlap of binding sites with other TFs (4-fold more significant than any other TF, even after correcting for total number of binding sites).

A core subnetwork linking hypoxia and redox adaptation and lipid metabolism

The network also begins to reveal interactions between transcription factors mediating the complex and dynamic responses of MTB to its environment (**Figure S3-9**). Of particular interest is a subnetwork involving responses to altered oxygen status and lipid availability (**Figure 3-4**). These responses, among the most extensively studied in MTB, have been viewed largely as separate, disconnected phenomena. DosR (Rv3133c) and Rv0081 mediate the initial response to hypoxia, while a larger stimulon termed the enduring hypoxic response (EHR) is induced later in hypoxia [54]. KstR (Rv3574) controls a large regulon mediating cholesterol degradation as well as lipid and energy metabolism [85, 108]. KstR was also identified as part of the EHR[54], but the biology linking these responses was unclear, especially as only autoregulation of DosR or KstR has been reported.

We identified two potential regulators for KstR: Rv0081 and Rv0324. Both regulators interact with KstR through an FFL. Rv0081 is predicted to repress both Rv0324 and KstR, while Rv0324 is predicted to activate KstR. Repression of Rv0081 or activation of Rv0324 would therefore be predicted to activate KstR. Of note, Rv0081 is the only regulator in the initial hypoxic response apart from DosR, and our network identifies an interaction underlying the known induction of Rv0081 by DosR. Rv0324 is a regulator implicated in the EHR [54]. Thus, the network suggests a direct and complex connection between the regulation of hypoxia adaptation and lipid catabolism.

We also identify several potential regulators of DosR: Rv2034, Rv3066 and PhoP (Rv0757). All three are regulators of the EHR, providing possible feedback from the enduring to the initial hypoxic response. Rv2034, in particular, is predicted to activate DosR, thus forming a positive feedback loop. PhoP mediates a range of responses including up-regulating hypoxia adaptation genes and DosR [152-154], though direct regulation of DosR by PhoP had not been previously demonstrated. PhoP binding to DosR is the strongest among the TFs identified, providing a mechanistic basis for this previously-described genetic link. PhoP also mediates pH adaptation and our data confirm direct binding between PhoP and the aprABC locus required for this [155]. PhoP is also known to modulate the production of virulence lipids. In this regard it is noteworthy that we predict PhoP to bind upstream of and directly regulate WhiB3 (Rv3416), a redox sensitive protein that directly regulates the production of virulence lipids [156, 157]. In addition to PhoP, both Rv0081 and Rv0324 also display binding to WhiB3, with possible activation by Rv0081 predicted. These interactions thus elucidate potential regulatory links between lipid biosynthesis and hypoxia and redox sensing.

Taken together, the data reveal an interconnected subnetwork linking hypoxic adaptation, lipid and cholesterol degradation, and lipid biosynthesis. The links were either revealed by the network itself, or the links were derived by integrating existing literature with the framework of the network. The topology of the network predicts that altered oxygen tension is tied to changes in lipid metabolism and

intracellular lipid pools. Moreover, the network predicts that hypoxia directly modulates lipid catabolism. Connected to these changes is a hub centered on Rv0081 and TFs within the EHR. Although initially described in the context of hypoxia, these regulators appear to link to a wider range of stress inputs along with stress associated TFs, and may thus mediate multiple stress responses (**Figure 3-4**).

Comprehensive Profiling of MTB during Hypoxia and Re-Aeration

To broadly assess the changes associated with altered O_2 availability, and assess the explanatory power of the regulatory network in these responses, we performed genome scale profiling of MTB during a time course of hypoxia and subsequent re-aeration using a microarray platform (**Figure S3-10** and methods). To compare RNA profiling results between and within the time-points, all measurements were normalized to baseline RNA levels at T0 prior to the induction of hypoxia.

Identification of regulators underlying gene expression during oxygen changes

Changes in oxygen availability result in expression changes to nearly one-third of all MTB genes [40, 54, 125] (**Figure 3-5A**). Consistent with a non-replicating state, two-thirds of differentially expressed genes are repressed, although of roughly 100 differentially regulated transcription factors, two-thirds are upregulated. The majority of genes return to baseline during re-aeration. To identify temporal trends and associate them with possible regulators, we clustered the expression data into paths using DREM [158]. We then assessed the consistency between each path, the expression of each TF binding genes in the path, and the predicted regulatory role of the TF based on the overexpression transcriptomics described above.

Strikingly, we identify Rv0081 as a candidate high level regulator broadly predictive of the overall expression of sets of genes during hypoxia and re-aeration. In particular, the regulatory role of Rv0081 with respect to individual genes, as determined independently by overexpression, matches the correlation in expression between Rv0081 and these genes during hypoxia and re-aeration. Rv0081 is induced during hypoxia, declines in expression throughout hypoxia, and expressed at a low level during

re-aeration. A broad regulatory role of Rv0081 is thus supported by three independent sources of evidence: induction of Rv0081 alters expression of a large number of genes, ChIP-Seq reveals a large number of binding sites, and the expression and predicted regulatory role of Rv0081 correlates with the expression of bound genes in an independent expression data set derived from a significantly different condition.

Predicting gene expression during hypoxia

We next sought to assess the degree to which the regulatory network could be used to predict changes in the expression of individual genes during hypoxia and re-aeration. We used the regression models described above - parameterized by the independent ChIP-Seq and TF overexpression transcriptomic data - to predict the pattern of target genes based on the expression of their predicted regulators (Methods). Only target genes displaying significant changes were tested, and predictions were compared to models based on random sets of TFs. Using the models, we generate predictions that are significantly better than random for 838 genes (52% of modeled genes, 33% of all genes with significant changes). Examples are shown in **Figure 3-5C**, and predictions for additional genes discussed in this manuscript are in (**Figure S3-11**). For example, we correctly predict the pattern of expression of KstR, confirming the implication of the network topology. Importantly, these data also indicate that the regulatory network, built from a baseline condition, can generalize to other conditions with predictive power.

Concluding Remarks

This report presents the first stage in the reconstruction of the MTB regulatory network and its integration with system-wide profiling of MTB during a time-course of hypoxia and re-aeration. The regulatory network represents an initial model based on 26% of MTB regulators. Although this initial model is necessarily incomplete, it confirms previously known physical interactions, provides possible mechanisms for known regulatory interactions, provides a broader framework for re-interpreting

existing data, and identifies network structures that have been shown to underlie complex dynamic behavior. The predictive models based on the network data take a first step towards systems modeling. And integration of the network model with profiling data provides new insight in the physiological consequences of the regulatory programs induced by changes in oxygen availability – an environmental perturbation relevant to adaptation of the microbe to oxygen-limited host microenvironments. These methods and results provide a foundation for ongoing efforts to map the complete regulatory network.

The results presented here identify compelling questions for further investigation. As with previous reports, the functions of the majority of the reproducible TF binding sites are not known. Our results suggest functions for some, and experiments are ongoing to study the potential functions of others. Moreover, the resolution of our binding data provides opportunities to study regulation in greater depth for even well-known regulators. Similarly, our profiling data suggest interpretations for changes in gene expression that require future targeted experimentation. Nonetheless, the system-wide nature of our data reveals a context for the data that enriches their interpretation and provides a more coherent map for guiding such targeted experiments. In addition, our data provide a consistent and comparable data set that can be used in the development of systems level modeling algorithms. Finally, it remains to be shown how the network connections and physiological alterations identified *in vitro* will be related to changes *in vivo*. Although previous literature suggests the importance of many of the processes described above, ongoing work is aimed at a systems-level profiling MTB and the host during the process of infection.

Methods

Strains

MTB H37Rv was used for all experiments. The specific strain was acquired from the Colorado State University TB Vaccine Testing and Research Materials Contract. This strain has been fully

sequence by the Broad Institute and the data are available at http://www.broadinstitute.org/science/projects/gscid/projects. As described below (and in detail in chapter 2 of this dissertation), a library consisting of the majority of MTB regulators under the control of an inducible promoter was generated (one clone per TF) and will be made publically available through BEI (www.beiresources.org/).

Culture Protocol for Hypoxia Time Course

MTB H37Rv was grown to sufficient biomass for multiple high-throughput analyses over multiple time points in standard Middlebrook 7H9 (Difco) supplemented with glycerol (0.2%), Tween80 (0.05%), and ADC supplement (Difco). Cells were pelleted, re-suspended in Sauton's media without Tween80 to an OD_{600} of 0.2, and cultured for two days in aerobic rolling culture. The detergent free media results in a dispersed culture of microclumps that make measures of OD and cfu less meaningful. At time point zero (T0) the culture was diluted in half to a calculated final OD of 0.1 to 0.2 and transferred to three-armed flasks for hypoxic culturing, as described previously [54]. Samples were taken after 1, 2, 3, 5, and 7 days of culturing in bacteriostatic hypoxic conditions, returned to aerobic rolling culture, and sampled after 1, 2, 5, and 7 days of re-aeration. Some experiments focused on a subset of time points. Three separate time course experiments were conducted, each with at least three biological replicates from each time point. Microarray analysis of the mRNA transcriptome was done with all experiments. In all experiments the very sensitive hypoxic responsive regulon controlled by DosR was induced at T0 due to stresses induced during transfer to the hypoxic culture system. Microclumps formed in the absence of detergent did not induce the DosR regulon, as can be seen at the later re-aeration time-points.

As described previously [159], there is no significant drop in colony forming units over this time frame. We verified the viability in our detergent free model by plating after detergent treatment to disrupt the microclumps (data not shown).

Chromatin Immunoprecipitation Followed by High Throughput Sequencing (ChIP-Seq)

Detailed methods are described in chapter 2 of this dissertation. Briefly, chromosome immunoprecipitation followed by sequencing (ChIP-Seq) [86, 115, 116] was performed using FLAG-tagged transcription factors episomally expressed under the control of the mycobacterial tetracycline-inducible promoter [78, 117-119]. Each of the 51 regulators studied were induced with 100.0 ng/mL anhydrotetracycline (ATc) during mid-log-phase growth, and ChIP was performed using a protocol optimized for mycobacteria. In the case of the KstR induction titration curve, expression was induced with 10.0, 1.0 and 0.1 ng/mL ATc. ChIP preparations were sequenced using Illumina, and reads were aligned to the MTB genome [77] and analyzed using a custom pipeline (Peterson, M.W. et al manuscript-in-preparation and **Figure S3-1**). Using a blind deconvolution approach [80], we then identified binding locations within regions of enrichment with single nucleotide resolution.

Sequencing

All sequencing was performed on an Illumina GAIIx sequencer at the Boston University Illumina Core Facility (http://www.bu.edu/iscf/). Single 40bp reads were generated. A single lane was used for each ChiP-Seq sample resulting in roughly 30-50 million reads per sample. All library preparation and sequencing was performed using standard Illumina protocols.

Sequence Analysis

Sequence reads were mapped to version two of the *M. tuberculosis* genome using MAQ[160]. The *pileup* command from the SAMTools [161] suite was used to calculate coverage along the forward and reverse strand along the genome. From this coverage, regions of enrichment along the genome were identified using a lognormal distribution. The lognormal distribution is defined by the probability density function (PDF)

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma}}$$

Here, μ and σ are the mean and standard deviation of the natural logarithm of the random variable, respectively. For each experiment, positions along the genome greater than 5 times the mean coverage or higher were excluded to avoid fitting outliers, and the parameters of the distribution were estimated using maximum likelihood, calculated as

$$\hat{\mu} = \frac{\sum_k \ln(x_k)}{n}$$

$$\hat{\sigma} = \frac{\sum_k (\ln(x_k - \hat{\mu})}{n}$$

The resulting distribution was then used to score each position of the genome to identify enriched positions along the genome. Coverage at each position was scored against the distribution, and positions with a p value of 0.01 or lower were called as enriched. Because TF binding is expected to result in contiguous regions of enrichment, only regions of enrichment of 100 nucleotides or longer were included in further analysis.

A cross-correlation filter was then applied to the resulting regions to identify those that have the expected signature of transcription binding in ChIP-seq experiments, identified by a shift between peaks in the forward and reverse strands. The cross-correlation function is calculated as

$$c[n] = \sum_{m=-L}^{L} f[m]r[n+m]$$

In this function, *f* represents the coverage on the forward strand, while *r* is the coverage on the reverse strand, and *n* is the amount of shift applied to the function. The shift between peaks in the forward and reverse was defined as the value *n* that maximized the cross-correlation. Regions with a shift of less than 60 nucleotides were removed from further analysis.

The resulting regions were then compared to regions identified using the procedures described above for two wild-type background lanes, as well as a ChIP experiment for the Lsr2 protein. Enriched regions that were found to overlap either a region called in one of the wild-type experiments or the Lsr2

experiment by at least 10 percent of the region length of 30 nucleotides (whichever is larger) and had a height of less than 5-fold (for the background) or 2-fold (for Lsr2) over median compared to the overlapping region were removed from analysis.

Finally, we used a modified version of the CSDeconv blind deconvolution algorithm [162] to identify binding sites within the enriched regions. Deconvolution was used to fit a model for a binding event from the 16 most enriched regions, and this model was used to identify peaks in the remaining regions. Next, a motif search was performed around these peaks using MEME [163]. The enriched regions were searched for this motif using FIMO [164], and the centers of the motif instances were used to seed a second round of blind deconvolution, with the position of binding constrained to the centers of the motif instances. Predicted binding events with a height greater than zero (and thus having evidence in both sequence and ChIP evidence) were selected as the final set of binding events.

Binding events were then assigned to genes and classified by both where they were located relative to genomic features (either *intergenic* or *genic)* and their location relative to potentially regulated genes (either *upstream*, *downstream*, or *in-gene*). For binding sites located within genic regions, the region bound, as well as the two flanking genes were included as potentially regulated, while the genes surrounding the intergenic regions were included for intergenic binding events.

Transcriptomics of TF Induction

Detailed methods are described in chapter 2 of this dissertation. Briefly, cells were cultured in Middlebrook 7H9 with the ADC supplement (Difco), 0.05% Tween80, and 50 µg/mL hygromycin B at 37° C with constant agitation. All experiments were performed under aerobic conditions and growth was monitored by OD600. Total RNA was isolated from TF-induced cultures 18 hours after treatment with 100 ng Anhydrotetracycline (ATc) per mL of culture or an equivalent volume of DMSO (in the case of uninduced controls) using the protocol described in chapter 2 of this dissertation. When interrogating the same culture for ChIP-Seq and RNA profiling, cells were divided immediately prior to sample

processing. In the case of biological replicate RNA samples, independent cultures were generated and transcription factor induction/RNA isolation was carried out as described above.

Assigning Regulation to Binding Peaks

TF overexpression data from 61 experiments was used to assign a probability of observing the expression level for each gene identified to be bound by a given transcription factor in the overexpression microarrays. For each gene, we built an empirical distribution from the transcriptomics, eliminating any experiments for which the transcription factor being overexpressed has been implicated in the regulation of the gene of interest. This distribution then was then used to score the expression values for the gene in the experiments for which the TFs bound were overexpressed. The p-value of a gene being regulated by a transcription factor through a binding site is the number of cases when the gene has higher expression value – compared to the value in the experiment where the transcription factor was induced – over total number of values in the background distribution. Every binding site can have multiple targets depending on the mapping distance. The final p-value for the binding site is the smallest one among all its targets. Bonferroni correction was applied to each case where a peak was associated with multiple genes. This procedure was performed for the entire set of peaks as a whole, as well as each of the different classes of regulation. To determine the expected percentage of binding events that have expression evidence purely by chance, random sets of peaks with the same size and composition are randomly selected from the genome excluding already known binding sites and 4000 nucleotides around them.

Gene Expression Modeling

Model Structure and Parameterization

We developed regression models relating the steady-state expression of individual genes to the expression of predicted regulators. For each target gene, a selection process was used to identify the

optimal subset of predicted regulators to be used as regressors. Then, for each set of regressors, we considered 8 possible model structures. To select the best combination of regressors and model structure, the accuracy of each combination for predicting the expression of the target gene in the TF induction data set was determined. The accuracy in predicting gene expressions is then assessed with cross-validation on the TF induction dataset as well as generalization to hypoxic time course data described below.

The overall model selection process was as follows. For each target gene:

1. The TFs predicted to potentially regulate the target gene were selected as described in the main text.
2. The associated TFs were sorted based on z-scores derived from the TF induction experiments. Z-scores reflect the degree to which induction of a TF induces a large expression change in the target gene.
3. For each target gene, the set of regressors was initialized to the one TF with the highest z-score. If there were any other TFs binding to target gene for which induction experiment transcriptomics data were not available, they would also be added to the initial regressor set.
4. For the current set of TF regressors, each of the 8 potential model structures – described below – was considered. For each possible model structure:
    a. The model was parameterized by fitting to TF expressions from all experiments in the TF induction dataset. An F-test evaluates the hypothesis that the proposed regression model fits the data well. A model structure is considered only if the p-value of its F-test is less than threshold of 0.005.
    b. Model selection was guided using AIC [165] and Lilliefors [166] test. AIC is a measure of goodness of fit of a statistical model that corrects for the number of parameters, allowing comparison between different model structures. Lilliefors tests the hypothesis

that the error remaining from model fitting comes from a normal distribution. From the models that passed this test, the structure with the minimum AIC is selected as optimal. If no model passed the normality test, the model with minimum AIC is chosen.

5. The set of regressors was then updated by adding the TF with the next highest z-score (from step 2), and step 4 repeated. If model accuracy improved, the updated set of regressors was chosen, and step 5 was repeated.
6. The model at step 4 was selected as the final optimal model if adding an additional regulator in step 5 did not improve prediction accuracy.

Model Structures

Since the exact relationship between target genes and TFs may vary [167], we considered 8 possible model structures for each target gene. These structures model the expression of a target gene (y) with linear regressions on TF expressions x_i for i=1 to T (where T is the number of regulating TFs), with and without interaction terms, sigma factors or polymerase genes The most general model structure is:

$$y = a + \sum_{i=1}^{T} b_i x_i + \sum_{i=1}^{T} \sum_{j=i+1}^{T} c_{ij} x_i x_j + d x_{sigA} + e x_{rpoA} + \varepsilon$$

where x_{sigA} is the expression of sigma factor sigA (Rv2703) and x_{rpoA} is the expression of RNA polymerase alpha chain rpoA (Rv3457c) and ε is the noise or error with normal distribution, zero mean and variable variance.

The expressions for all N number of experiments of a target gene ($\underline{y}_{N\times 1}$) can be written as:

$$\underline{y} = f(X) = \underline{a} + X\underline{b} + XCX^T + \underline{d}X_{sigA} + \underline{e}X_{rpoA} + \underline{\varepsilon}$$

Where $X_{N\times T} = [x_1, x_2, \dots, x_T]$ is the matrix of expressions of regulating TFs and X_{sigA}, X_{rpoA} are $N \times 1$ columns of expressions corresponding to sigA and rpoA. The $N \times 1$ vectors $\underline{a}, \underline{b}, \underline{d}, \underline{e}$ are linear regression

coefficients/parameters and C is a $T \times T$ triangular matrix of interaction coefficients with zero diagonal elements. Collinear columns of X are removed in regression.

All 8 model structures are common in the first two terms (zero term and linear TF terms). The addition of the next three terms (TF interactions, sigA, rpoA) generates $2^3 = 8$ possible structures for form an ensemble of models $y = \{f_1(X), f_2(X), \dots f_8(X)\}$ from which the optimal model is selected for each target gene.

It should be noted that other nonlinear model structures such as second or third order models and logistic regressions were also tested. However, the ensemble of models were limited to linear models only, because higher order models did not have significantly better performance for most genes while they would present more parameters with more complex models and could overfit data in some cases. Also, logistic functions could not capture induced expressions which were at the tails of expression distribution, whereas these data points were most reflective of TF regulation. Thus, linear models could create a balance between the two.

Also, sigA and rpoA were added as linear terms to the model, rather than normalizing by e.g. sigA (as a gene expected to have non-varying expression), to avoid generating biased models.

By parameterizing the optimal model, expression is predicted as $\underline{\hat{y}} = \hat{f}(X)$ and the error is calculated between the predicted and actual expression as $L\left(\underline{y}, \hat{f}(X)\right) = \sum_{j=1}^{N}\left(\underline{y}_j - \hat{f}\left(X_{row=j}\right)\right)^2$, which is sum of squares errors or prediction (SSE). AIC, as a measure of goodness of fit, is a function of maximized log likelihood of $\underline{y}$ and the number of model parameters k. Under the assumption of i.i.d. normally distributed errors, AIC can be calculated as:

$$AIC = Nln\left(L\left(\underline{y}, \hat{f}(X)\right)/N\right) + 2k$$

where k is the number of parameters and N is the total number of experiments.

AIC penalizes model uncertainty in the first term as well as number of parameters (k) in each model, to avoid over-fitting in assessment of models. Thus, the optimal model would be the one with minimum AIC.

Cross-Validation and Accuracy Assessment on TF Induction Data

The ability of predicting gene expressions was first evaluated by parameterizing models on a subset of the TF induction expression data set, and assessing the accuracy on the remaining subset through a 10-fold cross-validation as below. The data was preprocessed using robust multichip analysis (RMA) [81, 168, 169].

For each target gene:

1. The optimal model selected above was parameterized on the best regulator set by fitting to a training set consisting of a random 90% subset of the TF induction data set. The fitting assumption is that residuals (fit errors) follow a normal distribution with zero mean but variable variances.
2. The parameterized models were then assessed in their ability to predict the remaining 10% of the TF induction data set. Prediction accuracy was evaluated as sum of squared errors (SSE) between prediction and actual expression values.
3. Steps 1,2 were repeated 10 times – i.e. 10-fold cross-validation – and the overall accuracy of the model determined by averaging the results of each step 2 and correcting using AIC to penalize more complex model structures.

Comparison of Accuracy versus a Random Selection of TFs

To determine the degree to which our predicted regulatory network is responsible for the accuracy of the final selected models – rather than the model selection process – for each target gene we compared the accuracy of the predicted model to a model based on a random set of the same number of TFs.

Random TFs were initialized to the set of all possible MTB TFs excepting the TFs predicted to regulate the target. Moreover, to eliminate TFs predicted by our regulatory network to be correlated in their expression with the target gene, we removed from the random set all TFs that directly bind to or are bound by the TFs regulating the target gene. From the remaining set of TFs, 20 random sets were selected, where each set had the same number of TFs as used in the target gene model. For each random set, model selection and accuracy on 10-fold cross-validation was performed – thus the best fitting model structure was selected independently for each random set.

A distribution of prediction SSE was generated from these random TF sets. The SSE for the actual target gene model was then compared to random set SSEs with a standard score, where a negative z-score means the true model performs better than average random. Also, to correct for large variances, the rank of the true model compared to random SSEs is also calculated, i.e. indicating where it stands compared to 20 random SSEs sorted in ascending order.

Prediction of Hypoxia Time Course Expression Data

To evaluate the generalization of predictability to another independent dataset, models parameterized from the entire TF induction data were used to predict expressions in hypoxic condition. This hypoxia time course expression data was preprocessed using RMA and expression of days 1 to 14 were normalized relative to day 0. The RMA preprocessing was done independent of training data, i.e. TF induction data.

Accuracy was measured by SSE between predicted expression and the actual time course expression of a target gene both scaled between 0 and 1. The expressions were scaled by subtracting the minimum and normalizing by maximum values. Only genes whose expression changed by more than 2-fold, prior to this normalization, were considered.

Similarly, the predictive power of the regulatory network in an independent condition was assessed by comparing accuracy to that of randomly selected TFs. For each random TF set, the optimal

model was independently found and parameterized on the TF induction data and then assessed on hypoxia time course expressions. The SSE for actual TF set was compared to SSEs of random TFs with a standard score and rank, as described above.

Summary of Modeling Results

TF Induction Cross-validation Results: Out of 3002 genes which have binding with impulse height more than 1%, significant models can be generated for 1696 (48%) with p-value less than 0.005. Out of this 1696, 1411 have good predictive power compared to random, i.e. have negative z-score (83% of modeled genes = 47% of total).

Hypoxia Prediction Results: With removing genes that don't change significantly in hypoxic data, i.e. have less than 2 fold change in expression during the 14 day time course, there are 2506 genes in total in hypoxic condition dataset. For 1615 (64%) of these we can generate significant 'condition free' models from over-expression data with p-value less than 0.005. Out of these genes, 838 (52% of modeled genes = 33% of total varying genes) have good predictions compared to random, i.e. have negative z-score. Results are summarized in the **Table S3-3**.

Histogram of Model Structures: The distribution of model structures selected as optimal is presented in **Figure S3-12**. Model structure numbers are:

(1) Linear model without interaction terms

(2) Linear model without interaction term with a sigA expression term

(3) Linear model without interaction term with rpoA expression term

(4) Linear model without interaction term with sigA and rpoA expression terms

(5) Linear model with interaction terms

(6) Linear model with interaction terms and a sigA expression term

(7) Linear model with interaction terms and an rpoA expression term

(8) Linear model with interaction terms with sigA and rpoA expression terms

DREM Analysis

DREM models gene expression as a set of paths where at each time point each path can split into two or more subpaths as a consequence of TF regulation. Genes are associated to a path as a function of the TFs that bind the gene in the regulatory network, and similarity in expression to other genes in the path. DREM has two data inputs – microarray data and TF binding data. We used the hypoxia and re-aeration Nimblegen expression time course data over 7 days of hypoxia and 7 days of re-aeration. The TF binding list was generated based on our ChIP-Seq data under the threshold of 1% of maximum impulse coverage for a TF. TF binding sites in intergenic upstream, genic upstream and genic in-gene regions were considered. After decomposing the time course expression data into a set of expression paths we compared the path, the expression of each TF binding genes in the path, and the predicted regulatory role of the TF from the regulatory network (based on Z-score values) to assess the degree to which expression patterns might reflect the direct action of transcription factors.

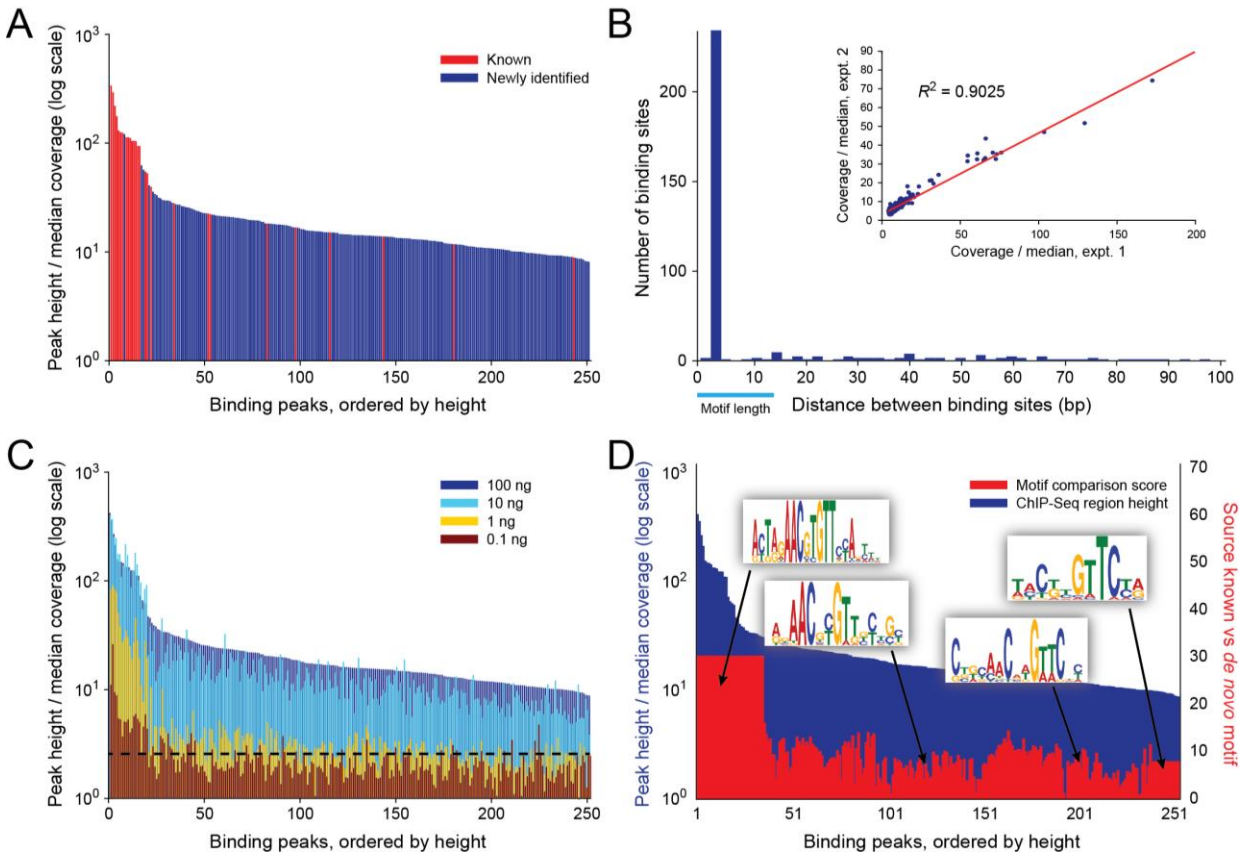


Figure 3-1: ChIP-Seq Binding Shows High Sensitivity, Reproducibility, and Sequence Specificity. (A). We identify all known binding sites for KstR. Binding site heights are plotted as bars and are ordered by peak height. Red bars indicated previously identified binding sites. Blue indicate newly identified sites by our method. We also identify all known sites for DosR (**Figure S3-3**). **(B)** Binding site identification is highly reproducible. Bar plot shows the distance between corresponding sites in two replicates for KstR. The line indicates the length of the KstR binding site motif. The majority of replicates fall within the known motif. Inset shows relationship of peak height between corresponding peaks in two replicates. Coverage for enriched sites is highly correlated between replicates (R^2 > 0.83 for all TFs). **(C)** Increasing TF expression increases peak height. Shown are plots of peaks identified at different levels of KstR induction, with the legend indicating concentration of ATc per mL of culture. Corresponding peaks are plotted at the same position on the horizontal axis. The solid line indicates the threshold required to call a peak in our system. With low KstR expression, only a small number of peaks are identified. As expression increases, overall peak heights increase. The strongest known peaks show marked increases in peak heights and additional lower peaks are revealed. At the highest levels of expression, the differences in peak heights begin to diminish. **(D)** Binding peak height correlated with motif structure. KstR is known to bind to a palindromic motif with accessory high information content bases. As expected, the canonical motif is identified in all strong binding sites. At weaker sites, however, we detect degraded motifs including motifs that lack accessory bases (known to play an important role in shaping affinity[170]) and motifs that include only one half of the palindrome[86]. Motif score is negative log10 of MEME e-value.

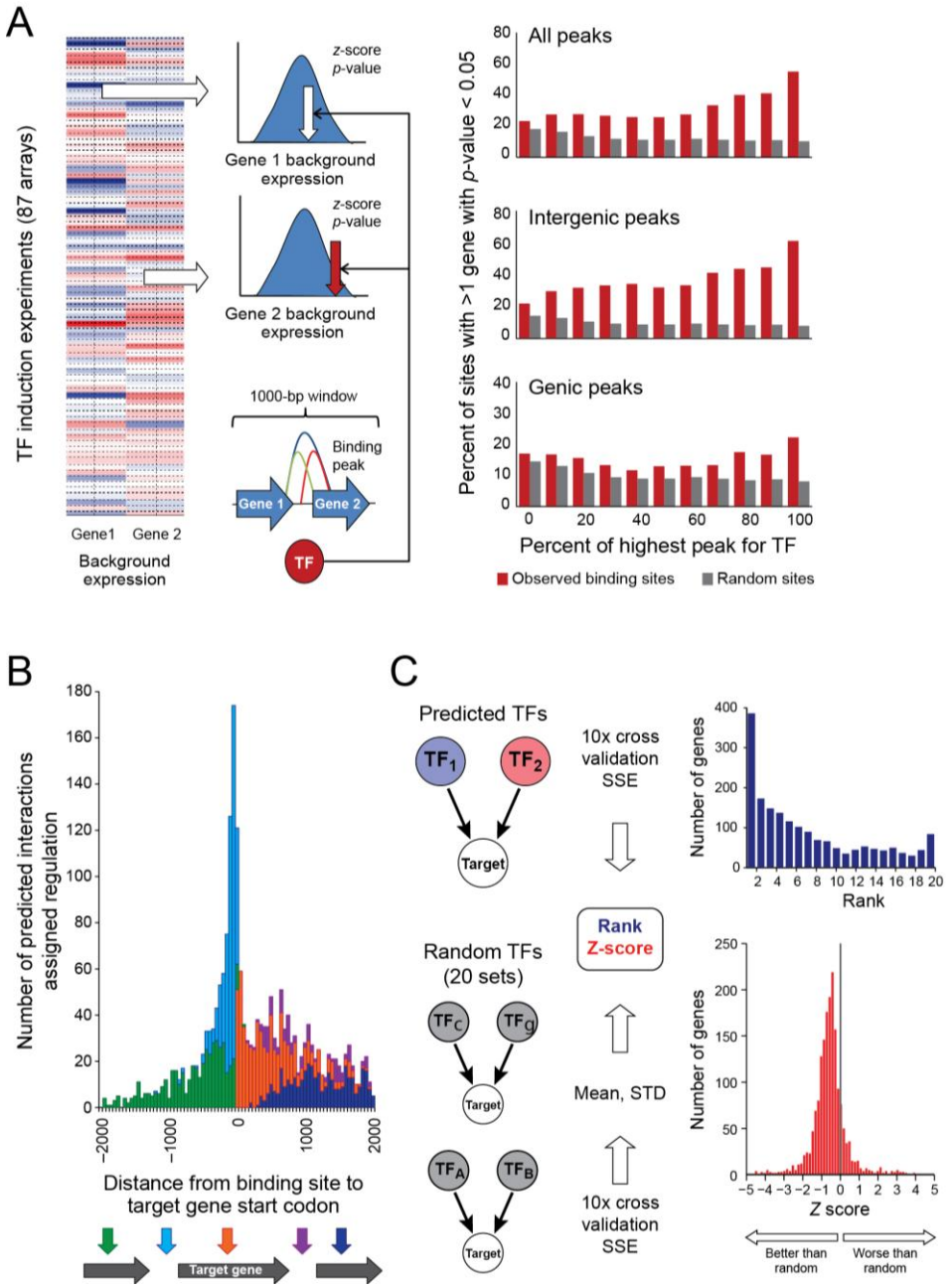

A
TF induction experiments (87 arrays)
Gene1 Gene 2
Background expression
z-score
p-value
Gene 1 background expression
Gene 2 background expression
1000-bp window
Binding peak
Gene 1
Gene 2
TF
Percent of sites with >1 gene with p-value < 0.05
All peaks
Intergenic peaks
Genic peaks
Percent of highest peak for TF
Observed binding sites
Random sites
B
Number of predicted interactions assigned regulation
Distance from binding site to target gene start codon
Target gene
C
Predicted TFs
Target
Random TFs (20 sets)
10x cross validation SSE
Rank
Z-score
Mean, STD
Number of genes
Rank
Z score
Better than random
Worse than random

Figure 3-2: Associating ChIP-Seq Binding with Regulation and Assessing Predictive Power. (A) Associating binding sites with potential regulation**.** For each site, we examine all target genes in a window of 1Kb around the site to determine if induced over-expression of the corresponding TF significantly increases or decreases expression of these genes. The expression of a target gene when the TF is induced is assigned a z-score and corresponding two-tailed p-value based on a background distribution determined from control experiments and the induction of TFs not associated with the target gene. If any gene in the window has a p-value<0.05 after multiple testing, the peak is classified as having a potential regulatory role. Positive z-scores suggest activation, negative repression. The proportion of genes that can be assigned a potential regulatory role as a function of peak height relative to the maximum for that TF, and genomic location is shown. The histograms display the fraction of peaks assigned regulation as a function of peak height threshold. Peaks equal or greater than threshold are included in each bin. **(B)** Distance between binding sites and associated target genes. Stacked histogram of the number of peaks assigned regulation as a function of the distance to the start codon of the predicted target gene and colored by peak location as shown. **(C)** Prediction of expression from TF induction experiments. We developed computational models relating the expression of target genes to the expression of TFs predicted to bind the target. Model selection and verification was performed as described in the methods. The sum-squared error (SSE) of prediction versus actual expression was determined using 10x cross-validation. Performance was assessed by comparison to a distribution of SSE from 20 models with random TF assignments and the same model structure. The top histogram (blue) displays the results as a histogram of the rank of the actual model relative to the 20 random models. The bottom histogram (red) displays a histogram of the results of these comparisons as a z-score of the actual model relative to the distribution of SSEs for the random model. Fully 83% of verified models display better accuracy as compared to random TFs.

-

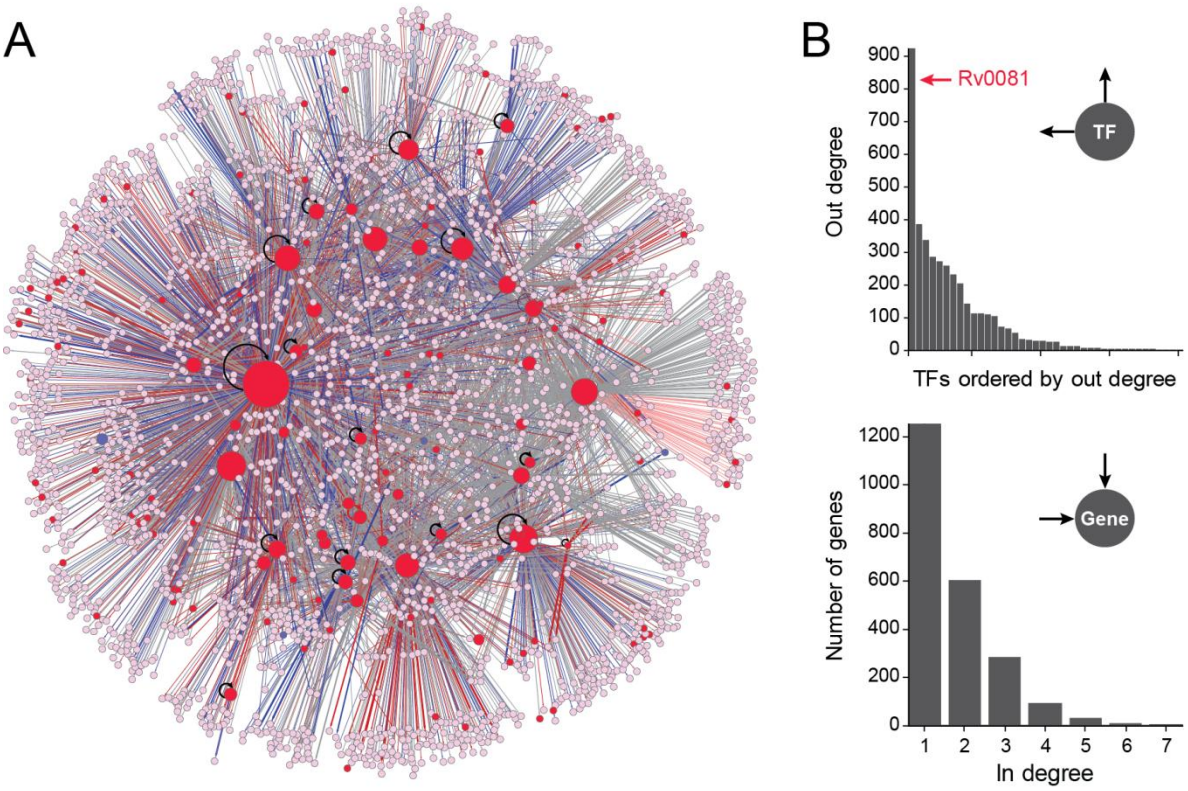


Figure 3-3: M. tuberculosis Regulatory Network Model. (A) Transcriptional regulatory network model based on ChIP-Seq binding and TF induction expression data for 51 transcription factors. The network encompasses 2704 genes, including 141 transcription factors, and 5521 TF-gene interactions based on 9865 binding sites from 6485 regions of enrichment. Nodes represent genes and red nodes are transcription factors. Edges indicate links between TFs and genes based on ChIP-Seq binding. Edges are colored by z-score (as described in **Figure 3-2**) with red edges indicating positive z-scores and activation, and blue indicating negative z-scores and repression. Grey edges indicate links without significant z-scores or TFs for which induction expression data was not yet available. The width of edges indicates the height of the corresponding binding site relative to the maximum binding site for the corresponding TF. The size of TF nodes is proportional to the TF out-degree. A TF-target gene link was included if the TF has a binding peak in either the upstream or downstream intergenic regions for the target gene, or in the gene itself. Links were also included for peaks in upstream genes if the peak was within 500 bp of the target gene and the interaction has a z-score>1 **(B)** Out-degree and in-degree for all TFs. Out-degree for each TF is plotted in the top figure ordered by out-degree. The Y-axis is fixed at 900 to better-visualize the rightward skew of distribution. Rv0081 binds 1120 sites in the genome. The bottom figure shows a distribution of in-degree for all target genes.

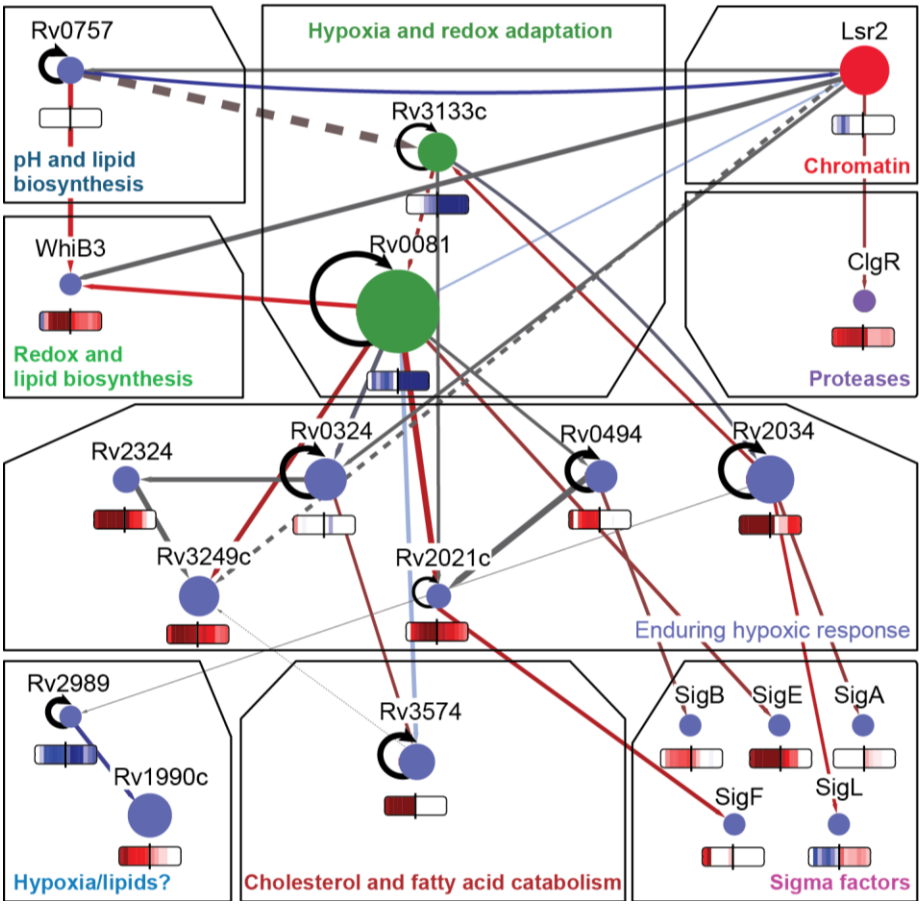


Figure 3-4: TF Regulatory Interaction Subnetwork Linking Hypoxia, Lipid Metabolism, and Protein Degradation. The network reveals a remarkably interconnected subnetwork linking hypoxia and redox adaptation, lipid and cholesterol degradation, lipid biosynthesis, and protein degradation that is supported by and links together reports in the literature, as described in the text. The figure shows a subset of the regulatory network model showing just interactions between selected transcription factors (the full TF-TF interaction network is shown in **Figure S3-9**). Edge color and width are as in **Figure 3-3.** Selected TFs are color coded by functional association and heat maps show expression data during hypoxia and re-aeration time-course as described in **Figure S3-10**.

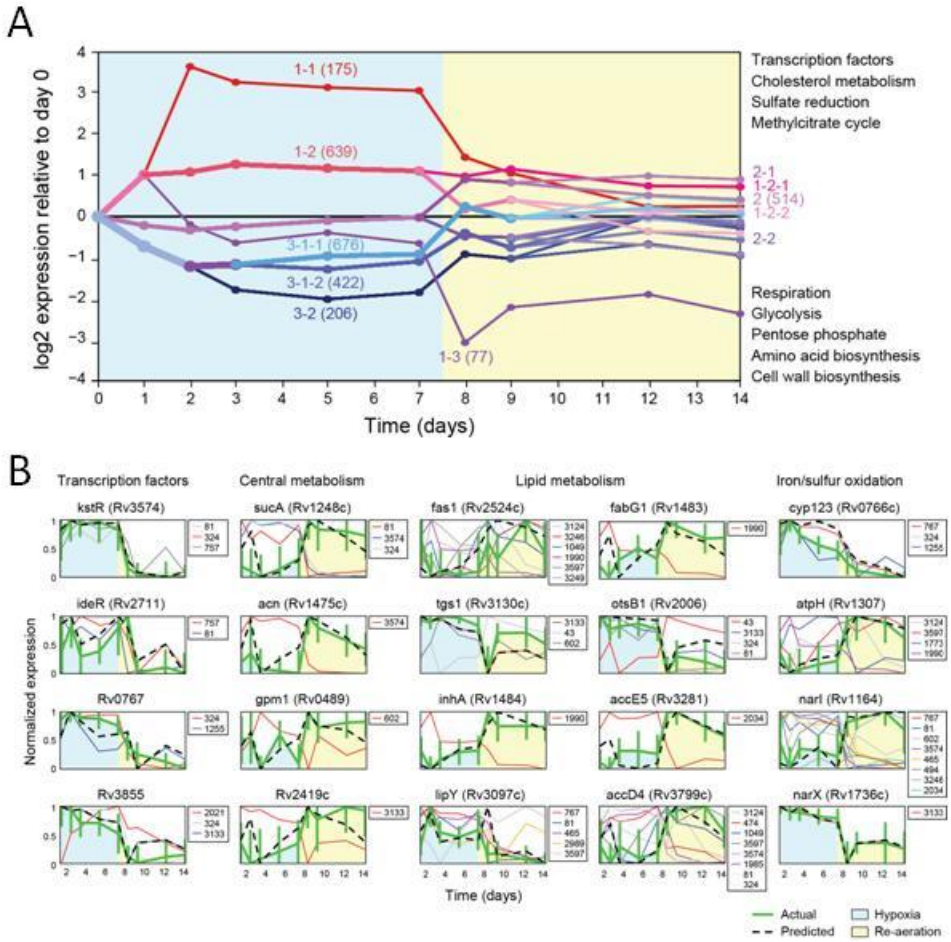


Figure 3-5: Predicting Regulators and Gene Expression during Hypoxia and Re-Aeration. (A) Hypoxia leads to widespread changes in expression. Clustering of gene expression into nested paths using DREM is shown. Each line indicates a path, or cluster of genes with similar expression profiles. Path names are shown with the number of genes in the path indicated in parentheses (e.g. 1-2 (639) is path 1-2 that has 639 genes). Line thickness is also proportional to the number of genes in the path. Text to the right of the figures indicates categories enriched in paths upregulated (top) or downregulated (bottom) during hypoxia. **(B)** Using the models described in **Figure 3-2**, we can predict the pattern of expression of 33% of genes whose expression changes during hypoxia and re-aeration. Selected examples are shown. Green filled lines are actual scaled expression and dotted black are predicted. Colored lines are TFs predicted to regulate the target gene based on the network model and used as regressors. Gray lines indicate other binding TFs not selected as regressors (see Methods). All expression data are normalized to maximum and minimum over the time-course. In all cases, predictions are better than for a random set of the same number of TFs. Additional predictions provided in **Figure S3-11**.

SUPPLEMENTARY FIGURES (S3-x)

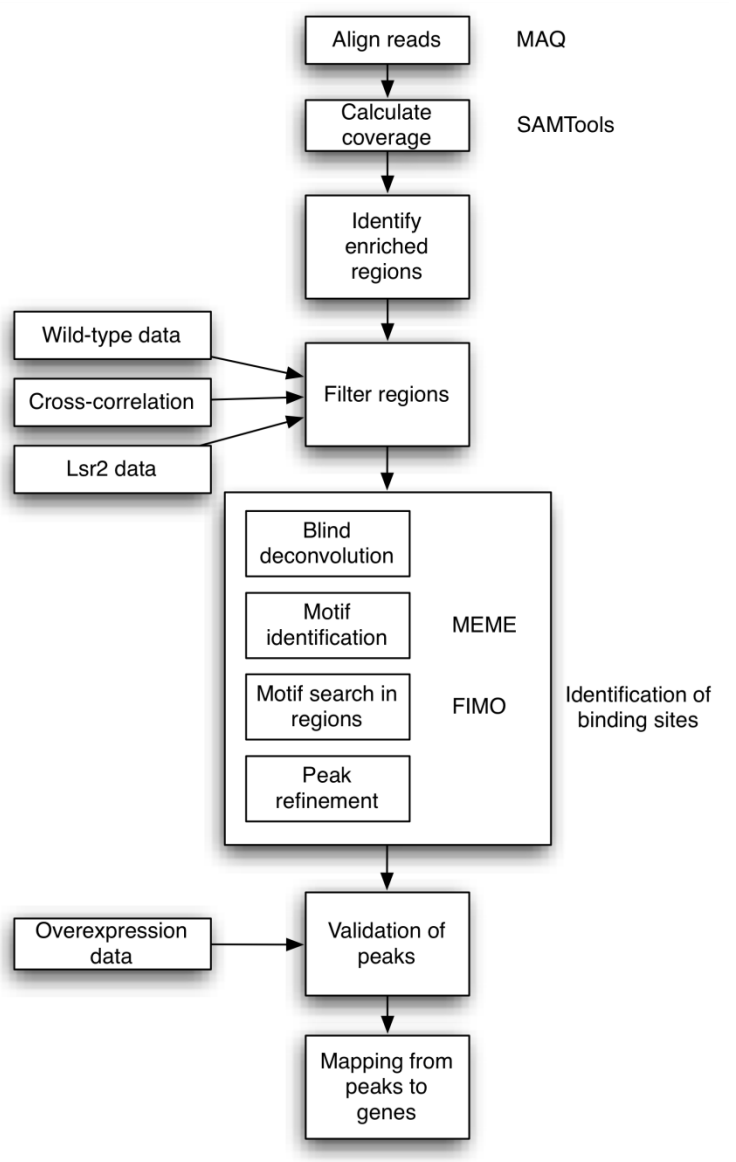


Figure S3-1: Overview of Analysis Pipeline. Reads are mapped to the H37Rv genome, scored against a lognormal distribution, and filtered to remove computational and experimental artifacts. Blind deconvolution and motif finding are used to identify binding events at single nucleotide resolution.

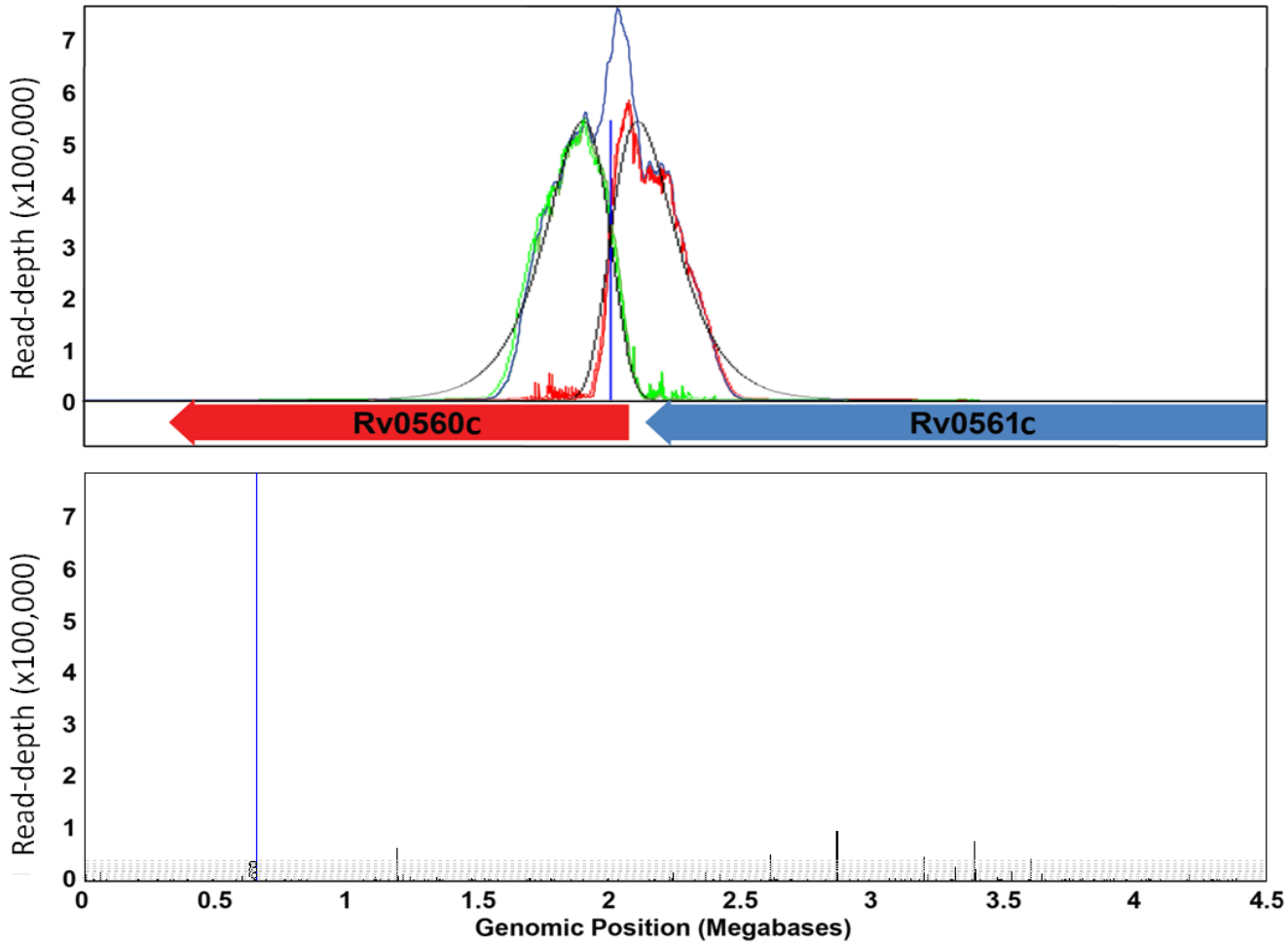


Figure S3-2: Example MTB ChIP-Seq Peak and Genome-Wide Binding for Rv2887. The top panel displays the read coverage for a single binding site of Rv2887. The total coverage is shown in blue. The forward and reverse strand coverage is shown in red and green, respectively. The solid black lines represent the blind deconvolution fit of this bimodal strand distribution with the vertical blue line indicating the centerpoint of binding at nucleotide 651,476. The maximum coverage for this binding site is 763,000 as compared to a genome-wide background (median) coverage of 150. The bottom panel displays the genome-wide fold coverage for the same ChIP experiment. The horizontal grey lines indicate increments of the standard deviation of background coverage.

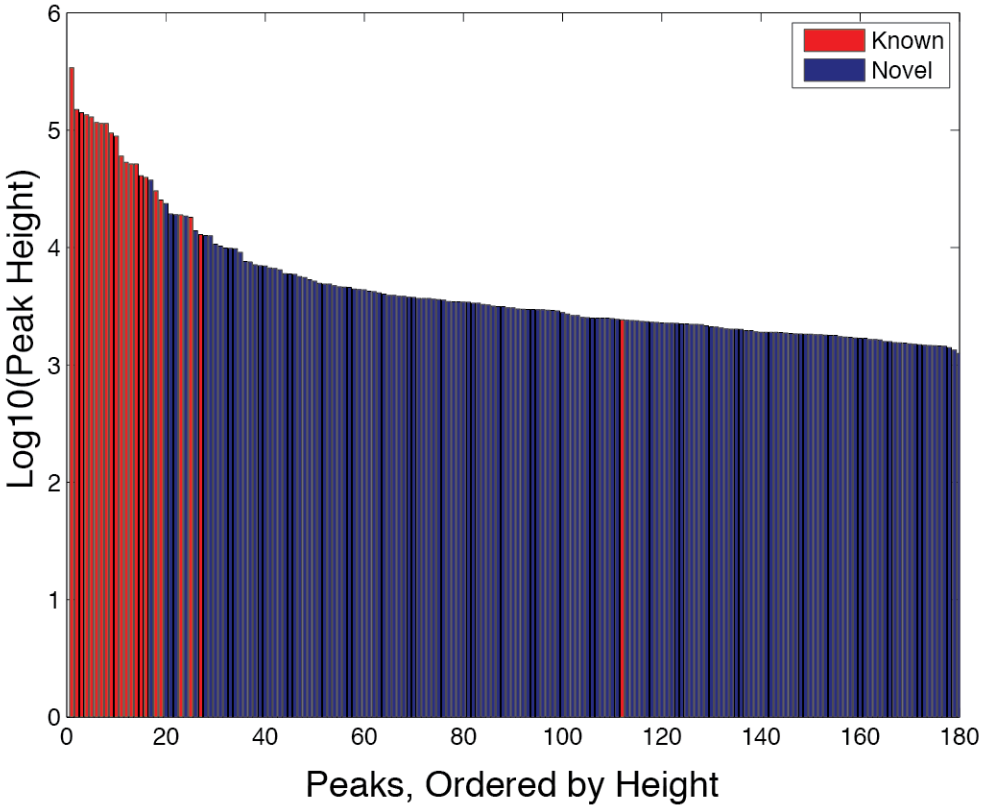


Figure S3-2: Distribution of Peak Heights and Identification of all known Binding Sites for DosR. We identify all known binding sites for DosR. Binding site heights are plotted as bars and are ordered by peak height. Red bars indicated previously identified binding sites. Blue indicate newly identified sites by our method

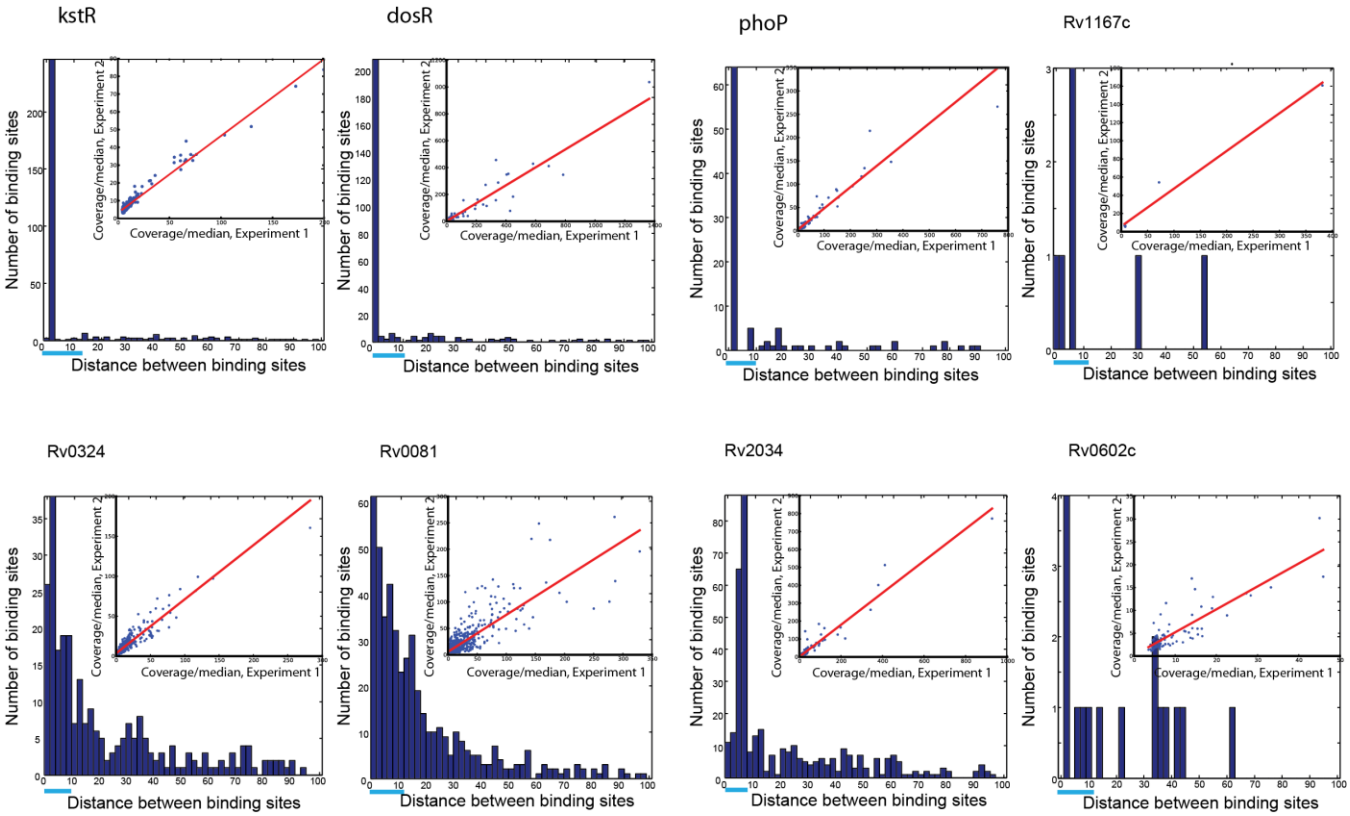


Figure S3-3: ChIP Binding Shows High Reproducibility in Peak Height and Location. The bar plot shows the distance between corresponding binding sites in two replicates for the same transcription factor. The blue line under the X-axis indicates the width of the predicted motif. The scatter plot shows the correlation of coverage in replicates.

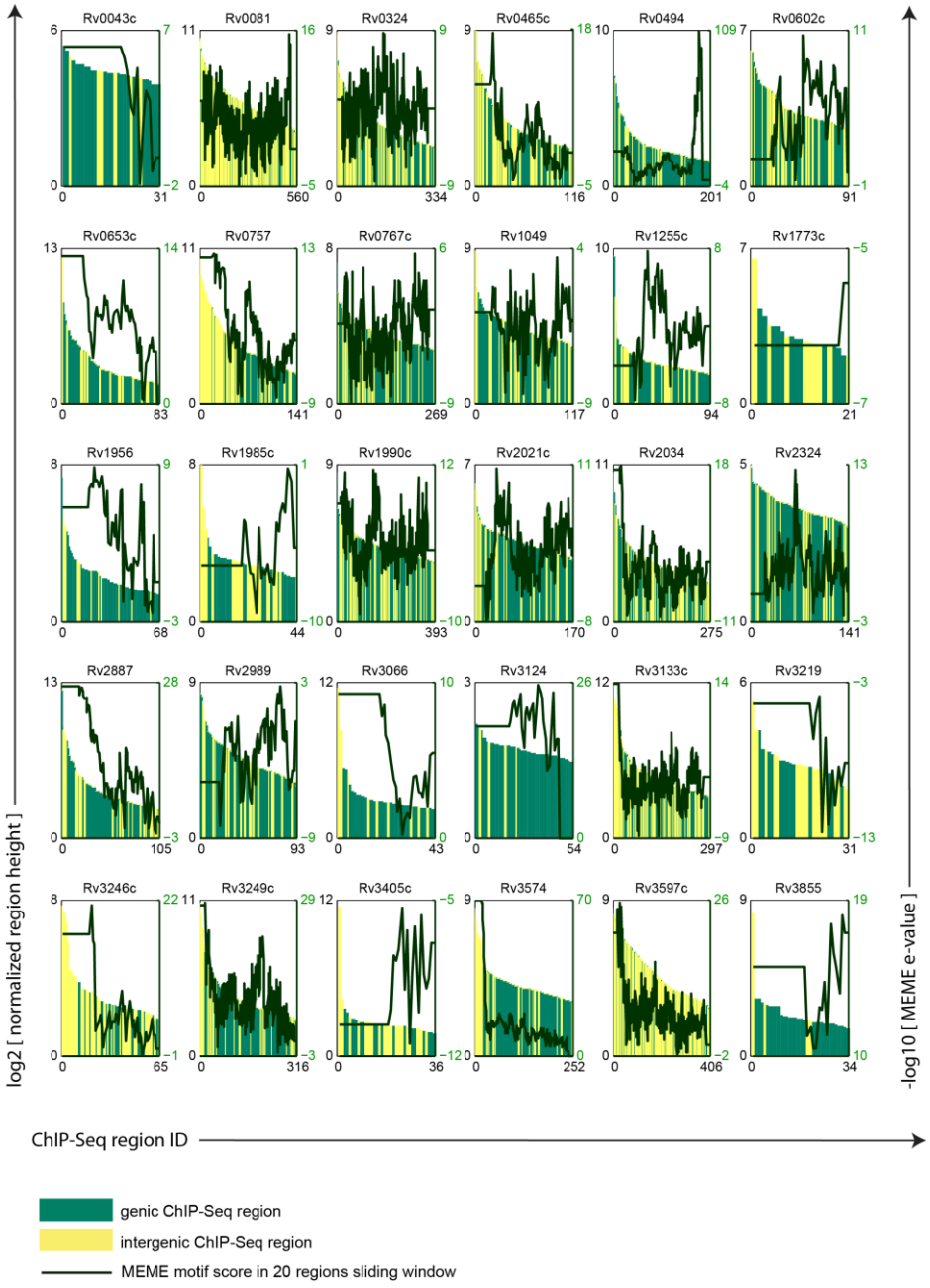


Figure S3-4: Binding Height Correlation with Motif Strength. For every ChIP-Seq experiment, we sort regions by height and predict the *de novo* motif in a sliding window of 20 peaks using MEME. Black line shows the *de novo* motif score (secondary Y-axis – negative log10 e-value of the motif). Green and yellow bars show genic and intergenic enriched regions respectively (primary Y-axis – log2 normalized region height).

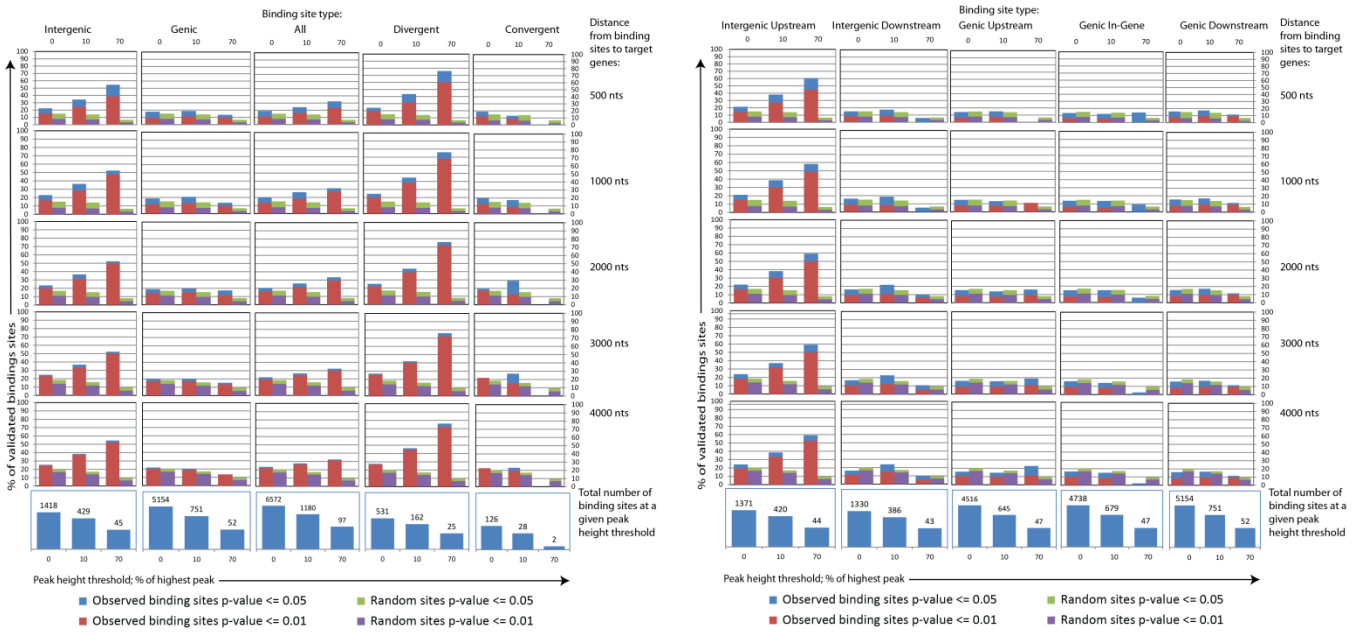


Figure S3-6: Summary of Overall Assignment of Regulation to Binding Events. Peaks are assigned potential regulation as described in **Figure 3-2**. X-axis of every plot is binding site coverage (normalized by coverage of the highest binding site of the experiment). We group binding sites in three groups by their relative coverage, 0-10%, 10-70%, 70-100%, and test these three groups independently. Y-axis of every plot is the percentage of binding sites within the chosen group that have at least one target gene validated. Blue and red bars correspond to two significance levels - 0.05 and 0.01. The green and purple bars show the estimated level of validation we expect at random for the same significance levels. At the bottom of the figure, bar plots show how many binding sites belong to every tested subgroup. We test five window widths - 500, 1000, 2000, 3000, and 4000 nucleotides – around the binding site. Graphs corresponding to the same window size are in the same row. We also compare the validation of binding sites from various subgroups. We group binding sites by their location relative to neighboring genes: intergenic, genic, convergent (in the intergenic region of two convergent genes), and divergent (in the intergenic region of two divergent genes). We also compare different methods of selecting target genes of the binding site within a given window size. Binding site can be located upstream or downstream of its target while being intergenic; or binding site can be upstream, in the gene, or downstream while being genic.

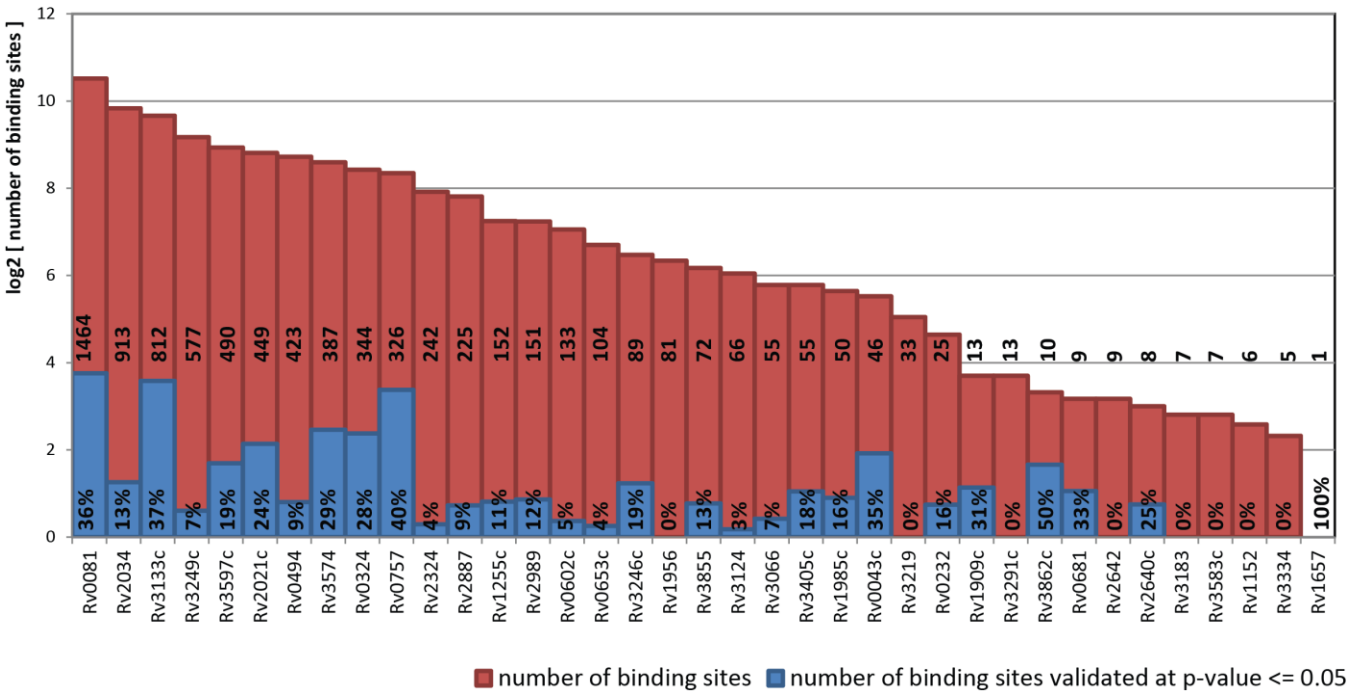


Figure S3-7: Variation in Degree of Regulatory Assignment between Transcription Factors. Red bars indicate the number of binding sites found for a transcription factor. Blue bars show the percentage of the binding sites that can currently be assigned regulatory roles given the approach and thresholds used in **Figure 3-2A**.

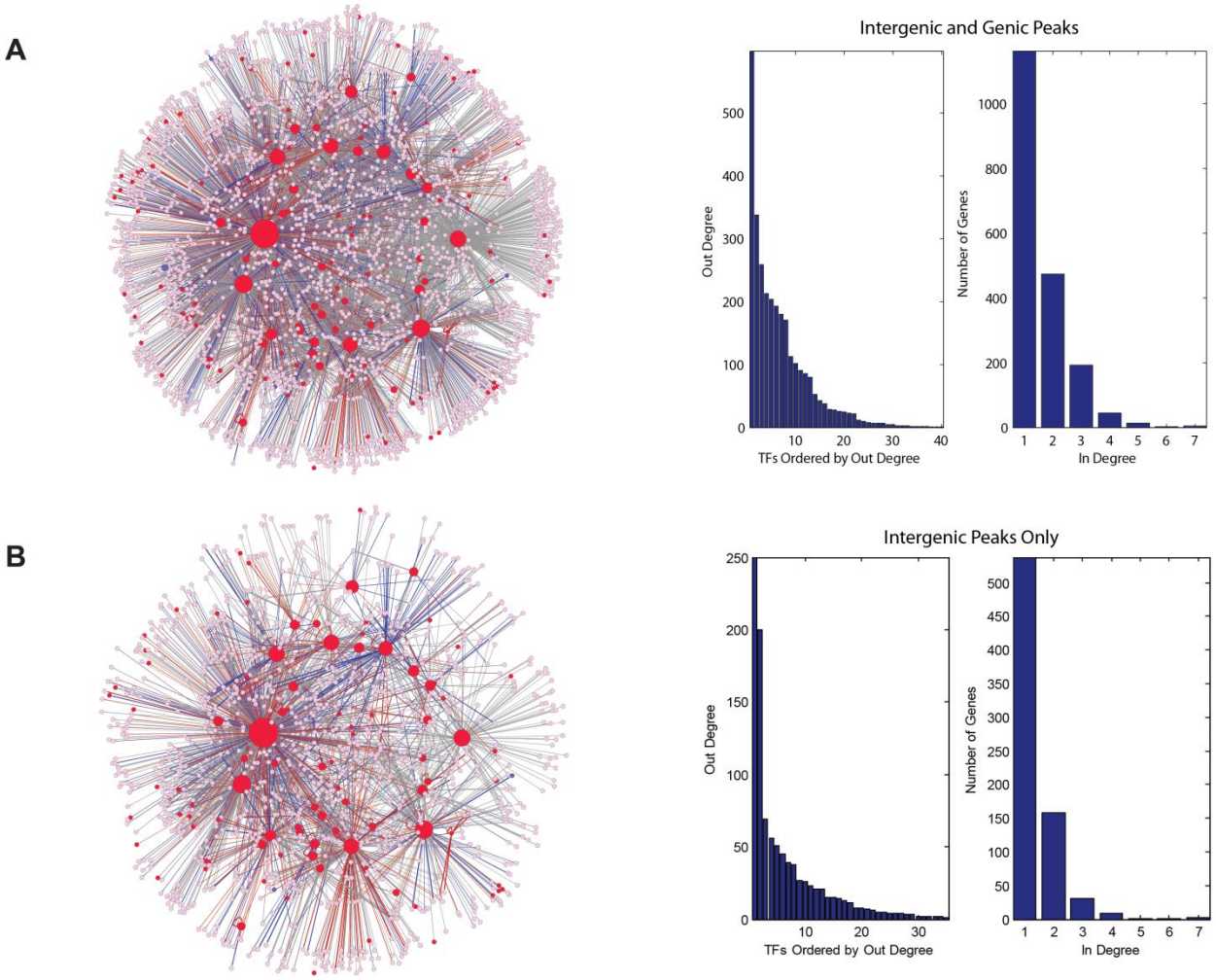


Figure S3-8: Regulatory Network Models using Different Criteria for Including TF-Gene Links. (A) Network including links between a TF and target gene if the TF binds in the upstream intergenic region or in the target gene itself independent of possible predicted regulatory role. **(B)** Network only including links if the TF binds in the upstream intergenic region independent of possible predicted regulatory role.

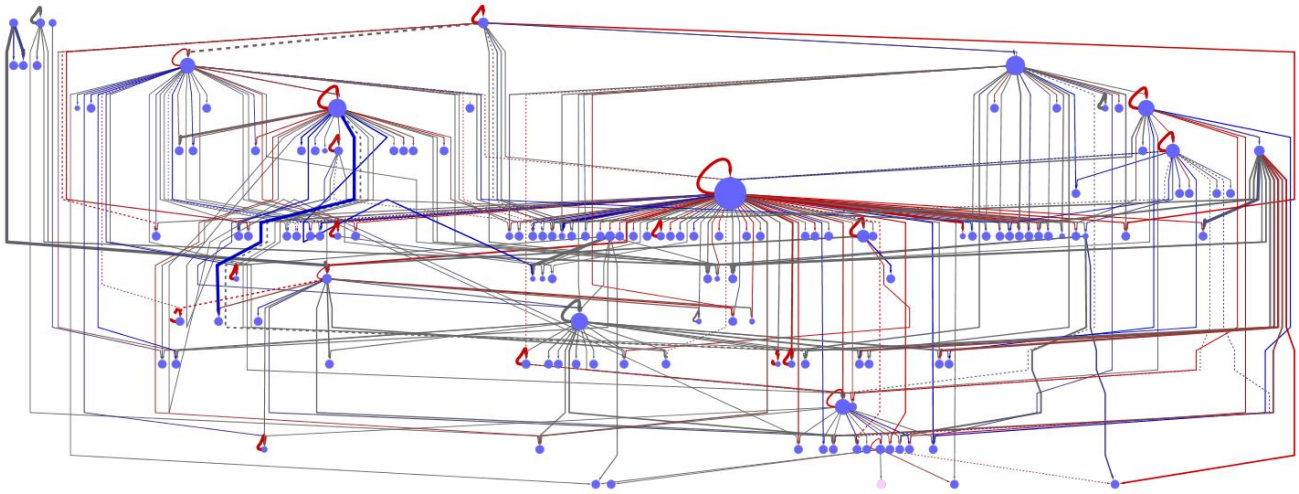


Figure S3-9: Regulatory Interaction Network Showing only Regulators. A Cytoscape map of regulatory interactions between transcription factors only. Edges indicate links between TFs based on ChIP-Seq binding. Edges are colored by z-score (as described in **Figure 3-2**) with red edges indicating positive z-scores and activation, and blue indicating negative z-scores and repression. Grey edges indicate links without significant z-scores or TFs for which induction expression data was not yet available. The width of edges indicates the height of the corresponding binding site relative to the maximum binding site for the corresponding TF. The size of TF nodes is proportional to the TF out-degree. A TF-target gene link was included if the TF has a binding peak in either the upstream or downstream intergenic regions for the target gene, or in the gene itself. Links were also included for peaks in upstream genes if the peak was within 500 bp of the target gene and the interaction has a |z-score|>1.

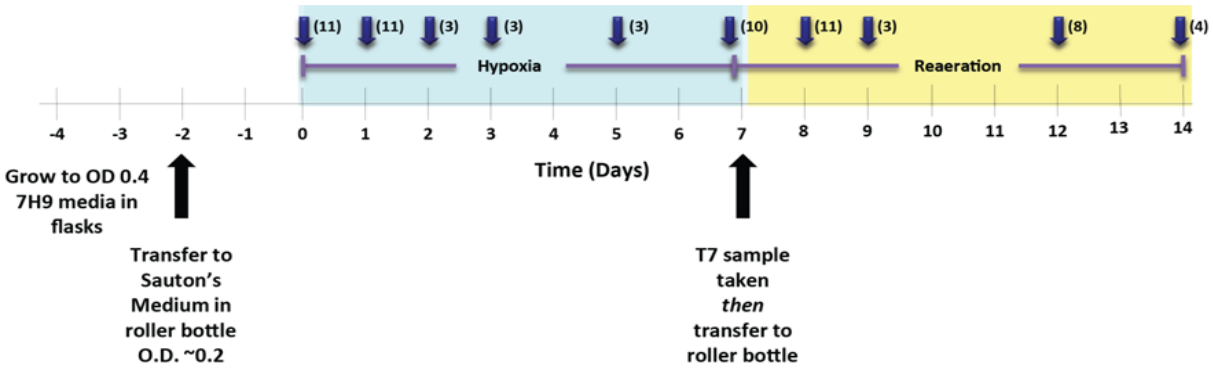


Figure S3-10: Hypoxia Profiling Sampling Protocol. Schematic of the time points sampled. For all experiments, MTB was grown to OD 0.4 in 7H9 medium, transferred to Sauton's medium (see Methods) in a roller bottle and grown for two days prior to experimental sampling. Sampling at T0 was performed from the roller bottle. The culture was then transferred to a spinner flask for 7 days during which time oxygen tension could be controlled ("hypoxia"). At day 7, the culture was returned to a roller bottle for 7 days of re-aeration. The T7 sample was taken prior to the transfer. Samples for profiling were taken at the time points indicated. The number in parenthesis indicates the number of replicates generated at each time point. Samples were generated from 3 different experiments with multiple biological replicates in each experiment.

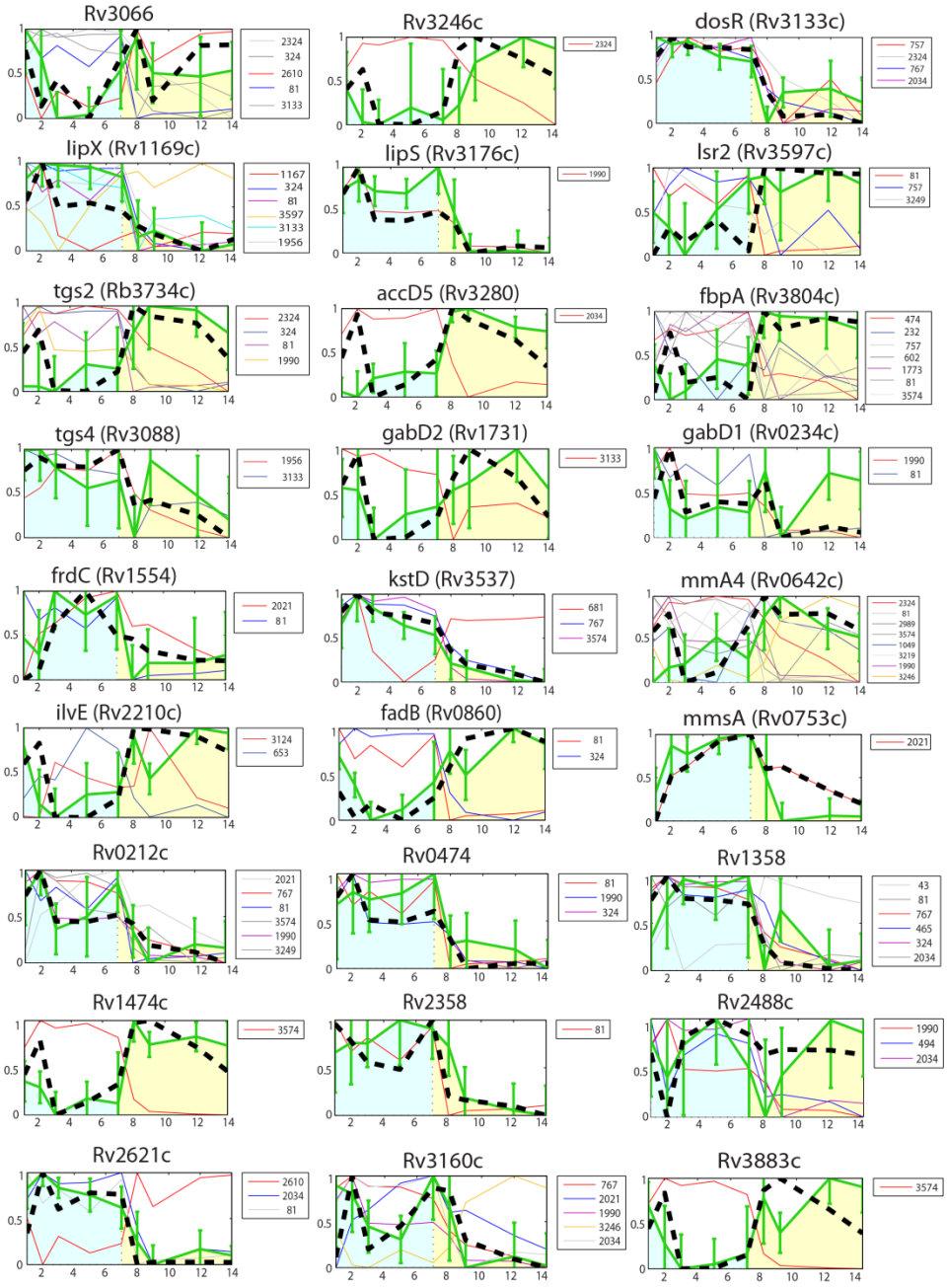


Figure S3-11: Prediction of Hypoxia and Re-aeration Gene Expression for Specific Genes Mentioned in Text. Prediction of expression patterns during hypoxia and re-aeration time course for additional genes mentioned in the main text. Data generated and plotted as describe in **Figure 3-5B** in the main text.

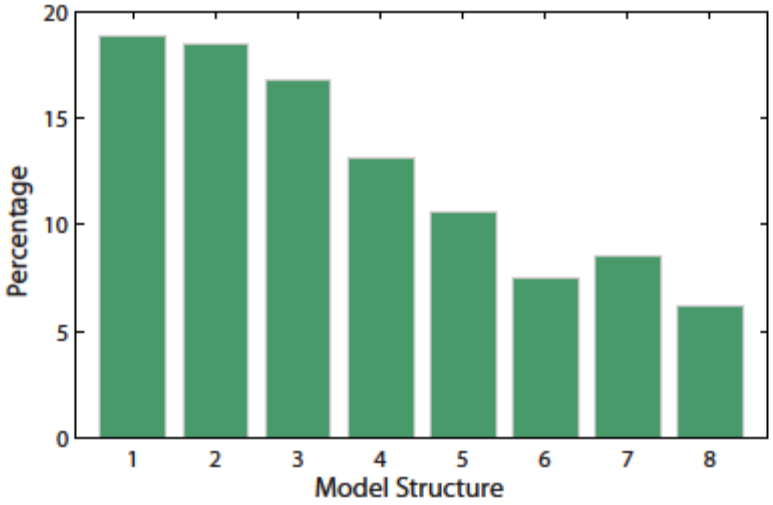
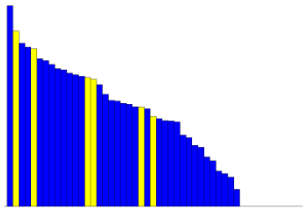
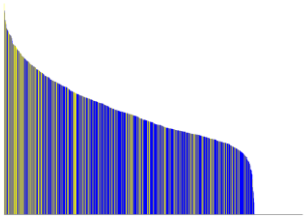
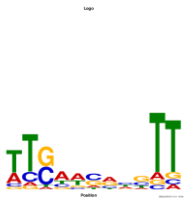
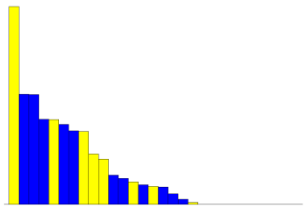
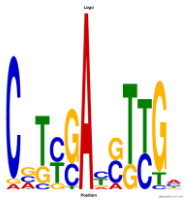


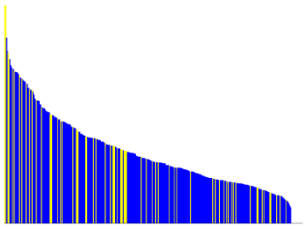
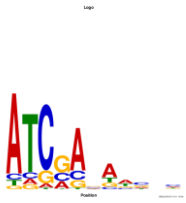
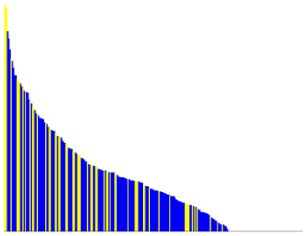
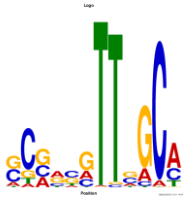
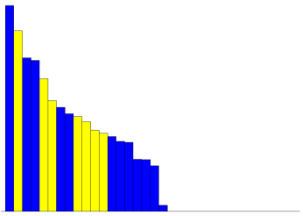
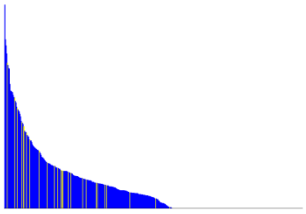
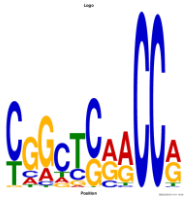
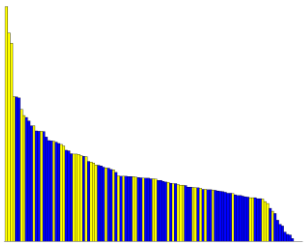
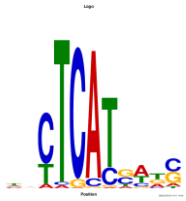
Figure S3-12: Histogram of Model Structures. The distribution of model structures selected as optimal, as described in the methods. Model structure numbers are: (1) Linear model without interaction terms. (2) Linear model without interaction term with a sigA expression term. (3) Linear model without interaction term with rpoA expression term. (4) Linear model without interaction term with sigA and rpoA expression terms. (5) Linear model with interaction terms. (6) Linear model with interaction terms and a sigA expression term. (7) Linear model with interaction terms and an rpoA expression term. (8) Linear model with interaction terms with sigA and rpoA expression terms.

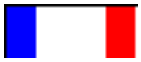
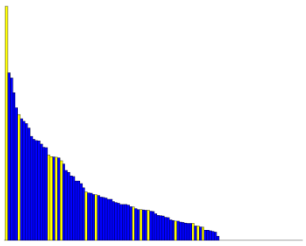
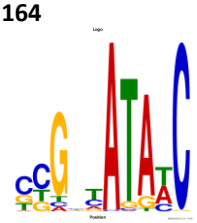
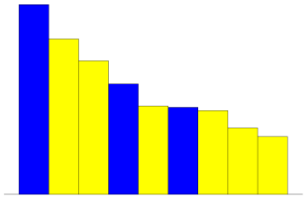
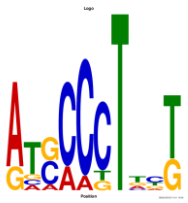
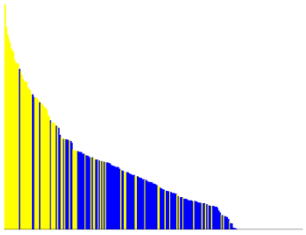
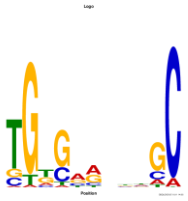
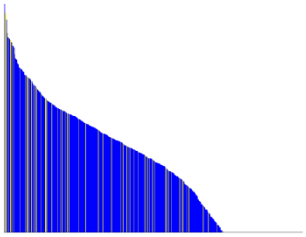
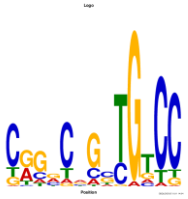
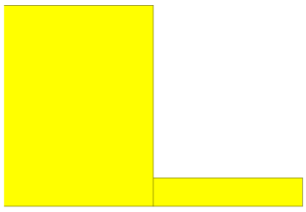
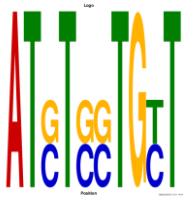
Table S3-1: Summary of ChIP-Seq Peak Calling Statistics

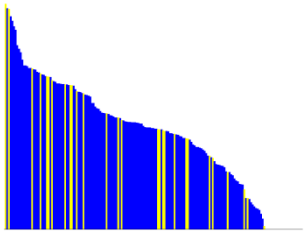
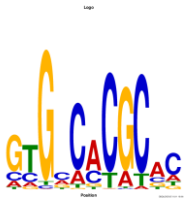
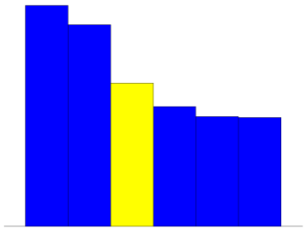
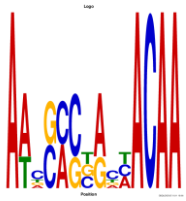
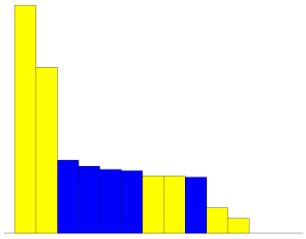
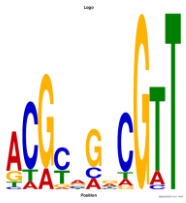
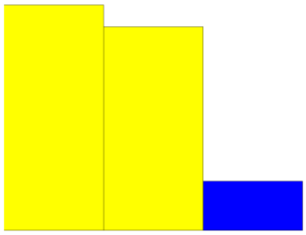
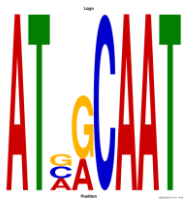
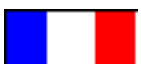
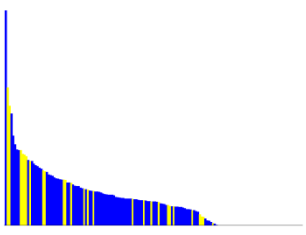
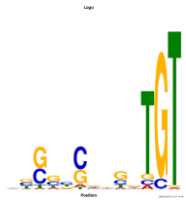
	Count	Fraction
Total number of enriched regions	19,369	-
Regions failing shift filter	8,478	44%
Number of regions failing background filter (WT1)	376	2%
Number of regions failing background filter (WT2)	451	2%
Number of regions failing Lsr2 filter	6,495	34%

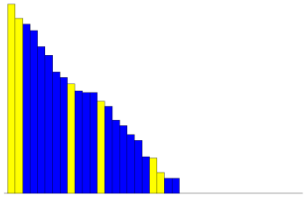
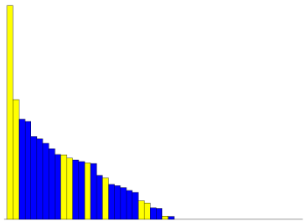
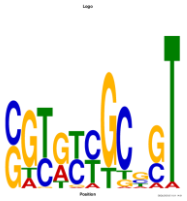
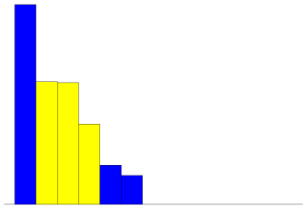
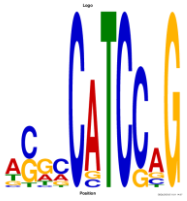
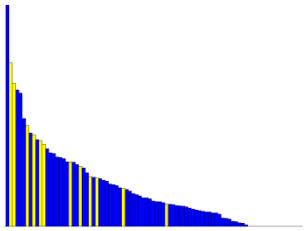
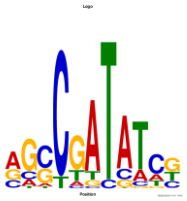
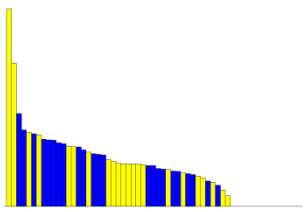
Table S3-2: Details of ChIP Mapping for Each Mapping Factor. The first column identifies the transcription factor interrogated. The second column identifies the number of up (red) and downregulated (blue) targets from the overexpression data (NE indicates no expression data was available). The third column shows the count and distribution of peak heights, and locations, with yellow bars indicating intergenic peaks and blue bars indicating genic peaks. The fourth column contains the motif identified from the entire set of peaks for each transcription factor. The number above the motif represents the negative log of the E-value for the motif.

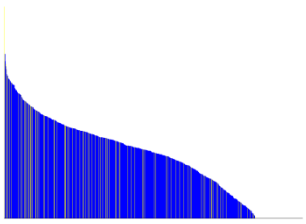
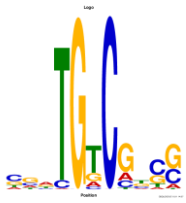
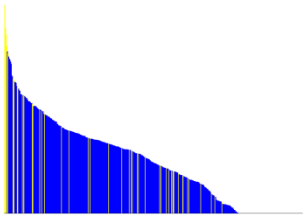
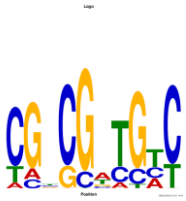
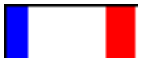
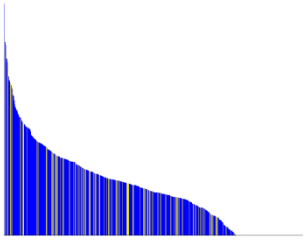
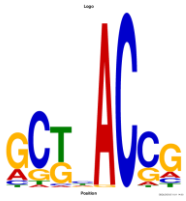
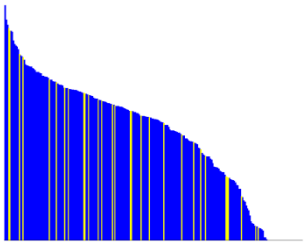
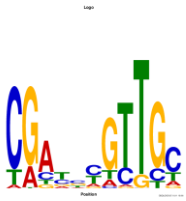
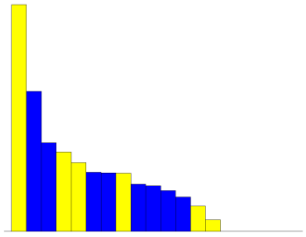
Rv0038		**1/1** 	**-2** 
Rv0043c		**7/39** 	**52** 
Rv0081		**329/680** 	**591** 
Rv0232		**9/16** 	**24** 

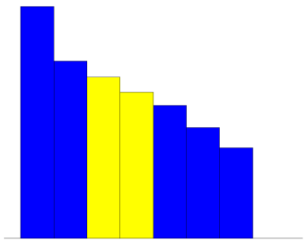
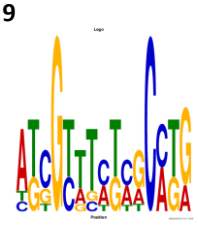
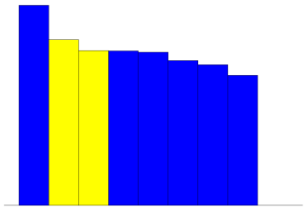
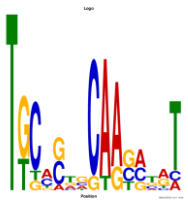
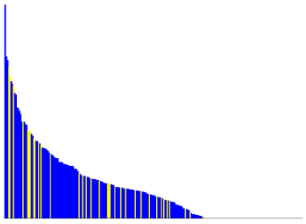
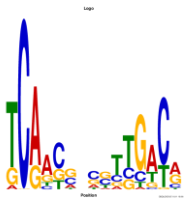
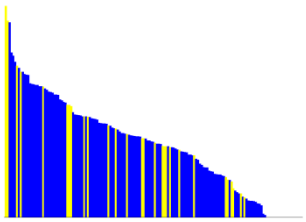
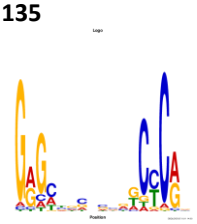
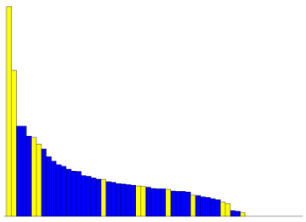
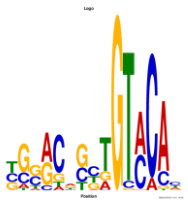
Rv0324		**68/276** 	**111** 
Rv0465c	**NE**	**62/155** 	**269** 
Rv0474	**NE**	**24/8** 	**27** 
Rv0494		**39/384** 	**542** 
Rv0602c		**46/72** 	**158** 

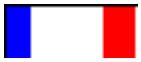
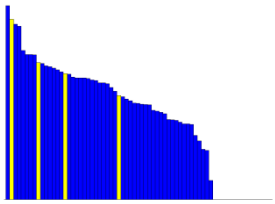
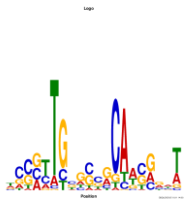
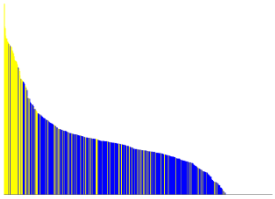
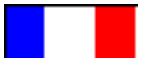
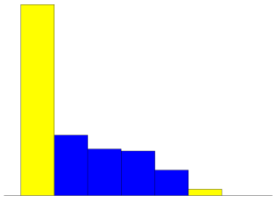
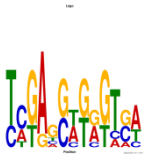
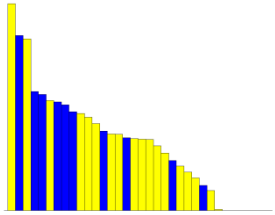
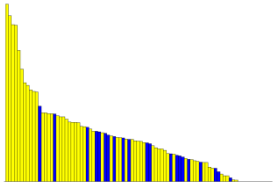
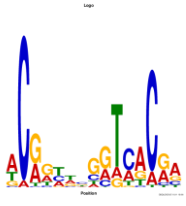
Rv0653c		**21/83** 	**164** 
Rv0681		**6/3** 	**1** 
Rv0757		**82/146** 	**84** 
Rv0767	**NE**	**72/540** 	**585** 
Rv0821c	**NE**	**2/0** 	**-1** 

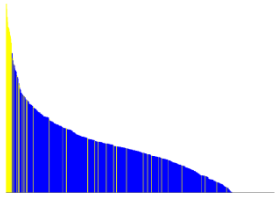
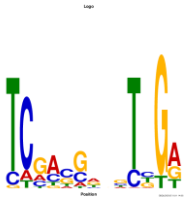
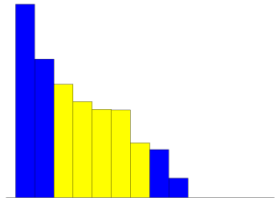
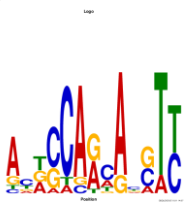
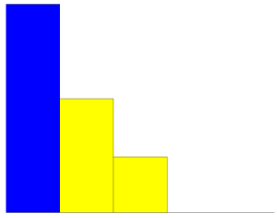
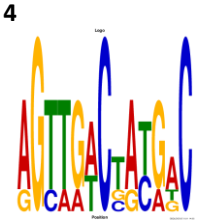
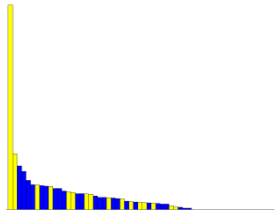
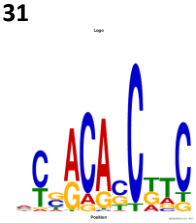
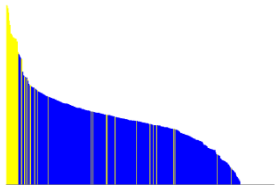
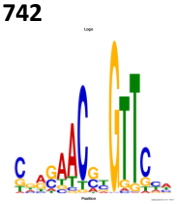
Rv1049	**NE**	**31/136** 	**149** 
Rv1152		**1/5** 	**7** 
Rv1167c	**NE**	**6/6** 	**21** 
Rv1176c	**NE**	**2/1** 	**1** 
Rv1255c		**33/119** 	**60** 

Rv1395	**NE**	**8/29** 	**17** 
Rv1773c	**NE**	**20/25** 	**39** 
Rv1909c		**6/7** 	**12** 
Rv1956		**13/68** 	**92** 
Rv1985c		**24/26** 	**26** 

Rv1990	**NE**	**94/753** 	**682** 
Rv2021c		**48/401** 	**360** 
Rv2034		**124/492** 	**502** 
Rv2324		**33/209** 	**256** 
Rv2610c	**NE**	**7/11** 	**14** 

Rv2640c		**2/6** 	**9** 
Rv2642		**2/7** 	**4** 
Rv2887		**39/186** 	**374** 
Rv2989		**30/121** 	**135** 
Rv3066		**15/40** 	**86** 

Rv3124		**4/82** 	**106** 
Rv3133c		**141/392** 	**812** 
Rv3183		**2/5** 	**-1** 
Rv3219		**21/12** 	**3** 
Rv3246c		**65/24** 	**101** 

Rv3249c		**57/520** 	**579** 
Rv3291c		**7/6** 	**7** 
Rv3334		**4/1** 	**4** 
Rv3405c		**22/33** 	**31** 
Rv3574		**44/343** 	**742** 

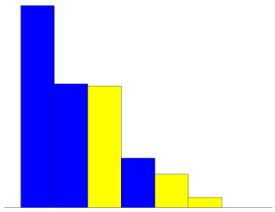
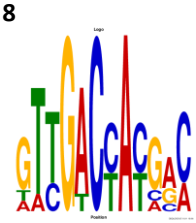
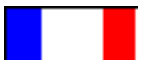
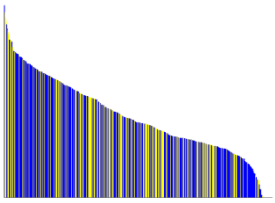
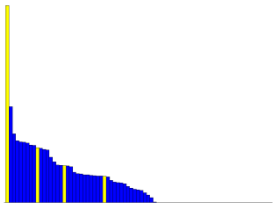
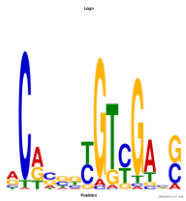
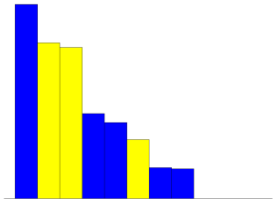
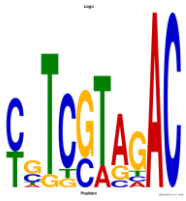
Rv3583c		**3/4** 	**8** 
Rv3597c		**187/303** 	**491** 
Rv3855		**4/64** 	**66** 
Rv3862c		**3/7** 	**7** 

Table S3-3: Summary of Modeling Results Presented in Methods Section

	Total genes with binding (impulse> 1%)	**10x cross-validation (same condition): TF Induction Data Prediction**		**Generalization (independent condition): Hypoxia Time Course Prediction**		
		Genes modeled (p-value <0.005)	Genes predicted better than random (Z-score < 0)	Genes with > 2 fold change in expression during time course	Genes modeled (p-value <0.005)	Genes predicted better than random (Z-score < 0)
Number of genes	3002	1696	1411	2506	1615	838
Percentage		56%	83% of modeled genes 47% of total genes	%84	%64 of changing genes	52% of modeled genes 33% of total changing genes

Acknowledgments

As indicated, the text of this dissertation chapter was modified from a submitted manuscript, and we wish to acknowledge the individuals who contributed to this study. The complete author list of the manuscript is:

James E. Galagan[1,2,3,†], Kyle Minch[4,5,*], Matthew Peterson[1,*], Anna Lyubetskya[1*], Elham Azzizi[1,*], Lindsay Sweet[6,*], Antonio Gomez[1,*], Tige Rustad[4], Gregory Dolganov[7], Irina Glotova[1], Thomas Abeel[3], Chris Mahwinney[1], Adam Kennedy[8], René Allard[9], William Brabant[4], Andrew Krueger[1], Suma Jaini[1], Brent Honda[1], Wen-Han Yu[1], Mark J. Hickey[4], Jeremy Zucker[3], Christopher Garay[1], Brian Weiner[3], Peter Sisk[3], Christian Stolte[3], Diogo Camacho[1], Jonathan Dreyfuss[1], Yang Lui[7], Anca Dorhoi[10], Hans-Joachim Mollenkopf[10,11], Paul Drogaris[9], Julie Lamontagne[9], Yiyong Zhou[9], Julie Piquenot[9], Sang Tae Park[2], Saha Raman[2], Stefan H.E. Kaufmann[10], Robert Mohney[8], Daniel Chelsky[9], D. Branch Moody[6], David R. Sherman[4,5,11,‡], Gary K. Schoolnik[7,13,‡]

[1]Departments of Biomedical Engineering and [2]Microbiology, Boston University, Boston, MA 02215 USA; [3]The Eli and Edythe L. Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [4]Seattle Biomedical Research Institute, Seattle, WA 98109, USA; Molecular and Cellular Biology Program, University of Washington, Seattle, WA 98195; [6]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [7]Departments of Medicine and of Microbiology and Immunology, Stanford Medical School, Stanford, CA, USA; [8]Metabolon Inc., Durham, NC 27713; [9]Caprion Proteomics, Inc., Montreal, Quebec H4S 2C8; [10]Department of Immunology, Max Planck Institute for Infection Biology, Berlin, Germany; [11]Core Facility, Max Planck Institute for Infection Biology, Berlin, Germany; [12]Interdisciplinary Program of Pathobiology, Department of Global Health, University of Washington, Seattle, Washington 98195, USA; [13]Division of Infectious Diseases and Geographic Medicine, Department of Medicine, Stanford Medical School, Stanford, CA, USA

[†]Corresponding Author; *Authors contributed equally; [‡]Co-Last Authors

Chapter 4 – *Mycobacterium tuberculosis* Growth Following Aerobic Expression of the DosR Regulon

The following text is from the article: Minch K, Rustad T, Sherman DR (2012) *Mycobacterium tuberculosis* Growth following Aerobic Expression of the DosR Regulon. **PLoS ONE** 7(4): e35935. Figure numbers have been updated to conform to the formatting of this dissertation; however, the remainder of the text is as published.

Abstract

The *Mycobacterium tuberculosis* regulator DosR is induced by multiple stimuli including hypoxia, nitric oxide and redox stress. Overlap of these stimuli with conditions thought to promote latency in infected patients fuels a model in which DosR regulon expression is correlated with bacteriostasis *in vitro* and a proxy for latency *in vivo*. Here, we find that inducing the DosR regulon to wildtype levels in aerobic, replicating, *M. tuberculosis* does not alter bacterial growth kinetics. We conclude that DosR regulon expression alone is insufficient for bacterial latency, but rather is expressed during a range of growth states in a dynamic environment.

Introduction

Mycobacterium tuberculosis is a remarkably successful pathogen that causes ~9 million new cases of active tuberculosis (TB) disease every year [3]. However, this pool of patients with clinical disease is dwarfed by a vast reservoir of latently infected individuals estimated to consist of ~1/3 of the world's population [2]. A prominent model of TB pathogenesis argues that a heterogeneous bacterial population within latently infected individuals enters into a reversible state of non-replicating persistence or dormancy, induced by diverse stimuli including nutrient deprivation, nitric oxide and hypoxia [171]. *In vitro* study of the latter condition has demonstrated that hypoxia can drive *M.*

tuberculosis in to a viable but bacteriostatic state with concomitant substantial remodeling of the transcriptional and metabolic profile of the cell [91]. One of the earliest mediators of this transcriptional shift is induction of the two-component response regulator, Rv3133c [38].

Rv3133c was initially identified as a regulator differentially expressed in a virulent strain *M. tuberculosis* (DevR) [39]. Subsequently this transcription factor was shown to be induced in response to hypoxia [38], nitric oxide [41], or redox stress [52], and was renamed dormancy survival regulator (DosR) [53]. Induction of DosR results in the expression of ~49 genes under its direct control. Furthermore, because each of these conditions is associated with aspects of bacterial dormancy *in vitro* and clinical latency, it was hypothesized that DosR induction initiated a genetic program that prepared *M. tuberculosis* for survival of bacteriostasis [41]. It appears, however, that DosR regulon induction during the initial hypoxic response is not absolutely required for survival during bacteriostasis (dormancy), as a DosR knock-out shows a modest survival defect upon exposure to short- or long-term defined hypoxic conditions [54]. In the Wayne model of gradual oxygen depletion, genetic manipulations of DosR lead to larger decreases in viability in *M. tuberculosis* and *M. bovis* BCG [53], though it is unclear to what extent nutrient depletion or toxic metabolic by-products impact these observations. Despite uncertainty about the role of DosR in the natural history of TB disease, expression of DosR regulon genes is sometimes cited as evidence of impending *in vitro* dormancy, with implications for the management of clinical latency [172]. This association persists despite the observation that *M. tuberculosis* strains of the W-Beijing lineage constitutively express the DosR regulon [173].

In this work we directly address the question of *M. tuberculosis* growth rate following DosR regulon expression. We demonstrate that under aerobic conditions, ectopic induction of DosR is sufficient to induce the DosR regulon to near wild-type levels, even in the absence of its usual cues. However, this induction does not cause bacteriostasis or otherwise alter the growth kinetics of replicating *M. tuberculosis*.

Materials & Methods

Strains and Culturing Conditions

All experiments were performed using an Rv3133c/DosR knockout *M. tuberculosis* strain, H37Rv:Δ*dosR*, generated in our lab and described previously [40]. In the present work, H37Rv:Δ*dosR* was transformed with an episomal plasmid containing Rv3133c/DosR under the control of the smyc, anhydrotetracycline-inducible, promoter described by Ehrt, et al. [78]. This vector contains a hygromycin B-resistance cassette and was modified to a contain Gateway Recombination™ (Invitrogen) cassette (kind gift of Eric Rubin). We further adapted this destination vector to contain an in-frame N-terminal FLAG epitope tag to create the vector pDTNF (plasmid Destination Tet. N-terminal Flag Tag). The *dosR* gene was transferred from the appropriate stock in an entry clone library (PFGRC, contracted by the NIAID) in to pDTNF to create the plasmid, pEXNF-3133c (EXpression N-terminal Flag Tag). Successful recombination was confirmed by sequencing (data not shown). The resulting ATc-inducible DosR complement strain, H37Rv:Δ*dosR*::*pEXNF-3133c*, was cultured in Middlebrook 7H9 with the ADC supplement (Difco),0.05% Tween80, and 50 μg/mL hygromycin B at 37° C with constant agitation. All experiments were performed under aerobic conditions and growth was monitored by OD600. Expression of pEXNF-3133c was induced using an ATc concentration of 10 ng/mL or 100ng/mL culture. For uninduced controls, an equivalent volume of sterile DMSO was added to cultures. ATc-dependent expression of DosR was confirmed by α-DosR/α-FLAG western blot (data not shown) as well as by microarray transcriptional profiling.

RNA Preparation

RNA was isolated as described previously [54]. Briefly, pellets in Trizol were transferred to a tube containing Lysing Matrix B (QBiogene, Inc.), and vigorously shaken at max speed for 30 seconds in a FastPrep 120 homogenizer (Qbiogene) three times, with cooling on ice between steps. This mixture was centrifuged at max speed for 1 minute and the supernatant was transferred to a tube containing 300 μL

chloroform and Heavy Phase Lock Gel (Eppendorf North America, Inc.), inverted for two minutes, and centrifuged at max speed for five minutes. The aqueous phase was then precipitated with 300 µL isopropanol and 300 µL high salt solution (0.8M Na citrate, 1.2M NaCl). RNA was purified using an RNeasy kit following manufacturer's recommendations (Qiagen). Total RNA yield was quantified using a Nanodrop (Thermo Scientific).

Microarray analysis

RNA was converted to Cy dye-labeled cDNA probes as described previously [54]. For all experiments described here, 1 µg of total RNA was used to generate probes. Sets of fluorescent probes were then hybridized to custom NimbleGen tiling arrays consisting of 135,000 probes spaced at ~100 bp intervals around the *M. tuberculosis* H37Rv genome (NCBI Geo Accession #: GPL14896). Three biological replicate experiments of both induced and uninduced cultures were hybridized to arrays. Arrays were scanned and spots were quantified using Genepix 4000B scanner with GenePix 6.0 software. These data were exported to NimbleScan for mask alignment, and ArrayStar for robust multichip average (RMA)[81] normalization and statistical analysis (NCBI GEO Accession #: GSE33752). Altered gene expression was considered significant if it produced a moderated t-test $P < 0.05$ after Benjamini Hochberg multiple testing correction.

Results

Aerobic ectopic induction of DosR results in expression of the DosR regulon.

The DosR regulon was originally described as those *M. tuberculosis* genes induced following 2 hours of hypoxic gas treatment (0.2% O_2) [38], a condition thought to mimic aspects of the granuloma during latent infection that results in bacteriostasis *in vitro*. Because expression of the DosR regulon represents the first broad transcriptional adaptation to hypoxia, it has been thought to play a critical role in driving a survival adaptation during this stress condition. To determine the growth-rate implications of aerobic DosR expression, we used a *ΔdosR M. tuberculosis* genetic background previously

generated in our lab [40], and created a complement strain in which DosR could be conditionally expressed with the addition of anhydrotetracycline (ATc). We first established if the addition of ATc was sufficient to induce the DosR regulon in rolling culture. Early log phase cultures were diluted to OD600 of 0.04 and incubated independently for 24 hours prior to chemical induction of DosR. At T0 cultures were treated with 10ng ATc (dissolved in DMSO) per mL of culture, or a volume-matched amount of sterile DMSO, and RNA was collected at 12 and 24 hours post-induction. Three biological replicates of induced and three uninduced samples were processed and hybridized to genome-wide tiling microarrays. Focused transcriptional changes were apparent at 12 hours (data not shown), and by 24 hours of ectopic DosR induction under aerobic conditions, 50 genes were significantly induced, including 44 genes of the traditionally-defined DosR regulon (**Figure 4-1)**. Interestingly, these genes were induced under conditions in which neither of the histidine kinases known to interact with DosR have documented activity – log phase aerobic growth. Two DosR regulon genes, Rv3126c (conserved hypothetical) and Rv3132c (DosS) approached but did not reach statistical criteria for significant induction. Rv3841, a gene noted previously to be mildly induced during early hypoxia and considered part of the DosR regulon but without a canonical DosR binding motif [40], was moderately repressed under these conditions, raising the possibility this gene is not a member of the DosR regulon, but rather is correlatively induced in hypoxia by another factor not under the control of DosR. Along with ~90% of the DosR regulon, 6 additional genes (Rv0085, Rv0086, Rv0087, Rv1519, Rv1735c, and Rv2386c) were found to be significantly upregulated in response to this treatment. With the exceptions of Rv1519 and Rv2386c, all of these genes are located in operons with, or adjacent to, genes of the DosR regulon with known DosR binding sites.

To examine the ectopic response more closely we next asked whether the aerobic induction of DosR and the DosR regulon produced an overall transcriptional landscape similar to that of bacilli after 2 hours of hypoxic gas treatment. To do this, we performed a meta-analysis comparing the

aerobic/ectopically induced DosR regulon expression profile to the earlier reported DosR transcriptional data [40]. We found a significant correlation between the expression levels observed under ectopic/aerobic conditions and native DosR/DosR regulon expression under hypoxic conditions (Spearman r = 0.672, P<0.0001). We thus conclude that ectopic/chemical induction of Rv3133c from the ATc-inducible plasmid pEXNF-3133c under aerobic conditions results in expression of the DosR regulon comparable to that of the initial hypoxic response.

Ectopic induction of the DosR regulon does not alter growth kinetics of *M. tuberculosis*

Having established conditions in which DosR regulon induction is uncoupled from its traditional signals, we next sought to determine the impact on the growth kinetics of replicating *M. tuberculosis*. Using the induction conditions described for transcriptional profiling above, **Figure 4-2A** shows a growth curve in which expression of the DosR regulon was induced in early log phase (T0). Over the entire time course it is apparent that DosR regulon expression had no impact on the doubling time or entrance into stationary phase when compared with the uninduced control. To investigate if replication rate was more sensitive to DosR regulon induction at different growth phases, we also assessed the impact of chemical DosR induction during mid-log phase. Using bacteria at OD600 of 0.4 and ten times more ATc (100 ng/mL) than before, growth rate was again unaffected (**Figure 4-2B**). We conclude that induction of the DosR regulon does not by itself establish a bacteriostatic phenotype, or alter *M. tuberculosis* growth kinetics.

Discussion

In this work we investigated the impact of DosR regulon expression on replicating *M. tuberculosis* under aerobic conditions. Utilizing an *M. tuberculosis* H37Rv *dosR* knockout strain complemented with an ATc-inducible copy of *dosR*, we found that ectopic expression of DosR under aerobic conditions resulted in a transcriptional pattern strikingly similar to that found when wild-type *M. tuberculosis* is exposed to 0.2% oxygen for 2 hours – a condition that results in bacterial growth arrest.

However, expression of the DosR regulon to these near-physiological levels had no effect on growth rates after induction compared to DosR-uninduced cultures. Over a period of several days, these cultures replicated with nearly identical kinetics and displayed no differences as they entered stationary phase. Previous reports indicate that disruption of DosR can produce a range of hypoxic survival phenotypes – from ~1.5 log decrease in *M. tuberculosis* viability [54] to a more profound ~4 log drop in survival for *M. bovis BCG* under long-term hypoxic conditions [53]. However, we have demonstrated that deletion of DosR had very little impact on the long-term transcriptional adaptation of *M. tuberculosis* to hypoxic environments [54]. Thus, the precise role of DosR regulon induction remains an open question, but the data reported here lead us to conclude that it is not sufficient to initiate bacteriostasis.

Results reported here differ from a study performed using *M. bovis BCG* in which a constitutively-expressed merodiploid copy of DosR induced fewer members of the DosR regulon, and resulted in a moderate growth defect [174]. Perhaps the difference in species used accounts for this discrepancy. In addition, we cannot exclude the possibility that DosR-dependent growth arrest requires precise induction of a particular “native” transcriptional profile not reproduced in the ectopic system described here. However, as noted above, the transcriptome generated by ectopic DosR expression is remarkably similar to that produced after 2 hours of hypoxia, and those few genes demonstrating the greatest expression differences in this system do not have functions clearly associated with growth arrest.

The observation that *M. tuberculosis* replication is not affected by expression of the DosR regulon does not preclude the possibility that genes from this regulon are expressed during periods of clinically latent disease. This scenario is consistent with recent hypotheses that the term "latent infection" is a broad classifier describing a number of different states [91, 175]; however, that these genes are expressed during active replication strongly argues that DosR regulon expression is not an

explicit marker of latent infection. Indeed, it appears that expression of the DosR regulon correlates with multiple stages of infection and disease in a dynamic host environment.

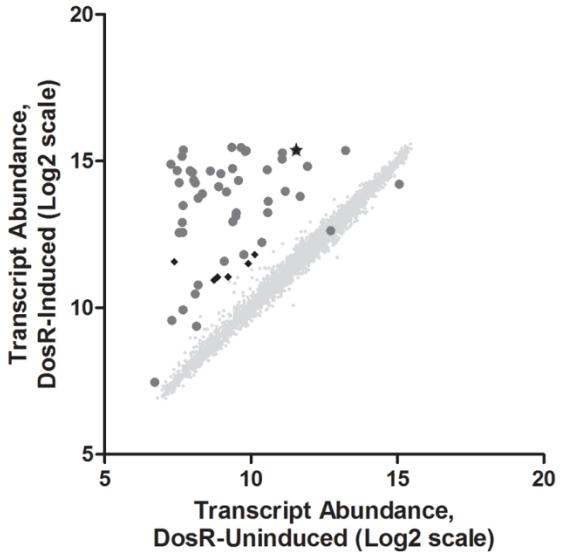


Figure 4-1: Ectopic expression of DosR induces the DosR regulon. Scatterplot displaying transcript levels of all *M. tuberculosis* genes after 24 hours of treatment with either 10 ng/mL Atc (induced) or an equivalent volume of sterile DMSO (uninduced). Three biological replicates were RMA-normalized and the median pixel intensity data are plotted on a $\log_2$ scale. Genes of the DosR regulon are represented as dark gray circles. Significantly induced genes (moderated t-test with Benjamini-Hochberg FDR correction, $p<0.05$) not part of the DosR regulon are presented as black diamonds, and the *dosR* transcript is indicated with a star.

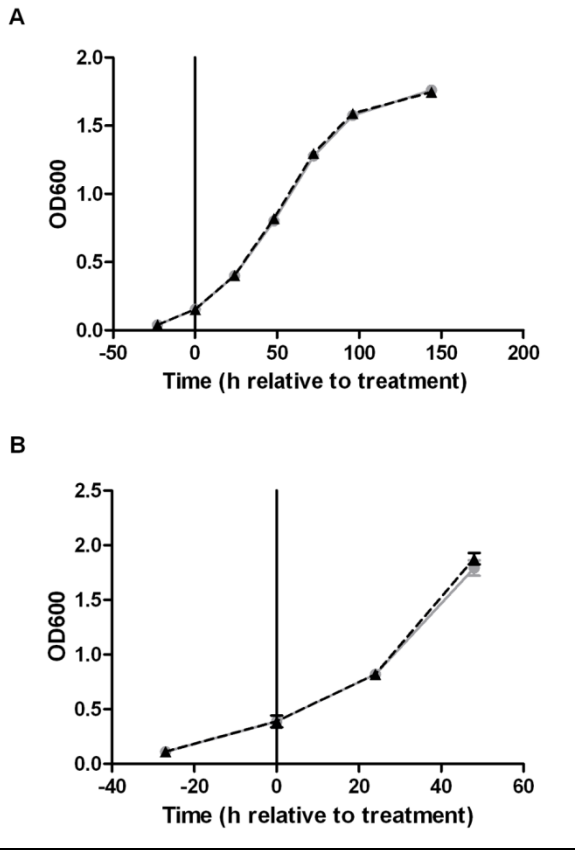


Figure 4-2: DosR regulon expression does not alter M. tuberculosis growth kinetics. **A**) Growth curves of cultures in which DosR was either ectopically induced with 10ng/mL ATc (gray circles connected by solid line) or treated with an equivalent volume sterile DMSO (uninduced, black triangles connected by dashed line). OD600 of three biological replicates were tracked for 6 days following chemical treatment. Displayed are mean OD600 +/- standard deviation. Doubling times for induced (21.97 hours) and uninduced (21.54 hours) cultures were calculated using exponential growth equation from nonlinear regression fit of un-/induced OD600 data points. Time points included in doubling time calculation were T0, T24, and T48. **B**) Growth curves of *M. tuberculosis* in which DosR was either ectopically induced with 100ng/mL ATc or treated with an equivalent volume of sterile DMSO. OD600 of three biological replicates were tracked for 48 hours following chemical treatment. Strain identifiers as described in 4-2A. Doubling times for induced (21.19 hours) and uninduced (20.33 hours) cultures calculated as above.

Chapter 5 - Global Analysis of mRNA Stability in *Mycobacterium tuberculosis*

The following text is from a manuscript submitted to the journal **Nucleic Acids Research:** Rustad, T.R., Minch, K.J., Rustad, T.R., Brabant, W., Winkler, J., and Sherman D.R. (2012) Global analysis of mRNA stability in *Mycobacterium tuberculosis*. At the time of this dissertation submission, requests for revisions from reviewers have been addressed. In the present volume figure numbers have been updated to conform to the formatting of this dissertation; however, the remainder of the text is as submitted. As in previous chapters of this dissertation, identification of the supplementary figure has taken the form "Figure S5-1."

Abstract

Mycobacterium tuberculosis (MTB) is a highly successful pathogen that infects over a billion people. As with most organisms, MTB adapts to stress by modifying its transcriptional profile. Remodeling of the transcriptome requires both altering the transcription rate and clearing away the existing mRNA through degradation, a process that can be directly regulated in response to stress. To understand better how MTB adapts to the harsh environs of the human host, we performed a global survey of the decay rates of MTB mRNA transcripts. Decay rates were measured for 2139 of the ~4000 MTB genes, which displayed an average half-life of 9.5 minutes. This is nearly twice the average mRNA half-life of other prokaryotic organisms where these measurements have been made. The transcriptome was stabilized further in response to lowered temperature and hypoxic stress, but not in response to starvation. The generally stable transcriptome described here, and the additional stabilization in response to physiologically relevant stresses, has far-ranging implications for how this pathogen is able to adapt in its human host.

Introduction

Despite more than 100 years of active research, the most widely-used vaccine in human history, and effective chemotherapeutics, tuberculosis (TB) continues to be a major global health problem leading to about 1.5 million deaths each year (http://www.who.int/tb/country/en/). *Mycobacterium tuberculosis* (MTB) can adapt to a variety of microenvironments within the host, survive a wide range of host defenses, and persist in the face of an extended multidrug regimen [176]. As in other organisms, transcriptional remodeling is a primary means of adaptation, which is achieved by changing the balance between new transcription and degradation of existing transcripts. Though transcriptional regulation is better understood, modulating mRNA stability is an alternate and complementary mechanism of regulating transcript levels. Messenger RNA half-life is a key variable needed to build models of transcriptional flux, and for interpreting physiologic changes in response to stress.

Prokaryotic mRNA degradation has been comparatively well studied in *Escherichia coli*, *Bacillus subtilis* and *Staphylococcus aureus*. In these studies rifampicin was used to interfere with the RNA polymerase beta subunit, allowing mRNA degradation to be measured in the absence of new transcriptional initiation. Using this method, 80% of mRNA transcripts in *E. coli* had measured half-lives between 3 and 8 minutes [177]. Similarly, in *B. subtilis* and *S. aureus* most mRNAs had half-lives of less than five minutes [178, 179]. The primary regulator of mRNA stability in *E. coli* in non-stress conditions is the endoribonuclease RNaseE, which sits at the center of a multi-protein mRNA degradasome complex [180]. *B. subtilis* and *S. aureus* do not have a direct homolog of RNaseE, but they do have RNases with similar function [178]. In addition to the degradasome there are at least three other mechanisms by which prokaryotic RNA stability is regulated: accessory endoribonucleases (including toxin-antitoxin modules), exoribonucleases, and interactions with RNA structure-modifying molecules (including pyrophosphatase and small non-coding RNAs) [178].

In response to stress, mRNA stability can be modulated both globally and specifically. Cold shock can extend the half-life of cold shock protein A (*cspA*) mRNA in *E. coli* from 10 seconds to hours [181, 182]. In *S. aureus* mRNA transcripts are globally stabilized in response to stationary phase as well as heat, cold, acid, and alkaline stress. For example, the percentage of transcripts with half-lives longer than 2.5 minutes in *S. aureus* goes from only 15% of all transcripts to over 50% in response to cold shock [183].

Very little is known about mRNA stability in mycobacteria. Unlike the Gram-positive organisms described above, MTB has a homolog of *E. coli* RNaseE, though it appears to associate with a different set of accessory proteins [184]. The catalytic domain of the MTB RNaseE cleaves a smaller subset of A/U rich sequences compared to the *E. coli* protein, though at a similar catalytic rate [185]. In *M. smegmatis*, a rapidly growing non-pathogenic relative of MTB, secondary structure in the 5′ UTR has been predicted to have a stabilizing effect on mRNA transcripts [186].

We present here a global survey of the mRNA decay rates in MTB. We find that the mRNA pool in MTB is very stable compared to other prokaryotes, with an average mRNA half-life over nine minutes. We also describe a global stabilization of MTB mRNA in response to both hypoxia and low temperature. This slow mRNA turnover places significant limitations on how MTB can remodel its transcriptome, and suggest that this pathogen may require novel adaptive mechanisms to degrade condition-inappropriate transcripts to adapt to commonly seen environmental stresses.

Materials and methods

Bacterial strains, media, and handling

All experiments used either *Mycobacterium tuberculosis* H37Rv (ATCC 27294) or *M. smegmatis* MC^2155, except for preliminary experiments to define relevant rifampicin concentrations, which used *Mycobacterium bovis* (BCG) Paris (ATCC#19015). The *dosR* induction experiments were done using a

strain carrying an episomal vector with a tet-inducible promoter driving the expression of a FLAG tagged *dosR* (Chapter 4 of this dissertation, [187]). For aerobic growth, cultures were grown in Middlebrook 7H9 with the ADC supplement (Difco) and 0.05% Tween80, at 37° C with constant agitation in a BioFlo110 fermentor.

Hypoxic experiments were done as described previously [54]. Briefly, a log phase culture was exposed to a constant flow of 0.2 cubic feet per hour of nitrogen with trace amounts of oxygen (2000 ppm) to create a hypoxic non-replicating culture. Low temperature experiments were performed in a rolling culture kept at 20° C for one hour before addition of rifampicin.

Transcription arrest, mRNA decay, and RNA isolation

To measure the mRNA degradation rate, transcription was stopped by addition of 50 mg/L rifampicin (Sigma). After arrest, samples were rapidly chilled in a dry ice/ethanol bath to inhibit further RNA degradation and kept at 4°C until the experiment was complete. Samples were pelleted for five minutes at 4750 x *g*, 4°C. Supernatant was discarded and pellets were resuspended in 1 mL TRIzol (Invitrogen), arresting all further degradation. Each experiment represents the average of at least three biological replicates.

RNA was isolated as described previously [188]. Briefly, pellets in TRIzol were transferred to a tube containing Lysing Matrix B (QBiogene, Inc.), and vigorously shaken at max speed for 30 sec in a FastPrep 120 homogenizer (Qbiogene) three times, with cooling on ice between steps. This mixture was centrifuged at maximum speed for 1 min and the supernatant was transferred to a tube containing 300 µL chloroform and Heavy Phase Lock Gel (Eppendorf North America, Inc.), inverted for one minute, and centrifuged at max speed for five minutes. The aqueous phase was then precipitated with 300 µL isopropanol and 300 µL high salt solution (0.8M Na citrate, 1.2M NaCl). RNA was purified using an RNeasy kit following manufacturer's recommendations, including an on-column DNase step (Qiagen). Total RNA yield was quantified using a Nanodrop (Thermo Scientific).

Microarray analysis

RNA was converted to Cy dye-labeled cDNA probes as described previously [54]. For all experiments described here, 3 μg of total RNA was used to generate probes. Sets of fluorescent probes (Cy3 and Cy5) were hybridized to custom 12x135K microarrays from Roche-Nimblegen. Each slide contains 12 identical arrays of 135,000 60-mer probes that tile most of the MTB genome with gaps of less than 200 bp between probes (average gap = 30 base pairs). Included in each array are 30,000 randomer probes of matched GC content. These randomer probes do not specifically hybridize with MTB RNA, and serve as a robust measure of background for each array. Our array design is publically available (Nimblegen ID 110405, NCBI GEO # GPL14824). A similar tiled array was used for *M. smegmatis* expression analysis (Nimblegen ID 110930, NCBI GEO #GPL15323). All array data can be downloaded from NCBI GEO (GSE36345).

Arrays were scanned using a fixed PMT gain on a GenePix 4000B scanner (Molecular Devices) at a 5 micron resolution using GenePix 4.1 image acquisition software. Each image was then burst into 24 individual arrays from each slide, and spot intensities calculated using the NimbleScan software (Roche-Nimblegen). Final analysis was done using ArrayStar (DNAstar) to collapse the probes covering each gene or inter-genic feature into a single value by averaging the probes tiled over each gene.

Loading equal amounts of total RNA allowed us to roughly normalize to rRNA, which is by far the most abundant species of RNA, with a half-life of >24 hours [189]. Standard microarray normalization methods depend on the assumption that the net fluorescence intensity in each channel of the microarray is equal, an invalid assumption in these experiments. Therefore the mRNA features of the microarray were normalized to the stable tRNA features. A set of 30,000 randomer probes was used to define background, and only genes that had overall intensity at least four-fold above background at T=0 were used for calculating ½ life. Data were plotted as mean array intensity ($\log_2$) over time, where the negative reciprocal of the slope of each line equals the half-life of that transcript. Transcripts whose

degradation did not fit a standard logarithmic degradation profile (R^2<0.8) were also excluded from calculations of average half-life. More than 50% of MTB genes met inclusion criteria and were included in average half-life determinations.

qRT-PCR

Quantitative real-time PCR was used to confirm the microarray results for a selected subset of transcripts. For each sample 20 ng of DNase-treated total RNA was converted to cDNA using SuperScriptIII (Invitrogen) and purified using a QiaQuick column (Qiagen). qRT-PCR was done using a BioRad CFX384 as directed by the manufacturer. In brief, 1/20 of the cDNA reaction was added to 12 wells of a 96-well plate along with 5 µl of Universal Master Mix (ABI) and 2 µl of water. Primers and probes specific to genes representing short, long, and average half-lives were then added to each well. Amplification conditions followed manufacturer recommendations.

Results

Global mRNA decay rate of the MTB transcriptome

To measure the global rate of mRNA decay in MTB, transcription initiation was arrested using a high concentration (~250X the minimum inhibitory concentration) of the transcriptional inhibitor rifampicin, which binds to the beta subunit of RNA polymerase [190], leaving degradation as the primary factor affecting transcript abundance. Degradation rates were calculated from the slope of a plot of $\log_2$ transcript abundance over time after transcriptional arrest.

The degradation rate calculation depends on the assumption that transcription is completely inhibited at the concentration of rifampicin used. To test this we used a strain of *M. bovis* BCG carrying a luciferase reporter fused to the hypoxia-responsive alpha-crystallin gene promoter [40]. This strain produces strong luciferase activity within one hour of exposure to hypoxic conditions (unpublished data). We treated a log phase culture of this strain with 50 µg/ml of rifampicin and measured the

inhibition of luciferase induction after one hour of hypoxia relative to an untreated control. The sample given 50 µg/ml rifampicin produced minimal luciferase activity (<3% of untreated) that did not increase with time (**Supplementary Figure S5-1**). Based on these results we used 50 µg/ml of rifampicin to stop new transcription in subsequent experiments.

To assess the mRNA half-life (T1/2) globally, we measured mRNA decay using tiled genomic microarrays comprised of 100,000 60-mers, or roughly 75% of the entire genome. Once rifampicin stopped transcription in a log phase MTB culture, aliquots were removed at regular intervals. RNA was isolated from these samples, and equal amounts of total RNA (mostly stable rRNA) were converted to fluorescently labeled cDNA and hybridized to microarrays. To insure that decay over at least two half-lives could be measured, transcripts that were not at least four-fold above background were excluded from analysis. Transcripts whose degradation pattern did not fit an exponential decay curve were also excluded, as exponential decay is implicitly assumed in a half-life calculation. After filtering the data this way we were able to measure the half-lives of 2139 transcripts of the roughly 4000 genes in the MTB genome. There was good correlation between four replicate experiments: on average the standard deviation was less than one minute and the coefficient of variance was <0.1. Excluded transcripts were not enriched for any functional category with the exception of the PE/PPE family of genes, for which only 33 of the 168 genes could be measured. The mean half-life of log phase mRNA transcripts in MTB is 9.5 minutes (**Figure 5-1**), substantially longer than the global half-life of other prokaryotes studied to date [178]. Over 80 percent of mRNA transcripts had half-lives between 8 and 12 minutes, and even the shortest half-lives in MTB were longer than the average transcript in *E. coli* [177]. Quantitative real-time PCR on a subset of 89 genes produced similar results (data not shown).

The doubling time of MTB is 16 to 22 hours, much longer than most well-studied prokaryotes. To see if replication rate and mRNA turnover may be related, we measured the degradation rate using methods described above for the non-pathogenic mycobacterium *M. smegmatis*, which has a much

shorter doubling time (~3-4 hours). The average half-life in *M. smegmatis* was 5.2 minutes, much closer to the mRNA decay rate described for other bacteria (**Figure 5-1**), raising the possibility that stable mRNA and slow growth may be correlated in mycobacteria.

We also examined the impact of position within an operon on mRNA stability. We compared opposite ends of polycistronic transcripts where both ends of the transcript had measured half-lives (352 of all 889 polycistronic transcripts). About 90% of intra-operon half-lives were within 2 minutes of each other (**Figure 5-2**). This pattern held true for transcripts with short, average, or long half-lives. In addition, the half-lives of genes at the 3' or 5' end of an operon did not show any consistent trend toward being more labile. This supports a model of RNA decay wherein the destabilization of a transcript, typically cleavage by an endonuclease, is the rate limiting step followed by rapid digestion by exonucleases [178].

Gene functional class and degradation rate.

In MTB three of the nine functional categories annotated by Cole et al. [77] have stabilities significantly different ($p < 0.001$) from the general pool of genes (**Table 5-1**). Transcripts of genes annotated as being involved in information systems, including ribosomal proteins, have significantly shorter half-lives over all. The PE/PPE family of genes that are specific to the MTB complex and have a common proline-glutamic acid repeat motif produce mRNAs that are generally more stable, though as mentioned above this category had fewer than expected measurable half-lives. Genes transcribed from mobile DNA elements of the insertion sequence and phage category are also significantly more stable than average. As discussed below, this may be related to the relatively low level transcription of these genes during log phase growth. The differentially stable functional categories were very different than those in *E. coli* [177]. For example, the ribosomal protein genes produce transcripts that are among the shortest half-lives in MTB, but the analogous transcripts have average half-lives in *E. coli*. As the functional categories in the original MTB annotation are very broad, we turned to a more recent re-

annotation publically available on TBDB (www.tbdb.org) to identify specific sub-categories and pathways with very stable or labile transcript stabilities [191]. We compared the 100 most labile and 100 most stable transcripts to the genome as a whole (**Table 5-2**). Among the transcripts with the shortest half-life are many 'housekeeping' genes including ribosomal proteins, subunits of ATP synthetase, the Clp proteases, as well as the principal log phase sigma factor, *sigA*. More stable transcripts are more diverse, but include a disproportionate number of genes involved in replication, recombination and repair, and in amino acid transport and metabolism.

Other factors affecting decay rate.

We compared the mRNA ½ life of each transcript to several factors that could contribute to transcript stability. For example, we compared decay rate to transcript length. A plot of the degradation rate of single gene operons compared to the length of the gene encoded did not show any correlation (**Figure 5-3A**). As with the analysis of polycistronic operons, a poor correlation between transcript length and decay rate supports a model in which destabilization, perhaps through cleavage by an endonuclease, is the rate limiting step in mRNA degradation. Similarly, a plot of the GC content of each ORF compared to mRNA decay showed no significant correlation (**Figure 5-3B**, $R^2<0.1$). The set of ~20 relatively AT-rich transcripts (%GC<57) have significantly shorter transcript half-lives compared to the average (7.8 vs 9.5 minutes, $p<0.001$) but there are too few AT-rich genes to affect the general pattern. The lack of a correlation between global mRNA stability and broad physical characteristics of messages echoes the pattern in *E. coli* [177].

The transcript characteristic that most accurately predicted half-life was abundance of that message at the time transcription was arrested (**Figure 5-3C**). A plot of abundance at time zero compared to half-life shows a strong inverse correlation that follows a power trend line ($R^2>0.8$). The 100 genes that are most abundantly transcribed have an average half-life of 6.7 minutes, and only one of these genes has a half-life above average. This correlation suggests two possibilities: either

transcripts highly expressed during log phase growth are inherently more labile or the half-life of a transcript depends to a large degree on the level of expression.

To explore this question, we generated an MTB strain in which a subset of genes could be ectopically induced. The MTB transcriptional regulator DosR controls induction of several genes in response to reduced oxygen tension and other stimuli [38, 40]. Using a copy of *dosR* under the control of a tetracycline-responsive promoter, we induced expression of *dosR* and concomitantly upregulated the DosR regulon (chapter 4 of this dissertation, [187]). With and without induction of the DosR regulon we measured half-lives for all transcripts as above. Most transcripts showed little change in expression or half-life, however the DosR-regulated genes were all highly induced and nearly all showed a corresponding drop in the measured half-life (**Figure 5-3D**). Of the 17 genes for which half-lives could be measured, 15 had a shorter half-life after induction, on average shorter by seven minutes.

Modest cold shock substantially stabilizes MTB mRNA

During transmission between hosts MTB can be exposed to a drop in temperature, a stress that is generally associated with some mRNA stabilization. To test the impact of temperature shifts on mRNA stability, we compared the mRNA degradation profile of cultures maintained at room temperature (20° C) to those observed at 37° C. Decay rates were generated as described above. As expected, the transcripts in the 37°C culture degraded normally; however no measurable mRNA degradation occurred during the first two hours of transcriptional arrest in the culture kept at 20°C (**Figure 5-4**). Five hours after arresting transcription, only 55 transcripts had decayed to half their starting intensity, indicating an average mRNA ½ life of >5 hours for MTB transcripts at room temperature (data not shown).

mRNA degradation in MTB is oxygen dependent

In response to hypoxia, MTB undergoes a major remodeling of its transcriptome. The transcriptional regulation of the initial response to hypoxia by the two-component response regulator

DosR has been well characterized [38, 40], but we hypothesized that the stability of some transcripts may change in response to hypoxia, thereby modifying the response rate of DosR-regulated genes. To test this we exposed bacteria to one hour of hypoxia before adding rifampicin and followed the subsequent mRNA decay in absence of oxygen. One hour of exposure to hypoxic conditions (1% atmospheric O_2) is sufficient to arrest growth and induce the initial hypoxic response [54].

Rather than modifying the mRNA decay rates of individual genes, we found that exposure to hypoxia led to a global stabilization of all transcripts. An hour after rifampicin treatment of hypoxic cultures, average mRNA half-life increased to ~30 minutes for the first hour, and subsequent decay slowed to below the limits of detection (**Figure 5-4**). A biphasic mRNA degradation curve invalidates accurate half-life calculations, however after five hours of transcriptional arrest mRNA had decayed only four-fold, or two half-lives (data not shown), indicating a net half-life of >150 minutes in hypoxia. If anything, this experiment overestimates the rate of decay in hypoxia, as some experiments show even less degradation.

We tested variants of the hypoxia-stabilization experiment by following a set of ten genes by qRT-PCR using Taqman probes (**Figure 5-5**). As in the global analyses, one hour of hypoxic exposure led to substantially inhibited mRNA decay. To test if this stabilization was a transient response to the initial hypoxic stress, we measured the decay rate by arresting transcription with rifampicin after five days of hypoxia, and again observed a general stabilization of mRNA. This hypoxic stabilization was immediately reversed if oxygen was reintroduced to the culture concurrent with transcriptional arrest (**Figure 5-5**).

Both hypoxic stress and lowered temperature induce stable and viable but non-replicating states in MTB. We turned to a third stress, starvation, which also results in a stable non-replicating state [192], to see if mRNA stabilization in non-replicating conditions was a general phenomenon. Log phase MTB was pelleted, washed, and resuspended in phosphate buffered saline with no carbon or nitrogen source. After one hour of starvation, rifampicin was added to stop transcription and the degradation

rate was measured as described above. No stabilization was seen in response to starvation, suggesting that mRNA stabilization during cold and hypoxic stress is not a general feature of non-replicative states (**Figure 5-5**).

Discussion

This study provides the first global survey of mRNA stability in the human pathogen *Mycobacterium tuberculosis*, providing half-life measurements for 2139 genes. The pool of mRNA in MTB degrades with a mean half-life of 9.5 minutes, significantly slower than in other bacteria that have been studied (**Figure 5-1**). The half-life of over 1000 messages were also measured in *M. smegmatis*, a related organism with a much shorter doubling time, and the decay rate was found to be very similar to that seen in other prokaryotes, just over five minutes. This suggests that the stable mRNA and slow growth may be linked in mycobacteria. However, previously published decay rates in organisms with highly variable doubling times have not shown this correlation, suggesting that this is not a common adaptation to slow growth [177, 179, 193, 194]. A slower rate of mRNA decay means that MTB cannot repress genes as quickly as other prokaryotes through limiting transcription and allowing the native degradation machinery to remove surplus transcripts. In return for having a slower rate of adaptation, MTB can conserve energy by limiting the rate of mRNA turnover to maintain transcripts at a given level.

We are currently exploring three potential mechanisms that may lead to the generally extended mRNA half-life in MTB. The enzymes that carry out degradation could limit the rate of mRNA decay, either through slower enzyme kinetics or lower abundance compared to other bacteria. We show here that mRNA half-life shows no bias toward the 3' or 5' end of transcripts (**Figure 5-2**) nor is it correlated with the size of a transcript (**Figure 5-3A**), supporting the hypothesis that the initial destabilization is the rate limiting step. In *E. coli* this destabilization step is catalyzed by the endonuclease RNaseE. The homolog to this gene in MTB is kinetically equivalent to the orthologous enzyme in *E. coli*, but the MTB version is more specific about the sites of cleavage [184]. This specificity could directly result in a slower

global rate of mRNA decay. Alternatively, features of MTB transcripts such as stabilizing secondary structure or post-transcriptional modifications could contribute to decay rates. Although MTB has a higher GC content than other bacteria with defined half-lives, this feature does not seem to predict mRNA half-life (**Figure 5-3B**). Finally, mRNA binding proteins or the ribosome may shelter MTB mRNA from the degradation machinery. Further biochemical analysis of the mRNA decay process in MTB is necessary to define the stabilizing mechanism.

The abundance of a transcript is dependent almost exclusively on its rates of transcription and degradation. One would predict therefore that half-lives would be positively correlated with abundance. A plot of those two variables however reveals a striking inverse correlation (**Figure 5-3C**). Previously, Bernstein et al. described a weak but statistically significant negative correlation between transcript abundance and the decay rate in *E. coli* [177]. In MTB this inverse correlation is much stronger (R^2>0.8). This raises the question of whether or not the mRNA degradation rate is dependent on the level of transcript abundance, or if abundant transcripts, as a class, have shorter half-lives. Overexpression a subset of genes showed that induction led to reduced stability of those transcripts. Clearly rapid turnover of abundant transcripts requires more energy, as the rate of transcription must be elevated to compensate. Perhaps rapid cycling allows tighter regulation of very highly expressed genes. We are currently developing models of mRNA decay that include this strong inverse relationship.

We also explored the role of differential mRNA degradation in response to physiologically relevant stresses including hypoxia, starvation, and low temperature. Hypoxic stress is of interest in MTB as it is a clinically relevant host-dependent environmental stress that arrests replication and has been suggested to play a role in the bacterial adaptation from active to latent disease [91]. In response to hypoxic stress MTB undergoes drastic remodeling of its transcriptome resulting in the altered transcription of nearly a quarter of all transcripts [91, 175]. Given the size of this rearrangement, we hypothesized that some of the regulated transcripts may have altered abundance due to modified

mRNA degradation rates. We found that, rather than a small set of hypoxia regulated genes having altered half-lives, exposure to a low oxygen environment results in a global stabilization of all mRNA transcripts (**Figure 5-4**). This raises the questions of how MTB mobilizes a large remodeling of the transcriptome with minimal mRNA degradation, and by what mechanism does this occur.

The native stability of MTB mRNA is well above average for bacteria, but the increase in global stabilization of mRNA in response to both hypoxia and decreased temperature is much larger than seen in other bacteria. The decay of MTB mRNA in response to hypoxic conditions is at least fifteen-fold slower than in aerobic conditions. The decay seems to follow a biphasic curve, with an initial half-life of ~50 minutes followed by a virtual arrest of mRNA degradation. Hypoxia-triggered mRNA stabilization of a small set of transcripts has been described in *E. coli*, though the effect was less striking [181]. Given the well characterized repression of hundreds of MTB genes in response to extended hypoxic stress [54], it is clear that degradation of specific transcripts occurs in spite of global stabilization of mRNA. It is tempting to speculate that the need to degrade specific transcripts rapidly in response to stress could help explain the evolution of the remarkable number of toxin-antitoxin modules in MTB, many of which encode RNases [195], as an alternative method of mRNA degradation in conditions that globally stabilize mRNA. We are currently exploring the question of repressed transcripts in the face of global stabilization by analyzing transcripts down-regulated in hypoxia.

Bacteria in droplets coughed out the lungs of patients with active TB are often exposed to a drop in temperature, dependent on the ambient environment. The arrested decay of mRNA at room temperature described here (**Figure 5-4**) suggests that the transcriptome of these bacteria in transit between hosts is likely to stabilize rapidly in response to that change of environment. Stabilization of mRNA at room temperature also has implications for researchers performing transcriptional analysis of MTB, as the lack of signal decay considerably lessens concerns about alteration of the transcriptome after sampling. In *E. coli* a cold-specific RNA helicase, CspA, is necessary to effectively degrade mRNA at

low temperature [181, 182]. The MTB ortholog to CspA is less ordered and less stable than similar genes in other prokaryotes [196]. This suggests that MTB may be unable to unravel the secondary structure of mRNA at lower temperatures, thereby inhibiting degradation.

As systems biology approaches begin to map the regulatory networks of MTB, information about the influence of mRNA degradation on transcript levels will be needed to create valid models. The innate stability of a transcript can affect the kinetics of transcriptional response: genes with a rapid response to stress have less stable transcripts than genes involved in enduring stress responses [197]. Characterizing the mRNA decay process in MTB has uncovered a number of unusual features of this central biological process. The long native mRNA half-life of MTB, and the massive stabilization in response to hypoxia and cold, will alter our understanding of how this critical pathogen is able to adapt and respond during infection.

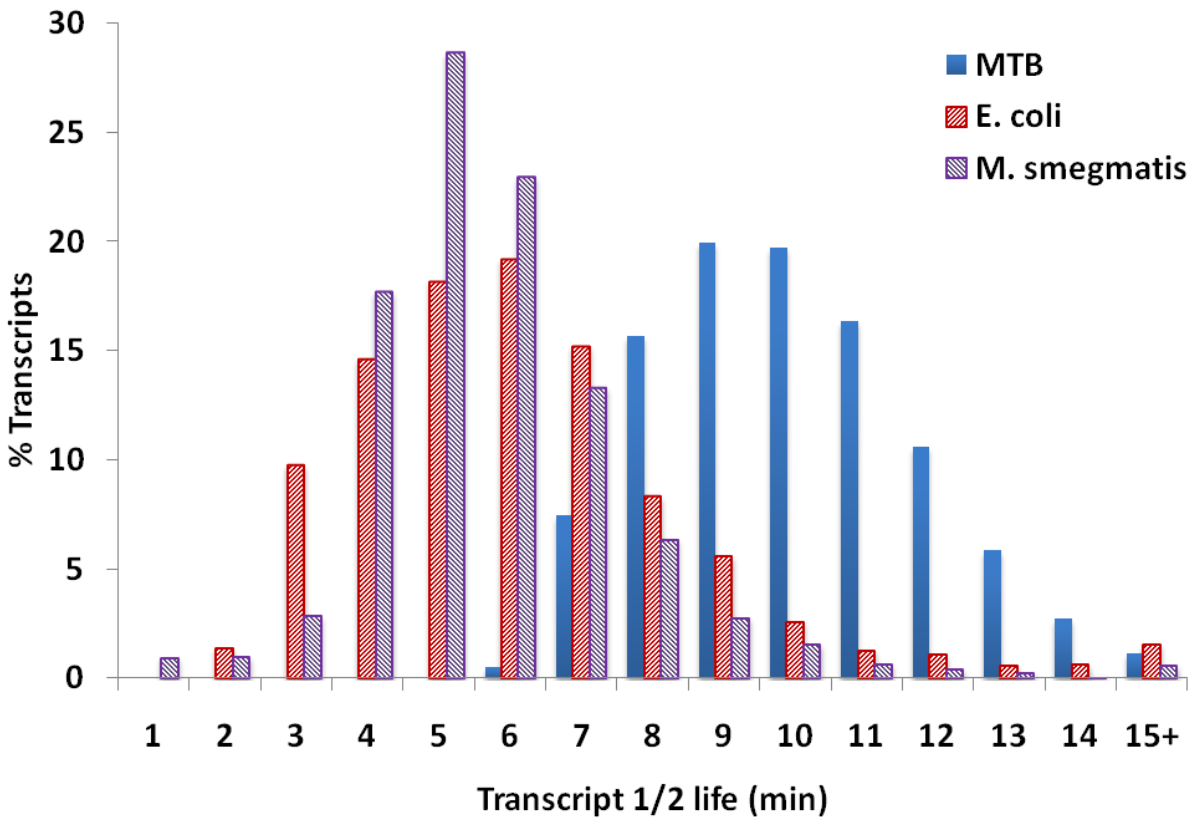


Figure 5-1: Histogram of transcript half-lives in log phase MTB. The distribution of MTB mRNA half-lives follows a normal distribution (blue), with a mean half-life of 9.5 minutes. This is substantially longer than that previously measured for *E. coli* (red, data shown from Bernstein 2002) (19). The fast growing mycobacterium *M. smegmatis* had a decay rate similar to previously described bacteria (purple). No decay rate was measured for genes that were not expressed at least four-fold above background or did not follow a logarithmic decay.

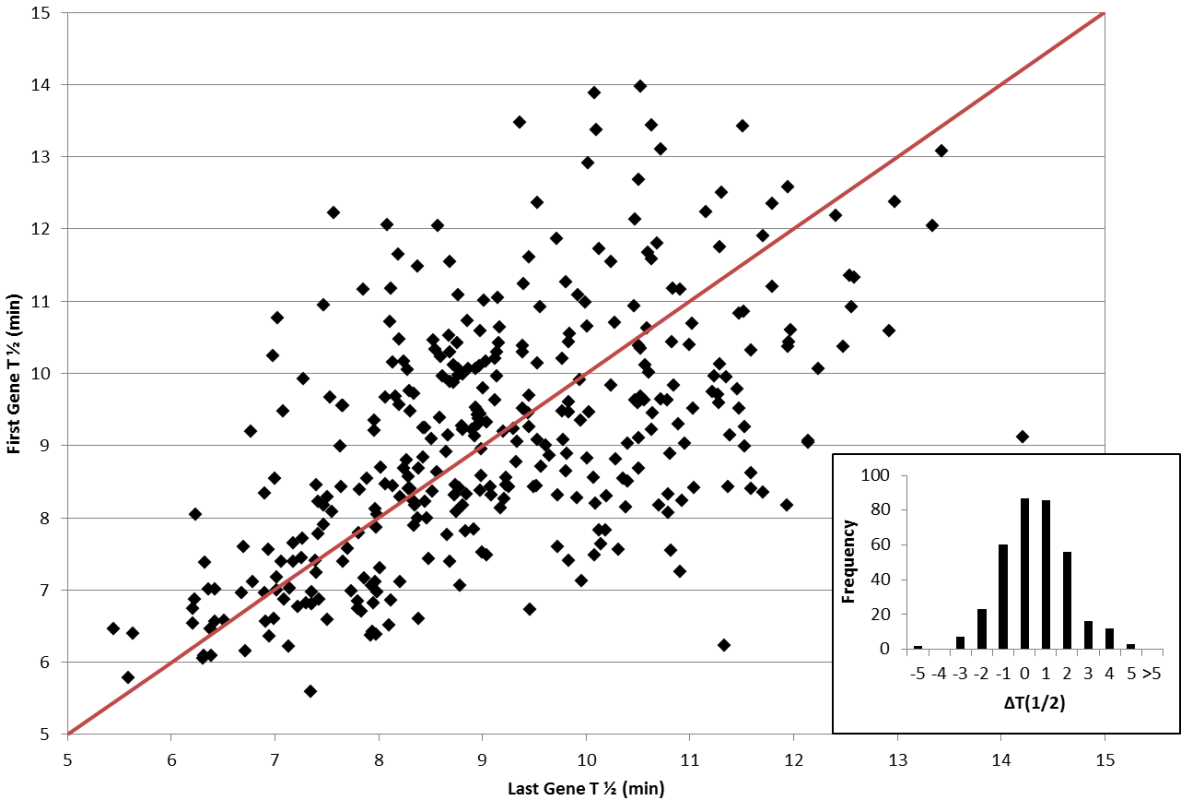


Figure 5-2: Conservation of degradation rate within operons. Half-lives of the 5' and 3' genes from operons show generally similar mRNA stabilities, with no trend toward either end of the transcript. Inset shows a histogram of the differences between first and last half-lives in minutes.

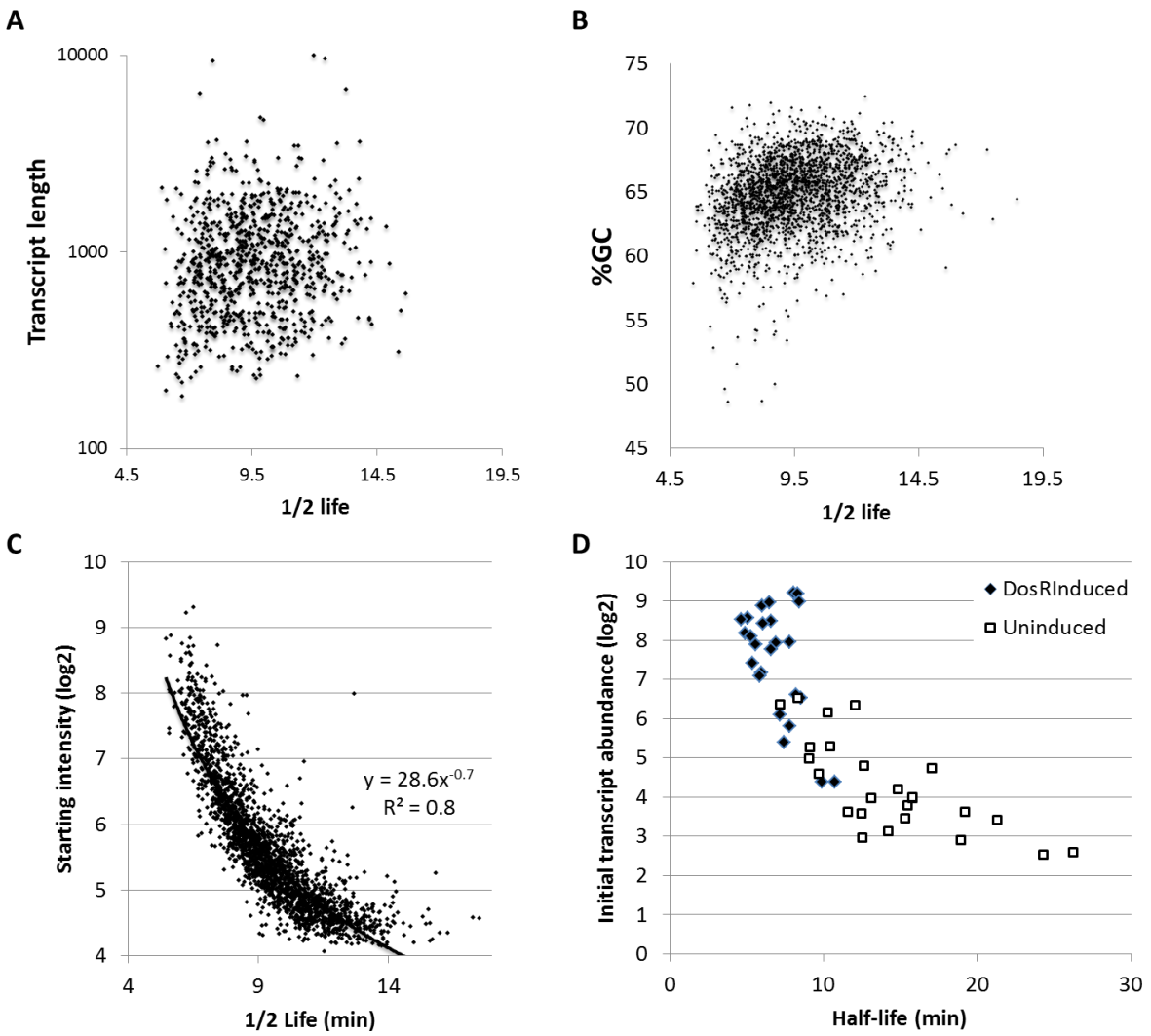


Figure 5-3: Impact of transcript attributes on mRNA half-life. Although little or no correlation is seen between mRNA stability and the length of single gene transcripts (A) or the %GC of mRNA messages (B), the abundance of a transcript at the time of transcription inhibition (C) shows a very strong negative correlation with mRNA half-life. The correlation follows a power rather than a linear regression line, with an R^2>0.8. Increased expression of the DosR regulon using a tet-inducible system in aerobic conditions results in a corresponding decrease in the mRNA half-lives of these genes (D).

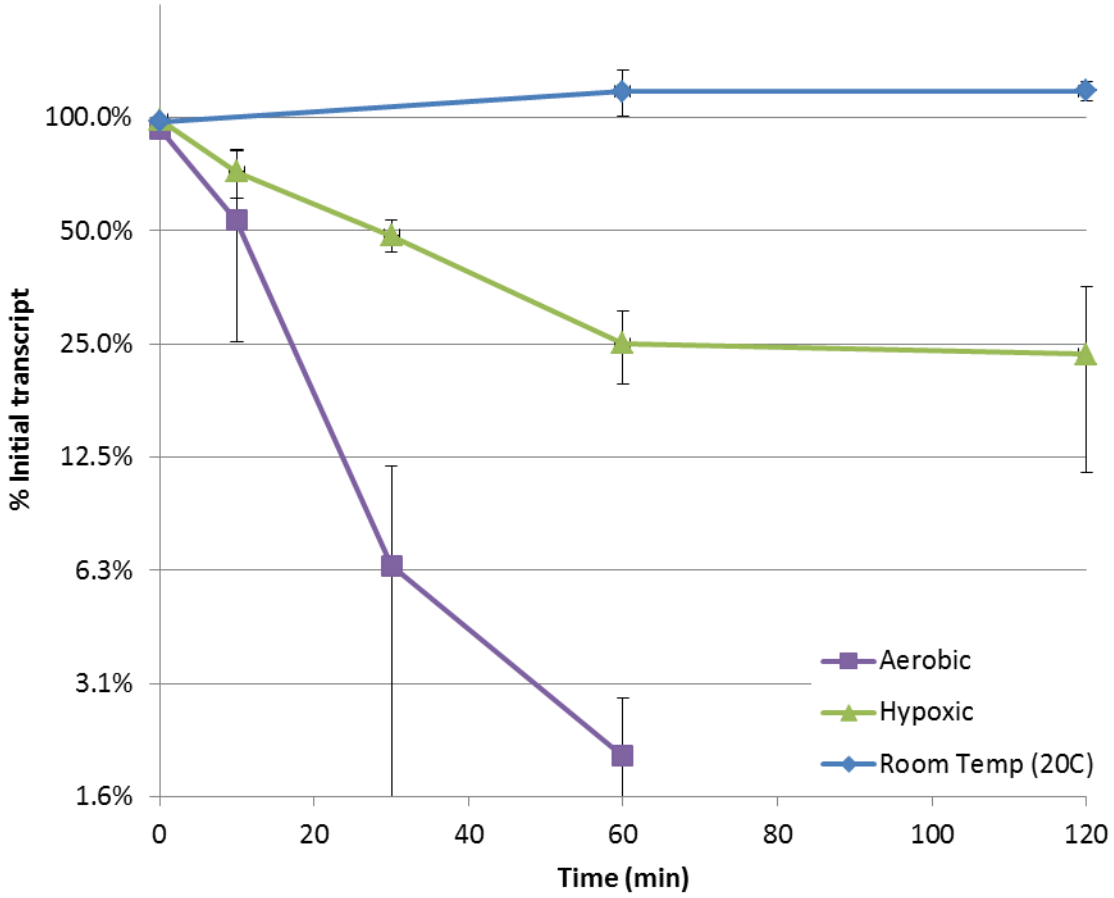


Figure 5-4: mRNA degradation during stress conditions. mRNA decay curves for MTB in aerobic log phase (squares), hypoxia (triangles), or room temperature (20°C, diamonds) measured by microarray. Shown are averages over all genes above background in four replicate experiments.

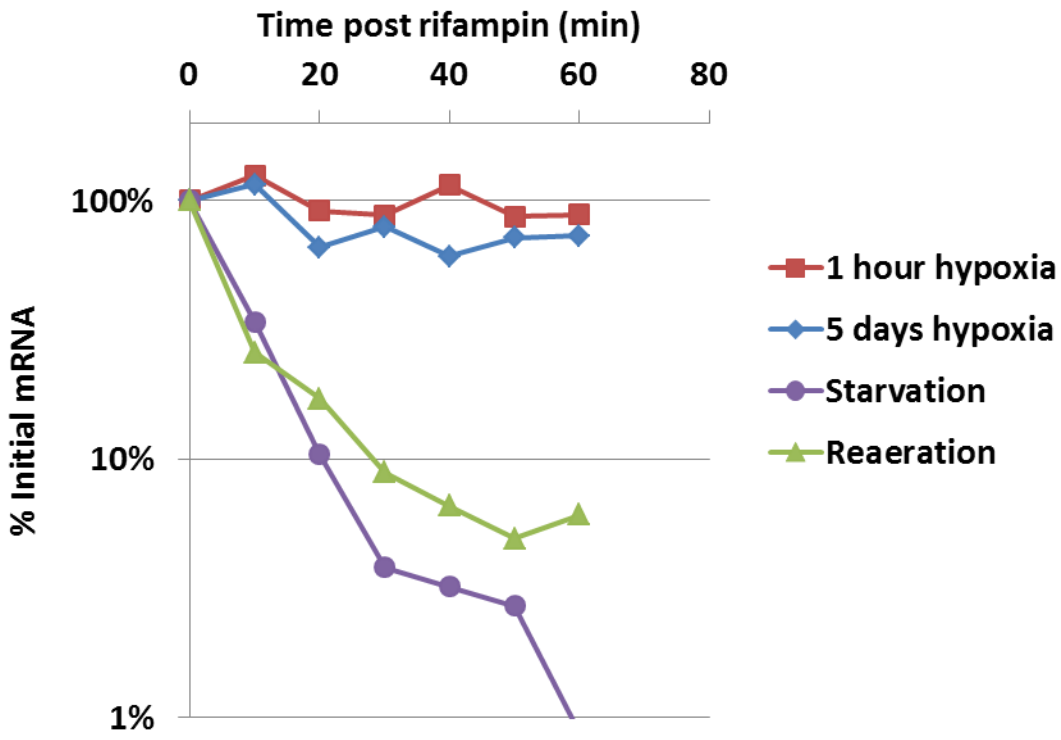


Figure 5-5: Variations of hypoxia-induced stabilization. Each line represents the average abundance of ten transcripts of average decay rates measured by QRT-PCR under a given condition. Abundance on the y-axis is shown as a percentage of starting mRNA. After exposure to one hour (red squares) or five days (blue diamonds) of growth arrest by hypoxic stress, mRNA decay is significantly retarded. The reaeration of a hypoxic culture immediately relieves inhibition of mRNA decay (green triangles). Starvation, a stress that similarly induces a non-replicating state, does not inhibit mRNA decay (purple circles).

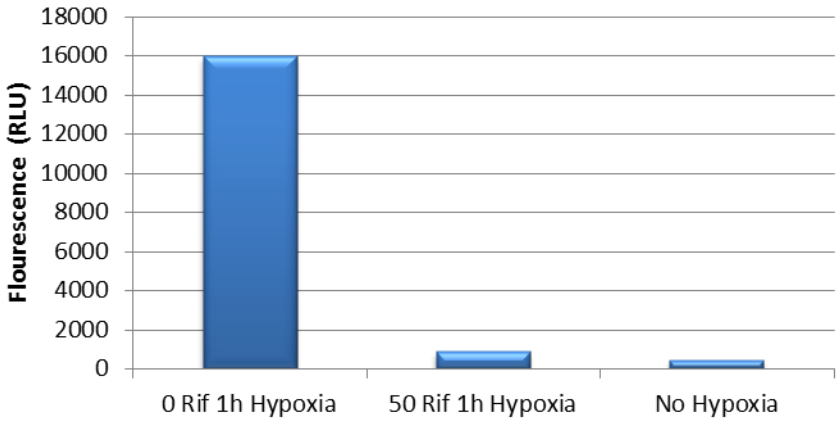


Figure S5-1: Treatment of *M. bovis* BCG with 50 µg/ml of rifampicin halts transcription. A reporter construct in which firefly luciferase was fused to the strongly hypoxia-responsive alpha-crystallin gene promoter was implemented to determine the concentration of rifampicin required to halt transcription. Under aerobic conditions there is negligible luciferase production; however, upon exposure to 1 hour of defined hypoxic environments luciferase expression is substantially induced. This induction can be ablated by the treatment of cells with 50 µg rifampicin per ml of culture prior to the initiation of hypoxic gas flow.

Functional Category	Mean half-life (minutes)
information pathways	**8.7***
virulence, detoxification, and adaptation	9.3
lipid metabolism	9.4
cell wall and cell processes	9.4
metabolism and respiration	9.5
hypothetical protein	9.6
Regulatory proteins	9.6
insertion seqs and phages	**10.4****
PE/PPE	**11.1****
*=significantly lower	
**=significantly higher	

Table 5-1: Gene functional class and degradation rate.

Functional subcategory	Specified Genes	Whole Genome	Specified Gene %	Whole Genome %	P Value
Functional Enrichment for the genes with the shortest 1/2 life					
Posttranslational modification, protein turnover, chaperones	9	94	12.8	2.3	0
Translation, ribosomal structure and biogenesis	15	121	21.4	3.0	0
Energy production and conversion	10	210	14.2	5.2	0.002
Intracellular trafficking, secretion, and vesicular transport	2	15	2.8	0.3	0.025
Functional Enrichment for the genes with the longest 1/2 life					
Replication, recombination and repair	9	194	12.8	4.8	0.004
Amino acid transport and metabolism	8	183	11.4	4.5	0.01
PE/PPE	2	11	2.8	0.2	0.014
Energy production and conversion	8	210	11.4	5.2	0.019

Table 5-2: Functional enrichment of the 100 most labile and 100 most stable transcripts vs. the whole genome.

Chapter 6 – Conclusions and Future Directions

M. tuberculosis in hypoxia: the key to persistence

M. tuberculosis resides in a dynamic niche throughout its infection cycle [175, 198]. As described in chapter 1 of this dissertation, decades of research have gone in to studying the response of *M. tuberculosis* to varying levels of oxygen. Oxygen limitation occurs during infection, and exposure to hypoxic environments *in vitro* renders bacteria viable but non-replicating [91]. Subsequent re-introduction of oxygen stimulates bacterial replication [159]. Correspondingly, we and others hypothesized that latent tuberculosis disease is in part driven by the creation and maintenance of hypoxic microenvironments. If so, then mistimed growth arrest or the illegitimate re-initiation of replication could potentially be deleterious to the bacterium as it adapts to varied environments within the host. The molecular events controlling these responses are the subject of this dissertation.

M. tuberculosis coordinates a complex transcriptional program in response to long-term hypoxia and reaeration *in vitro*, with the differential regulation of >20% of all putative transcriptional regulatory proteins in the genome [54, 91, 159]. We hypothesized that a combination of these regulators is responsible for the coordinated entrance into and exit from dormancy. The work to interrogate both the target-binding proclivities and the transcriptional impact of binding for these transcriptional regulators is described in chapters 2 and 3 of this dissertation. An intricate control logic governing the hypoxic response is evident in the data from those investigations: there is significant auto-regulation of expression, as well as hundreds of examples of regulatory cross-talk, and putative feed-forward loops. While it remains an open question if there is a programmed combination of regulatory proteins that halt replication, the TF-target identification data generated in this dissertation represent a new resource to investigate this possibility.

The two component response regulator Rv3133c/DosR was one of the first molecular correlates of hypoxic adaptation in *M. tuberculosis* [38]. This transcription factor is induced in response to a range of stimuli including hypoxia, nitric oxide and carbon monoxide [38, 40, 41, 52]. That this transcriptional regulator is involved in the initial hypoxic response is well-established; however, the significance of its role remains obscure. As described in chapter 4 of this dissertation, expression of this transcription factor and its regulon under aerobic conditions does not impact replication kinetics of logarithmically-growing *M. tuberculosis*. This, coupled with the observation that strains of the W/Beijing lineage *M. tuberculosis* constitutively express DosR [173, 199], lead us to conclude that DosR cannot be the principal *driver* of bacteriostasis *in vitro* or *in vivo*. This does not preclude the possibility that DosR regulon expression causes some level of metabolic remodeling, perhaps as part of a larger "developmental program" that conditions the cell to survive under oxygen-limiting conditions; however, given the described results the fitness cost of maintaining this poised state must be minimal.

Recent work has suggested that the flux of acetyl CoA may act as a pendulum to push *M. tuberculosis* into either an actively growing or quiescent state [200]. In this work the authors suggest that increased synthesis of triglycerides, principally through the activity of Rv3130c/triacylglycerol synthase 1 (*tgs1*), can divert carbon flow from the TCA cycle and promote the reduction of bacterial growth rate under limiting conditions. They subsequently demonstrate that the ectopic up-regulation of Rv0889c/citrate synthase II (*citA*) – which is responsible for the condensation of oxaloacetate with acetyl CoA to form citrate – can redirect carbon flow to the TCA cycle and induce continued replication even under growth-limiting conditions. Given these data, it is interesting to note that *tgs1* is strongly induced upon the induction of DosR in hypoxia and under the aerobic/ectopic conditions described in chapter 4 of this dissertation. This might explain why in some lineages of *M. tuberculosis* DosR regulon expression can be maintained in a poised state, as described above. It is possible that *dosR* and *tgs1* induction alone are not sufficient to induce bacteriostasis; however, the action of Tgs1 may siphon off a fraction of

the acetyl CoA pool from the TCA cycle even during conditions that promote rapid bacterial growth. Therefore, in the context of the normal cellular milieu, it may be that induction of triglyceride synthesis must be followed by additional stimuli from environmental conditions that promote bacteriostasis and trigger downstream regulatory switches – e.g. prolonged exposure to hypoxia and activation of the transcriptional regulators of the enduring hypoxic response.

One possible path to bacteriostasis might be to induce the synthesis of triglycerides through the action of Rv3133c/DosR *and* repress the expression of key enzymes in central metabolic pathways. The data generated in chapter 3 of this dissertation provide glimpses that this may be occurring in *M. tuberculosis*, as the induction of several transcriptional regulators results in the differential expression of TCA cycle enzymes. For example, the *M. tuberculosis* genome contains three homologs of TCA cycle enzyme succinate dehydrogenase, and the induction of Rv0757 (PhoP) causes concomitant down-regulation of two of these loci. The absence of proximal transcription factor binding sites indicates that the repressions observed are indirect, perhaps through the activity of a "lieutenant" transcription factor under the control of PhoP, and it is not yet clear how differential modulation of these loci affects enzyme function or metabolic flux. Nevertheless, from the existing data the combined induction of DosR and PhoP make attractive candidates for inducing programmed growth arrest in *M. tuberculosis*. The combinatorial increase of TAG synthesis (mediated by DosR induction) and repression of TCA cycle enzyme expression (mediated by PhoP induction) might decrease flux of acetyl CoA through this pathway and place a replicative block on the cell.

Systems biology of tuberculosis: from gene expression to phenotype

The studies described in this dissertation have bearing on three areas of systems biology research in *M. tuberculosis*: 1) characterizing genome-scale regulatory models, 2) predicting gene expression patterns, and 3) predicting phenotypes of perturbations that can be used to direct targeted

anti-tuberculosis interventions. We have made contributions to the first two points, and constructed a foundation to address the third.

In chapters 2 and 3 of this dissertation we describe the results of experimentally characterizing transcription factor binding by ChIP-seq and the concomitant transcriptional impact of that binding event by microarray. Interrogating 51 different transcription factors under a standard condition this combined method identified hundreds of TF-promoter interactions that altered expression of the target gene. Using these data we subsequently developed models of regulation for ~1400 tuberculosis genes that predicted gene expression better than the random assignment of regulators. We generalized these models by predicting gene expression in an unrelated data set from a different experimental condition: transcript measurements throughout a 14 day time course of hypoxia and reaeration. In the time course there were 2506 genes including 159 transcriptional regulators that exhibited ≥ 2-fold change in expression at one or more time points. Of the 159 regulators, we had binding and transcriptional data for 43 (27%). Using these input data we were able to generate models that accurately described the expression patterns of 838 genes (33% of all varying genes) in the hypoxic time course.

The values reported above nearly double the number of genes with described regulators [69], and the predictive power of the network far exceeds what was previously known about *M. tuberculosis* transcription regulation. However, these results describe only a fraction of the gene regulatory space and more work remains to increase our gene expression predictive power. The most straightforward approach to do this is to simply increase the number of transcription factors interrogated by the ChIP-seq/microarray method described in chapters 2 and 3. In favor of this, we have constructed a validated system for these experiments and many of the reagents are in place. If, however, the relationship between transcription factors interrogated and genes modeled remains consistent, this approach will result in only incremental increases in predictive power. Furthermore, these studies have focused on the prediction of gene expression changes, and not the impact of this regulation on metabolic pathways.

Translating expression regulation to alterations of the metabolic landscape is a key step in the prediction of macroscopic phenotypes to defined perturbations – the third focus of tuberculosis systems biology research listed at the outset of this section. Progress on this front requires the integration of gene regulatory and metabolic networks.

A promising approach to address the network integration challenge is the Probabilistic Regulation of Metabolism (PROM) described in [73]. This algorithm incorporates the control logic of a transcription regulatory network augmented by a compendium of transcriptional data with curated metabolic networks using a probabilistic Boolean formalism. In the simplest case this method assigns the probability of "enzyme X" being present in a system based on the abundance of "regulator A" which controls its expression. The maximum flux of substrates through enzyme X is then constrained by the levels of all relevant gene products. Values for all regulators and targets in the system are determined and the output from this union of transcriptional regulatory and metabolic networks is a prediction of bacterial growth rate scaled from a maximum as a result of defined perturbations (e.g., *in silico* rate inhibition or gene knockout).

PROM was applied to *M. tuberculosis* using the curated gene regulatory network described by Balázsi et al. [69] and the metabolic network of Jamshidi & Palsson [201]. Using this approach these authors were able to accurately predict the phenotypes for 23 of 24 transcription factor perturbations for which they had data. Because the PROM algorithm could only assess the impact of disrupting target genes/metabolic reactions for which regulators were known, the overall scope of the predictions were limited by the incompleteness of the underlying regulatory network model. As noted above, results described in this dissertation nearly double the size of the experimentally-characterized *M. tuberculosis* gene regulatory network. Furthermore, because PROM utilizes a probabilistic framework to parameterize gene product abundance, it should be straightforward to implement additional constraints based on experimental data. To this end, the degradation rates of individual transcripts reported in

chapter 5 of this dissertation could refine TF or target abundance estimates under different modeled conditions. Integration of these data sets describing many more interaction constraints with the PROM algorithm should yield many more predictions of growth phenotypes and intervention points in *M. tuberculosis.* This is a critical step in the development of tuberculosis systems biology.

Systems biology of tuberculosis: today and tomorrow

The results presented in this dissertation describe our efforts to create an experimental framework to define the control logic underpinning transcript regulation in *M. tuberculosis*, with attention to physiological implications of these programs. As described throughout chapters 2-5 we have made significant progress, and the next challenge lies with integrating different modeling modalities to transition from predictions of gene expression to predictions of phenotypes.

Our work highlights outstanding questions where systems modeling might be most profitably employed: what regulatory checkpoints need to be triggered in order to halt bacterial replication? Do similar regulatory switches exist to govern resumption of growth upon the re-introduction of oxygen? What metabolic functions are unique and necessary for the maintenance of dormancy? Finally, the differential stability of the *M. tuberculosis* RNA pool is an interesting phenomenon and has the potential to inform our understanding of disease transmission. In the natural history of tuberculosis disease, droplet nuclei containing bacteria are dispersed via aerosols from one host to another. The temperature stabilization of the transcriptome described in chapter 5 of this dissertation suggests that rather than rapidly adapting to the transmission environment outside of the host, individual bacteria may harbor some “transcript memory” of the infection environment from which they were transmitted. It is conceivable that bacteria originating from varying disease microenvironments demonstrate correspondingly variable transmission/colonization efficiency. As described in chapter 1 of this dissertation, roughly two-thirds of people exposed to *M. tuberculosis* (close contacts of active cases) never show any sign of disease [6], and the transcript memory of the bacterium may help to explain this

phenomenon. An interesting application of the modeling efforts described in this dissertation would be to elucidate which bacterial regulatory states are primed for successful transmission.

The ultimate utility of systems biology research lies in our ability to shed light on complex interactions across modalities that previously eluded reductionist approaches. For the experimentalist, the necessary corollary of this process is the transformation of regulatory networks from esoteric maps of interactions to testable hypotheses. In the case of *M. tuberculosis* there are obvious and pressing needs for greater understanding of the basic biology of the organism, as well as the identification of novel targets for intervention. This dissertation describes advancements we have made toward achieving this goal.

References

1. Herzog, B.H., *History of Tuberculosis.* Respiration, 1998. **65**(1): p. 5-15.
2. Corbett, E.L., et al., *The growing burden of tuberculosis: global trends and interactions with the HIV epidemic.* Arch Intern Med, 2003. **163**(9): p. 1009-21.
3. Dye, C., K. Floyd, and M. Uplekar, *Global tuberculosis control: surveillance, planning, financing : WHO report 2008*. 2008, World Health Organization: Geneva.
4. Raviglione, M.C., *The TB epidemic from 1992 to 2002.* Tuberculosis (Edinb), 2003. **83**(1-3): p. 4-14.
5. Cosma, C.L., D.R. Sherman, and L. Ramakrishnan, *The secret lives of pathogenic mycobacteria.* Annu Rev Microbiol, 2003. **57**: p. 641-76.
6. Parrish, N.M., J.D. Dick, and W.R. Bishai, *Mechanisms of latency in Mycobacterium tuberculosis.* Trends Microbiol, 1998. **6**(3): p. 107-12.
7. Gedde-Dahl, T., *Tuberculous infection in the light of tuberculin matriculation.* Am J Hyg, 1952. **56**: p. 139-214.
8. McShane, H., *Co-infection with HIV and TB: double trouble.* Int J STD AIDS, 2005. **16**(2): p. 95-100.
9. Gomez, J.E. and J.D. McKinney, *M. tuberculosis persistence, latency, and drug tolerance.* Tuberculosis (Edinb), 2004. **84**(1-2): p. 29-44.
10. Via, L.E., et al., *Tuberculous granulomas are hypoxic in guinea pigs, rabbits, and nonhuman primates.* Infect Immun, 2008. **76**(6): p. 2333-40.
11. Rao, S., et al., *The protonmotive force is required for maintaining ATP homeostasis and viability of hypoxic, nonreplicating Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2008. **105**(33): p. 11945-50.
12. Boshoff, H.I. and C.E. Barry, 3rd, *Tuberculosis - metabolism and respiration in the absence of growth.* Nat Rev Microbiol, 2005. **3**(1): p. 70-80.
13. Davis, J.M. and L. Ramakrishnan, *The role of the granuloma in expansion and dissemination of early tuberculous infection.* Cell, 2009. **136**(1): p. 37-49.
14. Adler, J.J. and D.N. Rose, *Transmission and pathogenesis of tuberculosis*, in *Tuberculosis*, W.N. Rom and S.M. Garay, Editors. 1996, Little, Brown and Co.: Boston. p. 129-140.
15. Kaplan, G., et al., *Mycobacterium tuberculosis growth at the cavity surface: a microenvironment with failed immunity.* Infect Immun, 2003. **71**(12): p. 7099-108.
16. Medlar, E.M. and K.T. Sasano, *A study of the pathology of experimental pulmonary tuberculosis in the rabbit.* Am Rev Tuberc, 1936. **34**: p. 456-476.
17. Rich, A.R. and R.H. Follis, Jr., *The effect of low oxygen tension upon the development of experimental tuberculosis.* Bull Johns Hopkins Hosp, 1942. **71**: p. 345-363.
18. Olender, S., et al., *Low prevalence and increased household clustering of Mycobacterium tuberculosis infection in high altitude villages in Peru.* Am J Trop Med Hyg, 2003. **68**(6): p. 721-727.
19. Adams, W.E. and A.J. Vorwald, *The treatment of pulmonary tuberculosis by bronchial occlusion.* J Thoracic Surg, 1934. **3**: p. 633-666.
20. Boyd, A.D., B.K. Crawford, and L. Glassman, *Surgical therapy of tuberculosis*, in *Tuberculosis*, W.N. Rom and S.M. Garay, Editors. 1996, Little, Brown and Co.: Boston. p. 513-523.
21. Flynn, J., *Lessons from experimental Mycobacterium tuberculosis infections.* Microbes Infect, 2006. **8**(4): p. 1179-88.
22. Scanga, C.A., et al., *Reactivation of latent tuberculosis: variations on the Cornell murine model.* Infect. Immun., 1999. **67**(9): p. 4531-8.

23. Aly, S., et al., *Oxygen status of lung granulomas in Mycobacterium tuberculosis-infected mice.* J Pathol, 2006. **210**(3): p. 298-305.
24. Gill, W., et al., *A Replication Clock for Mycobacterium tuberculosis.* Nat Med, 2009. **15**(2): p. 211-214.
25. Lin, P.L., et al., *Early events in Mycobacterium tuberculosis infection in cynomolgus macaques.* Infect Immun, 2006. **74**(7): p. 3790-803.
26. Wayne, L.G., *Synchronized replication of Mycobacterium tuberculosis.* Infect Immun, 1977. **17**(3): p. 528-30.
27. Wayne, L.G. and K.Y. Lin, *Glyoxylate metabolism and adaptation of Mycobacterium tuberculosis to survival under anaerobic conditions.* Infect Immun, 1982. **37**(3): p. 1042-9.
28. Wayne, L.G. and L.G. Hayes, *An in vitro model for sequential study of shiftdown of Mycobacterium tuberculosis through two stages of nonreplicating persistence.* Infect. Immun., 1996. **64**(6): p. 2062-9.
29. Wayne, L.G. and C.D. Sohaskey, *Nonreplicating persistence of Mycobacterium tuberculosis.* Annu Rev Microbiol, 2001. **55**: p. 139-63.
30. Hu, Y.M., et al., *Protein synthesis is shutdown in dormant Mycobacterium tuberculosis and is reversed by oxygen or heat shock.* FEMS Microbiol. Lett., 1998. **158**(1): p. 139-45.
31. Rodrigue, S., et al., *The sigma factors of Mycobacterium tuberculosis.* FEMS Microbiol Rev, 2006. **30**(6): p. 926-41.
32. Boshoff, H.I., et al., *Biosynthesis and recycling of nicotinamide cofactors in Mycobacterium tuberculosis. An essential role for NAD in nonreplicating bacilli.* J Biol Chem, 2008. **283**(28): p. 19329-41.
33. Bryk, R., et al., *Selective killing of nonreplicating mycobacteria.* Cell Host Microbe, 2008. **3**(3): p. 137-45.
34. Koul, A., et al., *Diarylquinolines are bactericidal for dormant mycobacteria as a result of disturbed ATP homeostasis.* J Biol Chem, 2008. **283**(37): p. 25273-80.
35. Kempner, W., *Oxygen tension and the tubercle bacillus.* Am. Rev. Tubercul., 1939. **40**: p. 157-168.
36. Canetti, G., *Growth of the tubercle bacilus in the tuberculosis lesion*, in *The tubercle bacillus in the pulmonary lesion of man*. 1955, Springer Publishing Co.: New York. p. 111-126.
37. Yuan, Y., et al., *The 16-kDa α-crystallin (Acr) protein of Mycobacterium tuberculosis is required for growth in macrophages.* Proc. Nat. Acad. Sci. USA, 1998. **95**: p. 9578-83.
38. Sherman, D.R., et al., *Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha-crystallin.* Proc Natl Acad Sci U S A, 2001. **98**(13): p. 7534-9.
39. Dasgupta, N., et al., *Characterization of a two-component system, devR-devS, of Mycobacterium tuberculosis.* Tuber. Lung Dis., 2000. **80**(3): p. 141-59.
40. Park, H.D., et al., *Rv3133c/dosR is a transcription factor that mediates the hypoxic response of Mycobacterium tuberculosis.* Mol Microbiol, 2003. **48**(3): p. 833-43.
41. Voskuil, M.I., et al., *Inhibition of respiration by nitric oxide induces a Mycobacterium tuberculosis dormancy program.* J Exp Med, 2003. **198**(5): p. 705-13.
42. Schnappinger, D., et al., *Transcriptional adaptation of Mycobacterium tuberculosis within macrophages: insights into the phagosomal environment.* J Exp Med, 2003. **198**(5): p. 693-704.
43. Karakousis, P.C., et al., *Dormancy phenotype displayed by extracellular Mycobacterium tuberculosis within artificial granulomas in mice.* J Exp Med, 2004. **200**(5): p. 647-57.
44. Shi, L., et al., *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. **100**(1): p. 241-6.

45. Shiloh, M.U., P. Manzanillo, and J.S. Cox, *Mycobacterium tuberculosis senses host-derived carbon monoxide during macrophage infection.* Cell Host Microbe, 2008. **3**(5): p. 323-30.
46. Kumar, A., et al., *Heme oxygenase-1-derived carbon monoxide induces the Mycobacterium tuberculosis dormancy regulon.* J Biol Chem, 2008. **283**(26): p. 18032-9.
47. Pang, X., et al., *Evidence for complex interactions of stress-associated regulons in an mprAB deletion mutant of Mycobacterium tuberculosis.* Microbiology, 2007. **153**(Pt 4): p. 1229-42.
48. Rohde, K., et al., *Mycobacterium tuberculosis and the environment within the phagosome.* Immunol Rev, 2007. **219**(1): p. 37-54.
49. Vasudeva-Rao, H.M. and K.A. McDonough, *Expression of the Mycobacterium tuberculosis acr-Coregulated Genes from the DevR (DosR) Regulon Is Controlled by Multiple Levels of Regulation.* 2008. **76**(6): p. 2478-2489.
50. Wisedchaisri, G., et al., *Structures of Mycobacterium tuberculosis DosR and DosR-DNA complex involved in gene activation during adaptation to hypoxic latency.* J Mol Biol, 2005. **354**(3): p. 630-41.
51. Roberts, D.M., et al., *Two sensor kinases contribute to the hypoxic response of Mycobacterium tuberculosis.* J Biol Chem, 2004. **279**(22): p. 23082-7.
52. Kumar, A., et al., *Mycobacterium tuberculosis DosS is a redox sensor and DosT is a hypoxia sensor.* Proc Natl Acad Sci U S A, 2007. **104**(28): p. 11568-73.
53. Boon, C. and T. Dick, *Mycobacterium bovis response regulator essential for hypoxic dormancy.* J Bacteriol, 2002. **184**(24): p. 6760-7.
54. Rustad, T.R., et al., *The enduring hypoxic response of Mycobacterium tuberculosis.* PLoS One, 2008. **3**(1): p. e1502.
55. Parish, T., et al., *Deletion of two-component regulatory systems increases the virulence of Mycobacterium tuberculosis.* Infect Immun, 2003. **71**(3): p. 1134-40.
56. Malhotra, V., et al., *Disruption of response regulator gene, devR, leads to attenuation in virulence of Mycobacterium tuberculosis.* FEMS Microbiol Lett, 2004. **231**(2): p. 237-45.
57. Converse, P.J., et al., *Role of the dosR-dosS two-component regulatory system in Mycobacterium tuberculosis virulence in three animal models.* Infect Immun, 2009. **77**(3): p. 1230-7.
58. Lee, B.Y., S.A. Hefta, and P.J. Brennan, *Characterization of the major membrane protein of virulent Mycobacterium tuberculosis.* Infect. Immun., 1992. **60**(5): p. 2066-74.
59. Reed, M.B., et al., *The W-Beijing lineage of Mycobacterium tuberculosis overproduces triglycerides and has the DosR dormancy regulon constitutively upregulated.* J Bacteriol, 2007. **189**(7): p. 2583-9.
60. de Jong, B.C., et al., *Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia.* J Infect Dis, 2008. **198**(7): p. 1037-43.
61. Balazsi, G., et al., *The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest.* Mol Syst Biol, 2008. **4**: p. 225.
62. Alon, U., *An introduction to systems biology : design principles of biological circuits*. Chapman & Hall/CRC mathematical and computational biology series. 2007, Boca Raton, FL: Chapman & Hall/CRC. xvi, 301 p., 4 p. of plates.
63. Thieffry, D., et al., *From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli.* Bioessays, 1998. **20**(5): p. 433-40.
64. Guelzim, N., et al., *Topological and causal structure of the yeast transcriptional regulatory network.* Nat Genet, 2002. **31**(1): p. 60-3.
65. Schlitt, T. and A. Brazma, *Current approaches to gene regulatory network modelling.* BMC Bioinformatics, 2007. **8 Suppl 6**: p. S9.
66. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology.* Nat Rev Genet, 2009. **10**(10): p. 669-80.

67. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.
68. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.
69. Balázsi, G.H., A.P.; Shi, L.; Gennaro, M.L., *The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest.* Mol. Syst. Biol., 2008. **4**: p. 225.
70. Sanz, J., et al., *The transcriptional regulatory network of Mycobacterium tuberculosis.* PloS One, 2011. **6**(7): p. e22178.
71. Gama-Castro, S., et al., *RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units).* Nucleic Acids Res, 2011. **39**(Database issue): p. D98-105.
72. Faith, J.J., et al., *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol, 2007. **5**(1): p. e8.
73. Chandrasekaran, S. and N.D. Price, *Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(41): p. 17845-50.
74. Shiloh, M.U. and P.A. DiGiuseppe Champion, *To catch a killer. What can mycobacterial models teach us about Mycobacterium tuberculosis pathogenesis?* Curr Opin Microbiol, 2010. **13**(1): p. 86-92.
75. Cosma, C.L., D.R. Sherman, and L. Ramakrishnan, *The secret lives of the pathogenic mycobacteria.* Annu Rev Microbiol, 2003. **57**: p. 641-76.
76. Neyrolles, O., et al., *Is Adipose Tissue a Place for Mycobacterium tuberculosis Persistence?* PLoS One, 2006. **1**: p. e43.
77. Cole, S.T., et al., *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence.* Nature, 1998. **393**(6685): p. 537-44.
78. Ehrt, S., et al., *Controlling gene expression in mycobacteria with anhydrotetracycline and Tet repressor.* Nucleic Acids Res, 2005. **33**(2): p. e21.
79. Valouev, A., et al., *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.* Nature methods, 2008. **5**(9): p. 829-34.
80. Lun, D.S., et al., *A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data.* Genome Biol, 2009. **10**(12): p. R142.
81. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
82. Janowski, R., et al., *Structural analysis reveals DNA binding properties of Rv2827c, a hypothetical protein from Mycobacterium tuberculosis.* J Struct Funct Genomics, 2009. **10**(2): p. 137-50.
83. Salzman, V., et al., *Transcriptional regulation of lipid homeostasis in mycobacteria.* Mol Microbiol, 2010. **78**(1): p. 64-77.
84. McGuire, A.M., et al., *Comparative analysis of Mycobacterium and related Actinomycetes yields insight into the evolution of Mycobacterium tuberculosis pathogenesis.* BMC Genomics, 2012. **13**(1): p. 120.
85. Kendall, S.L., et al., *A highly conserved transcriptional repressor controls a large regulon involved in lipid degradation in Mycobacterium smegmatis and Mycobacterium tuberculosis.* Mol Microbiol, 2007. **65**(3): p. 684-99.
86. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.
87. MacQuarrie, K.L., et al., *Genome-wide transcription factor binding: beyond direct target regulation.* Trends Genet, 2011. **27**(4): p. 141-8.

88. Moreno-Campuzano, S., S.C. Janga, and E. Perez-Rueda, *Identification and analysis of DNA-binding transcription factors in Bacillus subtilis and other Firmicutes--a genomic approach.* BMC Genomics, 2006. **7**: p. 147.
89. Busenlehner, L.S., M.A. Pennella, and D.P. Giedroc, *The SmtB/ArsR family of metalloregulatory transcriptional repressors: Structural insights into prokaryotic metal resistance.* FEMS Microbiol Rev, 2003. **27**(2-3): p. 131-43.
90. Ramos, J.L., et al., *The TetR family of transcriptional repressors.* Microbiol Mol Biol Rev, 2005. **69**(2): p. 326-56.
91. Rustad, T.R., et al., *Hypoxia: a window into Mycobacterium tuberculosis latency.* Cell Microbiol, 2009. **11**(8): p. 1151-9.
92. Ryndak, M., S. Wang, and I. Smith, *PhoP, a key player in Mycobacterium tuberculosis virulence.* Trends in microbiology, 2008. **16**(11): p. 528-34.
93. Chauhan, S., et al., *Comprehensive insights into Mycobacterium tuberculosis DevR (DosR) regulon activation switch.* Nucleic Acids Res, 2011. **39**(17): p. 7400-14.
94. Gupta, S., A. Sinha, and D. Sarkar, *Transcriptional autoregulation by Mycobacterium tuberculosis PhoP involves recognition of novel direct repeat sequences in the regulatory region of the promoter.* FEBS Lett, 2006. **580**(22): p. 5328-38.
95. Gupta, S., et al., *Mycobacterium tuberculosis PhoP recognizes two adjacent direct-repeat sequences to form head-to-head dimers.* Journal of bacteriology, 2009. **191**(24): p. 7466-76.
96. Cline, M.S., et al., *Integration of biological networks and gene expression data using Cytoscape.* Nat Protoc, 2007. **2**(10): p. 2366-82.
97. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
98. Organization, W.H., *Global Tuberculosis Control*. 2001.
99. Roberts, C. and J. Buikstra, *The Bioarchaeology of Tuberculosis: A Global View on a Reemerging Disease*. 2008: University Press of Florida.
100. Dannenberg, A.M., Jr., *Immunopathogenesis of pulmonary tuberculosis.* Hosp Pract, 1993. **28**(1): p. 51-8.
101. Manabe, Y.C. and W.R. Bishai, *Latent Mycobacterium tuberculosis-persistence, patience, and winning by waiting.* Nat Med, 2000. **6**(12): p. 1327-9.
102. McKinney, J.D., et al., *Persistence of Mycobacterium tuberculosis in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase.* Nature, 2000. **406**(6797): p. 735-8.
103. Flynn, J.L. and J. Chan, *Tuberculosis: latency and reactivation.* Infect Immun, 2001. **69**(7): p. 4195-201.
104. Yang, X., et al., *Cholesterol metabolism increases the metabolic pool of propionate in Mycobacterium tuberculosis.* Biochemistry, 2009. **48**(18): p. 3819-21.
105. Miner, M.D., et al., *Role of cholesterol in Mycobacterium tuberculosis infection.* Indian J Exp Biol, 2009. **47**(6): p. 407-11.
106. Chang, J.C., et al., *igr Genes and Mycobacterium tuberculosis cholesterol metabolism.* J Bacteriol, 2009. **191**(16): p. 5232-9.
107. Kendall, S.L., et al., *Cholesterol utilization in mycobacteria is controlled by two TetR-type transcriptional regulators: kstR and kstR2.* Microbiology, 2010. **156**(Pt 5): p. 1362-71.
108. Nesbitt, N.M., et al., *A thiolase of Mycobacterium tuberculosis is required for virulence and production of androstenedione and androstadienedione from cholesterol.* Infect Immun, 2010. **78**(1): p. 275-82.
109. Uhia, I., et al., *Characterization of the KstR-dependent promoter of the first step of cholesterol degradative pathway in Mycobacterium smegmatis.* Microbiology, 2011.

110. Daniel, J., et al., *Mycobacterium tuberculosis uses host triacylglycerol to accumulate lipid droplets and acquires a dormancy-like phenotype in lipid-loaded macrophages.* PLoS Pathog, 2011. **7**(6): p. e1002093.
111. Low, K.L., et al., *Triacylglycerol utilization is required for regrowth of in vitro hypoxic nonreplicating Mycobacterium bovis bacillus Calmette-Guerin.* J Bacteriol, 2009. **191**(16): p. 5037-43.
112. Russell, D.G., H.C. Mwandumba, and E.E. Rhoades, *Mycobacterium and the coat of many lipids.* J Cell Biol, 2002. **158**(3): p. 421-6.
113. Kaur, D., et al., *Chapter 2: Biogenesis of the cell wall and other glycoconjugates of Mycobacterium tuberculosis.* Adv Appl Microbiol, 2009. **69**: p. 23-78.
114. Balazsi, G., et al., *The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest.* Mol Syst Biol, 2008. **4**.
115. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-560.
116. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Meth, 2007. **4**(8): p. 651-657.
117. Ehrt, S. and D. Schnappinger, *Controlling gene expression in mycobacteria.* Future Microbiol, 2006. **1**(2): p. 177-84.
118. Klotzsche, M., S. Ehrt, and D. Schnappinger, *Improved tetracycline repressors for gene silencing in mycobacteria.* Nucleic Acids Res, 2009. **37**(6): p. 1778-88.
119. Guo, X.V., et al., *Silencing Mycobacterium smegmatis by using tetracycline repressors.* J Bacteriol, 2007. **189**(13): p. 4614-23.
120. Kim, J., et al., *An extended transcriptional network for pluripotency of embryonic stem cells.* Cell, 2008. **132**(6): p. 1049-61.
121. Mazzoni, E.O., et al., *Embryonic stem cell-based mapping of developmental transcriptional programs.* Nat Methods, 2011. **8**(12): p. 1056-8.
122. Saini, D.K., et al., *DevR-DevS is a bona fide two-component system of Mycobacterium tuberculosis that is hypoxia-responsive in the absence of the DNA-binding domain of DevR.* Microbiology, 2004. **150**(Pt 4): p. 865-75.
123. Flores Valdez, M.A. and G.K. Schoolnik, *DosR-regulon genes induction in Mycobacterium bovis BCG under aerobic conditions.* Tuberculosis (Edinb), 2010. **90**(3): p. 197-200.
124. Chauhan, S., et al., *Comprehensive insights into Mycobacterium tuberculosis DevR (DosR) regulon activation switch.* Nucleic Acids Res, 2011.
125. Sherman, D.R., et al., *Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -crystallin.* Proc Natl Acad Sci U S A, 2001. **98**(13): p. 7534-9.
126. Vasudeva-Rao, H.M. and K.A. McDonough, *Expression of the Mycobacterium tuberculosis acr-coregulated genes from the DevR (DosR) regulon is controlled by multiple levels of regulation.* Infect Immun, 2008. **76**(6): p. 2478-89.
127. Fernandez, P.C., et al., *Genomic targets of the human c-Myc protein.* Genes Dev, 2003. **17**(9): p. 1115-29.
128. Farnham, P.J., *Insights from genomic profiling of transcription factors.* Nat Rev Genet, 2009. **10**(9): p. 605-16.
129. Tanay, A., *Extensive low-affinity transcriptional interactions in the yeast genome.* Genome Res, 2006. **16**(8): p. 962-72.
130. Zhong, M., et al., *Genome-wide identification of binding sites defines distinct functions for Caenorhabditis elegans PHA-4/FOXA in development and environmental response.* PLoS Genet, 2010. **6**(2): p. e1000848.

131. Zeitlinger, J., et al., *Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo.* Genes Dev, 2007. **21**(4): p. 385-90.
132. Li, X.Y., et al., *Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm.* PLoS Biol, 2008. **6**(2): p. e27.
133. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.
134. Cao, Y., et al., *Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming.* Dev Cell, 2010. **18**(4): p. 662-74.
135. Gautam, U.S., S. Chauhan, and J.S. Tyagi, *Determinants outside the DevR C-terminal domain are essential for cooperativity and robust activation of dormancy genes in Mycobacterium tuberculosis.* PloS One, 2011. **6**(1): p. e16500.
136. Niu, W., et al., *Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans.* Genome Res, 2011. **21**(2): p. 245-54.
137. Cho, B.K., et al., *Deciphering the transcriptional regulatory logic of amino acid metabolism.* Nat Chem Biol, 2011.
138. Vilar, J.M. and L. Saiz, *DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise.* Curr Opin Genet Dev, 2005. **15**(2): p. 136-44.
139. Barnard, A., A. Wolfe, and S. Busby, *Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes.* Curr Opin Microbiol, 2004. **7**(2): p. 102-8.
140. Gold, B., et al., *The Mycobacterium tuberculosis IdeR is a dual functional regulator that controls transcription of genes involved in iron acquisition, iron storage and survival in macrophages.* Mol Microbiol, 2001. **42**(3): p. 851-65.
141. Rodriguez, G.M., et al., *ideR, An essential gene in Mycobacterium tuberculosis: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response.* Infect Immun, 2002. **70**(7): p. 3371-81.
142. Schroder, J. and A. Tauch, *Transcriptional regulation of gene expression in Corynebacterium glutamicum: the role of global, master and local regulators in the modular and hierarchical gene regulatory network.* FEMS Microbiol Rev, 2010. **34**(5): p. 685-737.
143. Shen-Orr, S.S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nat Genet, 2002. **31**(1): p. 64-8.
144. Mangan, S. and U. Alon, *Structure and function of the feed-forward loop network motif.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 11980-5.
145. Mangan, S., A. Zaslaver, and U. Alon, *The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks.* J Mol Biol, 2003. **334**(2): p. 197-204.
146. Milo, R., et al., *Network motifs: simple building blocks of complex networks.* Science, 2002. **298**(5594): p. 824-7.
147. Kashtan, N., et al., *Topological generalizations of network motifs.* Phys Rev E Stat Nonlin Soft Matter Phys, 2004. **70**(3 Pt 1): p. 031909.
148. Gordon, B.R., et al., *Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2010. **107**(11): p. 5154-9.
149. Colangeli, R., et al., *The multifunctional histone-like protein Lsr2 protects mycobacteria against reactive oxygen intermediates.* Proc Natl Acad Sci U S A, 2009. **106**(11): p. 4414-8.
150. Colangeli, R., et al., *Transcriptional regulation of multi-drug tolerance and antibiotic-induced responses by the histone-like protein Lsr2 in M. tuberculosis.* PLoS Pathog, 2007. **3**(6): p. e87.

151. Travers, A. and G. Muskhelishvili, *Bacterial chromatin.* Curr Opin Genet Dev, 2005. **15**(5): p. 507-14.
152. Gonzalo-Asensio, J., et al., *PhoP: a missing piece in the intricate puzzle of Mycobacterium tuberculosis virulence.* PLoS One, 2008. **3**(10): p. e3496.
153. Gonzalo Asensio, J., et al., *The virulence-associated two-component PhoP-PhoR system controls the biosynthesis of polyketide-derived lipids in Mycobacterium tuberculosis.* J Biol Chem, 2006. **281**(3): p. 1313-6.
154. Ryndak, M., S. Wang, and I. Smith, *PhoP, a key player in Mycobacterium tuberculosis virulence.* Trends Microbiol, 2008. **16**(11): p. 528-34.
155. Abramovitch, R.B., et al., *aprABC: a Mycobacterium tuberculosis complex-specific locus that modulates pH-driven adaptation to the macrophage phagosome.* Mol Microbiol, 2011. **80**(3): p. 678-94.
156. Singh, A., et al., *Mycobacterium tuberculosis WhiB3 maintains redox homeostasis by regulating virulence lipid anabolism to modulate macrophage response.* PLoS Pathog, 2009. **5**(8): p. e1000545.
157. Singh, A., et al., *Mycobacterium tuberculosis WhiB3 responds to O2 and nitric oxide via its [4Fe-4S] cluster and is essential for nutrient starvation survival.* Proc Natl Acad Sci U S A, 2007. **104**(28): p. 11562-7.
158. Ernst, J., et al., *Reconstructing dynamic regulatory maps.* Mol Syst Biol, 2007. **3**: p. 74.
159. Sherrid, A.M., et al., *Characterization of a Clp protease gene regulator and the reaeration response in Mycobacterium tuberculosis.* PLoS One, 2010. **5**(7): p. e11622.
160. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res, 2008. **18**(11): p. 1851-1858.
161. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
162. Lun, D.S., et al., *A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data.* Genome Biology, 2009. **10**(12): p. R142.
163. Bailey, T.L., et al., *MEME: discovering and analyzing DNA and protein sequence motifs.* Nucleic Acids Research, 2006. **34**(Web Server issue): p. W369-73.
164. Grant, C.E., T.L. Bailey, and W.S. Noble, *FIMO: scanning for occurrences of a given motif.* Bioinformatics. **27**(7): p. 1017-1018.
165. Akaike, H., *A new look at the statistical model identification.* IEEE Transactions on Automatic Control 1974. **19**(6): p. 716.
166. Lilliefors, H., *On the Kolmogorov–Smirnov test for normality with mean and variance unknown.* Journal of the American Statistical Association, 1967. **62**: p. 339.
167. Alon, U., *An Introduction to Systems Biology: Design Principles of Biological Circuits*. 2006: Chapman and Hall/CRC.
168. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. **31**(4): p. e15.
169. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
170. Fordyce, P.M., et al., *De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis.* Nat Biotechnol, 2010. **28**(9): p. 970-5.
171. Chao, M.C. and E.J. Rubin, *Letting sleeping dos lie: does dormancy play a role in tuberculosis?* Annu Rev Microbiol, 2010. **64**: p. 293-311.

172. Roupie, V., et al., *Immunogenicity of eight dormancy regulon-encoded proteins of Mycobacterium tuberculosis in DNA-vaccinated and tuberculosis-infected mice.* Infect Immun, 2007. **75**(2): p. 941-9.
173. Fallow, A., P. Domenech, and M.B. Reed, *Strains of the East Asian (W/Beijing) lineage of Mycobacterium tuberculosis are DosS/DosT-DosR two-component regulatory system natural mutants.* Journal of bacteriology, 2010. **192**(8): p. 2228-38.
174. Flores Valdez, M.A. and G.K. Schoolnik, *DosR-regulon genes induction in Mycobacterium bovis BCG under aerobic conditions.* Tuberculosis (Edinburgh, Scotland), 2010. **90**(3): p. 197-200.
175. Barry, C.E., 3rd, et al., *The spectrum of latent tuberculosis: rethinking the biology and intervention strategies.* Nat Rev Microbiol, 2009. **7**(12): p. 845-55.
176. Russell, D.G., C.E. Barry, 3rd, and J.L. Flynn, *Tuberculosis: what we don't know can, and does, hurt us.* Science, 2010. **328**(5980): p. 852-6.
177. Bernstein, J.A., et al., *Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays.* Proc Natl Acad Sci U S A, 2002. **99**(15): p. 9697-702.
178. Anderson, K.L. and P.M. Dunman, *Messenger RNA Turnover Processes in Escherichia coli, Bacillus subtilis, and Emerging Studies in Staphylococcus aureus.* Int J Microbiol, 2009. **2009**: p. 525491.
179. Hambraeus, G., C. von Wachenfeldt, and L. Hederstedt, *Genome-wide survey of mRNA half-lives in Bacillus subtilis identifies extremely stable mRNAs.* Mol Genet Genomics, 2003. **269**(5): p. 706-14.
180. Cohen, S.N. and K.J. McDowall, *RNase E: still a wonderfully mysterious enzyme.* Mol Microbiol, 1997. **23**(6): p. 1099-106.
181. Georgellis, D., et al., *Retarded RNA turnover in Escherichia coli: a means of maintaining gene expression during anaerobiosis.* Mol Microbiol, 1993. **9**(2): p. 375-81.
182. Prud'homme-Genereux, A., et al., *Physical and functional interactions among RNase E, polynucleotide phosphorylase and the cold-shock protein, CsdA: evidence for a 'cold shock degradosome'.* Mol Microbiol, 2004. **54**(5): p. 1409-21.
183. Anderson, K.L., et al., *Characterization of the Staphylococcus aureus heat shock, cold shock, stringent, and SOS responses and their effects on log-phase mRNA turnover.* J Bacteriol, 2006. **188**(19): p. 6739-56.
184. Kovacs, L., et al., *Mycobacterial RNase E-associated proteins.* Microbiol Immunol, 2005. **49**(11): p. 1003-7.
185. Zeller, M.E., et al., *Quaternary structure and biochemical properties of mycobacterial RNase E/G.* Biochem J, 2007. **403**(1): p. 207-15.
186. Unniraman, S., R. Prakash, and V. Nagaraja, *Alternate paradigm for intrinsic transcription termination in eubacteria.* J Biol Chem, 2001. **276**(45): p. 41850-5.
187. Minch, K., T. Rustad, and D.R. Sherman, *Mycobacterium tuberculosis Growth following Aerobic Expression of the DosR Regulon.* PloS One, 2012. **7**(4): p. e35935.
188. Rustad, T., et al., *RNA Isolation*, in *Mycobacteria Protocols Handbook*, T. Parish and A. Brown, Editors. 2007, Wiley.
189. Cangelosi, G.A. and W.H. Brabant, *Depletion of pre-16S rRNA in starved Escherichia coli cells.* J Bacteriol, 1997. **179**(14): p. 4457-63.
190. Levin, M.E. and G.F. Hatfull, *Mycobacterium smegmatis RNA polymerase: DNA supercoiling, action of rifampicin and mechanism of rifampicin resistance.* Mol Microbiol, 1993. **8**(2): p. 277-85.
191. Reddy, T.B., et al., *TB database: an integrated platform for tuberculosis research.* Nucleic Acids Res, 2009. **37**(Database issue): p. D499-508.

192. Betts, J.C., et al., *Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling.* Mol Microbiol, 2002. **43**(3): p. 717-31.
193. Hundt, S., et al., *Global analysis of mRNA decay in Halobacterium salinarum NRC-1 at single-gene resolution using DNA microarrays.* J Bacteriol, 2007. **189**(19): p. 6936-44.
194. Andersson, A.F., et al., *Global analysis of mRNA stability in the archaeon Sulfolobus.* Genome Biol, 2006. **7**(10): p. R99.
195. Ramage, H.R., L.E. Connolly, and J.S. Cox, *Comprehensive functional analysis of Mycobacterium tuberculosis toxin-antitoxin systems: implications for pathogenesis, stress responses, and evolution.* PLoS Genet, 2009. **5**(12): p. e1000767.
196. D'Auria, G., et al., *Dynamical properties of cold shock protein A from Mycobacterium tuberculosis.* Biochem Biophys Res Commun, 2010. **402**(4): p. 693-8.
197. Shalem, O., et al., *Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation.* Mol Syst Biol, 2008. **4**: p. 223.
198. Lin, P.L., et al., *Quantitative comparison of active and latent tuberculosis in the cynomolgus macaque model.* Infect Immun, 2009. **77**(10): p. 4631-42.
199. Reed, M.B., et al., *The W-Beijing lineage of Mycobacterium tuberculosis overproduces triglycerides and has the DosR dormancy regulon constitutively upregulated.* Journal of bacteriology, 2007. **189**(7): p. 2583-9.
200. Baek, S.H., A.H. Li, and C.M. Sassetti, *Metabolic regulation of mycobacterial growth and antibiotic sensitivity.* PLoS Biol, 2011. **9**(5): p. e1001065.
201. Jamshidi, N. and B.O. Palsson, *Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets.* BMC Syst Biol, 2007. **1**: p. 26.

VITA

Kyle James Minch was born and raised in northern Illinois. In 2005 he graduated magna cum laude from Augustana College with a bachelor of arts in biology. Following his undergraduate education he spent time working in Oxford, England and Blantyre, Malawi. After a brief, but noteworthy, stop in Mozambique he could be found exploring the Rocky Mountains of Colorado following his varied interests including backpacking, photography, and skiing. In 2007 he matriculated in to the Molecular and Cellular Biology Ph.D. program at the University of Washington in Seattle, Washington. He was a National Science Foundation Graduate Research Fellow while studying genome-scale regulatory networks in *Mycobacterium tuberculosis*. In Washington State he found quiet, necessary, places in the Cascade Mountains, and made several pilgrimages every hiking season. In the spring of 2012 he earned a doctor of philosophy in Molecular and Cellular Biology.