HHS Public Access

Author manuscript

Nat Commun. Author manuscript; available in PMC 2015 March 20.

Published in final edited form as:

Nat Commun. ; 6: 5829. doi:10.1038/ncomms6829.

The DNA binding network of *Mycobacterium tuberculosis*

Kyle J. Minch[1,2,*], **Tige R. Rustad**[1,*], **Eliza J.R. Peterson**[3], **Jessica Winkler**[1], **David J. Reiss**[3], **Shuyi Ma**[1,3,4], **Mark Hickey**[1], **William Brabant**[1], **Bob Morrison**[1], **Serdar Turkarslan**[3], **Chris Mawhinney**[5], **James E. Galagan**[5,6,7,8], **Nathan D. Price**[3], **Nitin S. Baliga**[3], and **David R. Sherman**[1,2,§]

Kyle J. Minch: kyle.minch@seattlebiomed.org; Tige R. Rustad: tige.rustad@seattlebiomed.org; Eliza J.R. Peterson: Eliza.Peterson@systemsbiology.org; Jessica Winkler: Jessica.winkler@seattlebiomed.org; David J. Reiss: dreiss@systemsbiology.org; Shuyi Ma: shuyima1@illinois.edu; Mark Hickey: willclone4food@gmail.com; William Brabant: william.brabant@cepheid.com; Bob Morrison: bob.morrison@seattlebiomed.org; Serdar Turkarslan: sturkarslan@systemsbiology.org; Chris Mawhinney: cmawhinney@gmail.com; James E. Galagan: jgalag@bu.edu; Nathan D. Price: nprice@systemsbiology.org; Nitin S. Baliga: nbaliga@systemsbiology.org; David R. Sherman: david.sherman@seattlebiomed.org

[1]Seattle Biomedical Research Institute, Seattle, Washington 98109, USA

[2]Interdisciplinary Program of Pathobiology, Department of Global Health, University of Washington, Seattle, Washington 98195, USA

[3]Institute for Systems Biology, 401 Terry Avenue North, Seattle, Washington 98109, USA

[4]Department of Chemical and Biomolecular Engineering, University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA

[5]Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA

[6]Department of Microbiology, Boston University, Boston, Massachusetts 02215, USA

[7]Bioinformatics Program, Boston University, Boston, Massachusetts 02215, USA

[8]The Eli and Edythe L. Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA

Abstract

Mycobacterium tuberculosis (MTB) infects 30% of all humans and kills someone every 20 – 30 seconds. Here we report genome-wide binding for ~80% of all predicted MTB transcription factors (TFs), and assayed global expression following induction of each TF. The MTB DNA binding network consists of ~16,000 binding events from 154 TFs. We identify >50 TF-DNA

§Corresponding author.

*These authors contributed equally to this work.

AUTHOR CONTRIBUTIONS

KJM and TRR conceived of the study, generated data, analyzed the results, and drafted the manuscript. EJRP performed promoter window analysis. JW generated strains and data. DJR performed consensus motif analysis. SM performed network assembly and statistical analyses. MH and WB generated strains and data. BM developed peak calling algorithm. ST performed network analyses. CM sequenced ChIP samples. JEG oversaw sequencing facility and assisted in study design. NDP and NSB assisted in editing the manuscript and analyzing data. DRS conceived of the study, led the design, organized the data analysis, and drafted the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests

consensus motifs and >1,150 promoter binding events directly associated with proximal gene regulation. An additional ~4,200 binding events are in promoter windows and represent strong candidates for direct transcriptional regulation under appropriate environmental conditions. However, we also identify >10,000 “dormant” DNA binding events that cannot be linked directly with proximal transcriptional control, suggesting that widespread DNA binding may be a common feature that should be considered when developing global models of coordinated gene expression.

INTRODUCTION

Mycobacterium tuberculosis (MTB) is a remarkably successful pathogen that infects an estimated 1.5 billion people and kills 1.3 million people each year[1]. Throughout TB disease, both bacterium and host engage in a dynamic series of adaptations to modulate local environments. For the pathogen, adaptation is principally mediated through the ~214 DNA binding proteins encoded in the MTB genome. These proteins interact with small molecule chemical messengers, other proteins, and the DNA to shape the transcriptional landscape of the cell and convert cascading stimuli into coordinated effector gene responses. Several approaches to understanding the wiring and connectivity of interacting macromolecular components of MTB have been described, including gene expression pattern-driven identification of regulatory subnetworks[2,3], metabolic reconstructions[4,5], integration of expression data from diverse experimental and environmental conditions[6], and hybrid networks that seek to bridge transcription regulation with metabolic outputs and cellular fitness[7]. In each case the goal of these approaches is to constrain the universe of potential interactions within cells through an iterative process of experimentation, data collection and computational approaches that result in network reconstruction.

Various groups have probed the gene regulatory landscape of MTB by characterizing the regulons of individual transcription factors (TFs). The most widely applied approach has been gene knockout and phenotyping or transcriptional profiling of the resultant mutant[8–10]. More recently however, technologies such as chromatin immunoprecipitation followed by microarray hybridization or high-throughput sequencing (ChIP-chip and ChIP-seq, respectively) have been applied to MTB[11–16]. These approaches identify directly sites of TF-DNA binding, and in conjunction with transcriptional profiling and/or meta-analyses offer a powerful window in to the global regulatory capacity of individual proteins. Employing ChIP-seq and transcriptional profiling, we recently described an analysis of the binding profile for 50 MTB TFs assessed in a uniform condition[17]. This preliminary network reconstruction showed good concordance with published results, as well as common features of regulatory networks from other organisms, such as robust network construction, connectivity and DNA binding motif structure[18–20].

Here we expand efforts to characterize the MTB gene regulatory network. We report the DNA binding and transcriptional regulatory profile of ~80% of all predicted MTB TFs (>150 proteins). From these data we derive high-confidence DNA consensus motifs for >50 TFs. We show that chromosomal regions proximal to coding sequence or transcription start sites are enriched for binding, allowing us to define functionally a genome-wide promoter window size for MTB. We identify 5,400 protein-DNA interactions within this window with

high probability for direct transcriptional control of proximal targets, and 1,162 binding events that regulate proximal gene expression in the experimental condition assayed. However, we also note even more DNA binding that cannot be linked directly with transcriptional control. Further, we characterize one TF in which widespread binding events, most of which are not directly associated with gene expression changes, are nonetheless dictated by specific DNA sequence motifs that can be validated by an independent experimental approach. We propose the phrase “dormant binding” to describe sequence-specific protein-DNA interactions without a proximal effect on gene expression, and suggest that this class of binding may exert proximal regulatory control under different environmental conditions, but may also contribute more subtly to the regulatory landscape of the cell. Altogether, this work presents an experimentally constrained protein-DNA interaction framework for MTB that reveals thousands of DNA binding events, many of which we can link to proximal regulatory events. Our pan-genome survey indicates that widespread, dormant TF-DNA binding is very common, and suggests that the control of gene expression in bacteria may involve a layer of complexity that is currently unappreciated.

RESULTS

We recently described a preliminary MTB gene regulatory network based on the DNA binding patterns of 50 TFs (23% of the 214 TFs of MTB)[17]. Here, we present a substantially more complete transcriptional regulatory network that incorporates updated peak calling algorithms, stringent controls/filters to define high quality TF binding (see Materials and Methods), and includes 80% of the MTB TFs (workflow in Supplementary Fig. 1).

We cloned 206 (of the estimated 214) DNA binding genes into an anhydrotetracycline - inducible Gateway shuttle vector to contain an N- or C-terminal FLAG epitope tag. The remaining 8 genes proved refractory to our sub-cloning efforts. For added inclusiveness, this list was compiled through gene annotation data from Tuberculist[21], TBDB[22], and PATRIC[23], as well as manual curation[24]. Once transformed, we cultured MTB strains to a uniform growth stage and induced expression of the gene-of-interest for 18 hours – approximately one cell division. We then harvested chromatin samples for ChIP-seq as well as total RNA for high-density transcriptional profiling by custom tiled microarray. For microarray analysis, induction and experiments were repeated with at least three biological replicates[24]. For ChIP-seq samples we employed a custom algorithm for read alignment and ChIP peak calling (Methods).

ChiP-seq data set and controls

Previously we showed that DNA binding events reproduced with high fidelity in eight of eight replicate ChIP samples[17]. In addition to the experimental ChIP samples we created a negative control composite data set against which we filtered experimental ChIP data sets. Because no single control captures all known or potential ChIP artifacts we designed this negative control compendium to include 10 diverse samples/sequencing datasets: wild-type H37Rv chromatin immunoprecipitated with and without anti-FLAG antibody (input DNA and mock IP controls, respectively), chromatin samples from uninduced expression-vector

bearing cells immunopreciptated with and without anti-FLAG antibody (basal expression from chimeric inducible promoter and mock IP controls, respectively), as well as chromatin samples from induced non-TF genes immunoprecipitated with anti-FLAG antibody (specificity of FLAG IP). We subjected each control data set to peak calling, creating an experimentally-derived negative control peak set consisting of ~2000 scored final peaks. We then compared each experimental peak with this negative control peak set to define a collection of pass-filter DNA binding events (Methods). This approach identified both global and local binding patterns for every TF assayed with associated significance scores for every ChIP peak (Supplementary Fig. 2).

Some genomic regions appeared to be hotspots for ChIP enrichment, irrespective of the significance threshold. Recent reports from yeast[25,26] suggest that loci with high transcriptional activity can be artificially enriched in ChIP assays. We compared our MTB high-occupancy sites against the absolute $\log_2$ expression value of transcripts derived from more than 700 microarrays[24] and did not observe any such correlation with hyper-enriched regions and transcript abundance (Supplementary Fig. 3). Nevertheless, we know of no biological mechanism for why these loci should be enriched across TF class and experiment. Therefore, any 50-bp region bound by more than 50 different TFs was flagged as a provisional experimental artifact and removed from subsequent analysis. This step culled 1006 peaks at 5 gene loci (Rv1088, Rv1115, Rv1396c, Rv2190c, Rv3622c-3623).

We considered the possibility that artificially high TF gene induction from our ectopic expression system might result in more DNA binding than would be observed in wild-type (WT) cells. In addressing that question, we previously demonstrated good concordance between DNA binding following ectopic induction of tagged TFs and published genome-wide binding studies that relied on native conditions and/or antibodies[12,14,17,22]. Specifically we compared data from our over-expression system to results of ChIP-seq experiments using WT cells and antibodies directed at native BlaI[12], DosR[11], or EspR[14]. In each case, approximately the same number of peaks was identified, and peak position and height were well conserved[17]. To assess this question more broadly, we compared here the number of binding sites per TF and the magnitude of TF ectopic induction, and found no correlation (Supplementary Fig. 4a). In addition, we compared TF expression levels in our over-expression system to a compendium of >2,300 published microarrays. We found that >80% of TFs were induced to a higher level by one or more experimental condition in WT cells (Supplementary Fig. 4b and[24]). Thus, while we cannot exclude the possibility that over-expression sometimes produced non-physiological DNA binding, we conclude that such spurious DNA associations are rare in our data sets.

Network topology and characteristics

We analyzed genome-wide binding profiles for all TFs at p-value cutoffs of <0.05, <0.01, <0.001, and <0.0001. As expected, the number of protein-DNA interactions shrinks as we progress to more stringent inclusion thresholds (Supplementary Fig. 5). While binding events in the range $0.05>p>0.01$ have binding scores stronger than at least 95% of all negative control peaks and are clearly distinguishable from background, they generally possess lower signal to noise ratios and skewed read distributions (Supplementary Fig. 2).

Testing showed that DNA binding events with a p~0.01 could be confirmed by independent experiment, where peaks with p~0.05 were less consistently validated. We therefore chose a cutoff of p<0.01 to filter peaks for subsequent analyses. With this threshold the physical DNA-binding map includes 15,980 protein-DNA interactions from 156 MTB TFs. Supplementary Table 1 provides all TF-target interactions, with associated peak binding metrics, genome coordinates, and confidence scores. In addition, all raw and filtered data can be found at: http://networks.systemsbiology.net/mtb.

We mapped the center of each binding event peak, and the global distribution of all TF-binding events was visualized on a circularized map of the MTB chromosome (Figure 1a). Although thousands of genome-binding events were mapped, visualization at high resolution revealed that the chromosome in general is sparsely bound (Figure 1b). The vast majority of the genome (~3.8 million base pairs, 86%) was not associated with any TF-binding, whereas ~0.6 million base pairs contained at least one binding site, and locations with more binding events were progressively fewer. Regions with multiple TFs binding in close proximity are prime candidates for combinatorial regulation. For example, our data recover the well-characterized binding of Rv3133c/DosR upstream of both Rv3134c and Rv2031c[8,11], but we also note in both regions a strong binding signature from hypoxic-responsive TF Rv1985c[27].

For individual TFs in this study, the number of DNA binding events per protein ranged from 0 to >850 (Figure 1c). No binding sites were detected for 24 TFs. There were also 7 proteins with >500 binding sites each on the chromosome, and 14 TFs accounted for ~50% of all binding in the network. For proteins that do not bind DNA as well as for prolific binders, no single gene family describes these TFs.

Correlating DNA binding with regulation of transcription

We explored binding locations relative to translation start sites of annotated genes. About 25% (nearly 4000 out of ~16,000 binding sites) were within intergenic regions. While this is roughly 2.5× what would be observed by random chance (cumulative hypergeometric mean p<0.001), the relatively low 25% intergenic enrichment was unexpected and caused us to investigate further binding site distribution characteristics. Peaks with the highest quality scores were slightly more likely to be intergenic. For instance, among the 800 best peaks, the proportion within intergenic regions rose to 29% (Supplementary Fig. 6). However, even when considering only the highest scoring peaks (top 20%) on a per-TF basis, the binding site distribution is highly idiosyncratic (Supplementary Table 2). About one-fourth of TFs exhibit 80% intergenic binding or more, while another one-fourth show at least 80% binding within coding sequences. Of the proteins with strong intergenic bias in this analysis, nearly all bind 3 or fewer times on the MTB chromosome. We cannot exclude the possibility that the prevalence of within-gene DNA binding we report is somehow a function of our approach; however, the trends observed here are broadly consistent with other genome-wide DNA binding studies in MTB[14,16], and with some reports in other bacteria[28,29].

We also analyzed the binding locations relative to an experimentally determined map of transcriptional start sites (TSS) in MTB, many of which are not consistent with traditionally defined coding region boundaries[30]. We observed a striking enrichment of TF-binding proximal to TSSs, with the highest density of binding at −18 nucleotides upstream to TSSs

(Figure 2a). To associate TFs with direct regulation of target genes, we analyzed instances where TF over-expression resulted in significantly altered expression of genes proximal to TF binding locations (Methods and[24]). By performing this analysis over different sized genomic segments, we determined that a consensus promoter spanning 150 bp upstream to 70 bp downstream of starts yielded maximal sensitivity vs. specificity (Supplementary Fig. 7). All binding events within this window were considered functional, i.e. – capable of directly regulating downstream gene expression in the right environmental context. 5,400 binding sites for 143 TFs were located within promoter windows. Because a single binding site could be associated with more than one promoter, altogether there were 7,248 TF-promoter interactions within 2,848 promoters. There were 1,243 promoters with a single TF binding site, and the median was 2 binding sites per promoter. Overexpression of TFs under reference growth conditions validated that 1,162 TF-DNA interactions can directly regulate proximal genes (Supplementary Table 3). Thus, despite the known conditional nature of gene regulation, we were able to validate over 20% of all promoter-proximal binding events using only one reference laboratory growth condition. By extension, a large fraction of the >7,200 promoter-proximal TF-DNA interactions are likely to regulate gene expression directly in the appropriate environmental context, and can even be used to refine promoter predictions. For example, expression of the putative benzoquinone methlytransferase Rv0560c was previously predicted to be controlled by an unknown repressor of the MarR family[31]. We found 5 TFs that bind near the start of this gene, but of those only over expression of the MarR family TF Rv2887 resulted in repression of Rv0560c (Supplementary Fig. 8). However, the other TFs are strong candidates to regulate Rv0560c in other contexts. Mapping TF-DNA binding and expression changes in other environments should expand further the list of interactions with corresponding identifiable downstream expression changes[6].

While 5,400 DNA binding events are located in the promoter window, roughly 66% (>10,500) of binding sites are outside this region. Altogether 109 different TFs exhibit promoter-distal binding. While there are examples of prokaryotic proteins binding outside of promoters and exerting regulatory effect at a distance (Figure 2a, and[32–35]), as a class these binding events are less likely to exert direct influence on gene expression. To explore globally the link between TF-DNA binding and transcription, we compared the number of binding events per TF and the number of expression changes associated with each TF (Figure 2b). Of 178 MTB TFs in this study, nearly 40% exhibit an approximately linear relationship between the number of DNA binding events and transcriptional changes. Two of the most well-characterized DNA binding proteins in MTB (Rv3133c/DosR[8] and Rv3849/EspR[14]) behave this way. For roughly 30% of proteins, induction is associated with a disproportionately large impact on transcription relative to the number of binding sites. These proteins may regulate other TFs and initiate a transcriptional cascade. Alternatively some of these TFs may be poor candidates for ChIP analysis. In contrast, there are approximately the same number of proteins whose induction results in prolific DNA binding but comparatively few transcriptional changes. The regulatory circuits of these genes may be complex, perhaps requiring one or more partner TF(s) or another co-factor to reconcile DNA binding and expression profiles. These proteins belong to a wide range of TF families, including TetR, ArsR, and GntR, along with one nucleoid associated protein Lsr2.

Identifying DNA consensus motifs from ChIP-seq data

We searched for conserved motif signatures for each TF. We queried all DNA binding data using MEME[36] and default parameters. We performed each motif search twice for each grouping – one unconstrained and one constrained to detect only palindromes. After filtering motifs for MEME E-values (E <= 1) and peak locations within the queried sequence (p <= 0.05) we could identify significant motifs for a total of 57 (71%) out of the 80 TFs that had ≥14 ChIP-seq peaks. We report the two motifs detected for each TF, along with all related statistics, in Supplementary Table 4. TFs with a greater number of binding sites were more likely to have an identified consensus motif. The average number of binding sites for TFs with a motif was 246 (range 14 to 859), compared to an average of 28 peaks (range 3 to 437) for those TFs where a significant consensus motif could not be identified. For TFs with previously characterized DNA binding motifs, this analysis corresponded well with previous reports (eg. – Rv2506[37,38], Rv2359[39], DosR[8], KstR[9], and EspR[14]). In cases where the data set was of sufficient size to parse by location within or outside of a promoter, the identified consensus motifs tend to share the dominant sequence features of the motif derived from the aggregate sequences (for example, Rv1255c); however, in this context subtle sequence variations are likely to have functional consequences.

Rv0494 as an example of widespread binding

As indicated above, about 30% of the TFs in this study bind prolifically around the chromosome both within and outside of promoters, but affect relatively few transcriptional changes. To investigate this behavior, we focused on a representative member, Rv0494 (Figure 3). Rv0494 is a GntR-family regulator[40,41] whose induction correlated with 10 transcriptional changes at 7 genomic loci (Figure 3, blue-red ring) including binding at the Rv3094c-Rv3095 locus (Figure 3, grey ribbon); however, there are 77 Rv0494 binding events distributed around the MTB chromosome (Figure 3 – internal lines). DNA pattern searching using MEME[36] on the entire data set yielded two significant consensus motifs (Supplementary Table 4). We observed that the Rv0494-bound regions contributing to the longer (17mer) motif have more significant ChIP binding scores, whereas the bound regions contributing to the shorter (~9mer) motif have strong but less significant scores. We stratified ChIP binding sites by score and searched for consensus motifs in two tranches: p<0.001 (higher peak quality scores; 36 input regions, purple lines in Figure 3) and 0.001 < p < 0.01 (lower peak quality scores; 41 regions, yellow lines in Figure 3). We saw a striking division in the consensus motifs derived. Of the 36 highly significant binding sites, 35 contained a close variant of the 17-mer consensus motif (motif E-value = 8.4×10^{-51}, Figure 3, purple ribbon). Of the 41 less significant bound sequences, 28 contained the 9-mer consensus motif (motif E-value = 1.7×10^{-31}, Figure 3, yellow ribbon). Combining the bound regions that did not contribute to either motif initially, we found that these peaks had p-values in the middle of the distribution (0.0015 < p < 0.004, Supplementary Fig. 9). Repeating the MEME pattern search on these 14 regions showed that 13 sites contained a close variant of the 17-mer consensus motif (motif E-value = 8.3×10^{-5}).

We next analyzed expression from Rv0494-induced cultures. Of the 10 differentially expressed genes following Rv0494 induction, two of these loci (6 genes) are immediately adjacent to Rv0494 binding sites. These are strong candidates for direct regulation by

Rv0494, and both these loci show highly significant binding (Figure 3, grey ribbon, and Supplementary Table 3). However, we also find examples of binding to the strong consensus motif with no obviously associated change in gene expression.

Validating Rv0494 binding to different motifs

From these analyses, the vast majority of Rv0494 binding sites – 76 of 77 bound regions – are described by one of two consensus motifs. We sought to validate this binding by an alternate approach. Employing purified, recombinant Rv0494 protein we developed a "universal" electrophoretic mobility shift assay (uEMSA) in which a uniform DNA scaffold was modified to contain a 5′ IR680 (red) or IR800 (green) IR tag (Figure 4a). This approach allows simultaneous visualization of target, non-specific, and specific competitor DNAs in an *in vitro* electrophoretic mobility shift assay[42]. The Rv3094c-Rv3095 intergenic region contains a variant of the 17-mer motif (Figure 3), and in uEMSA experiments both the 17-mer consensus motif and Rv3094c-Rv3095 DNA sequences are tightly bound by recombinant Rv0494 protein (Figure 4b). Binding is specific, as confirmed by a persistent gel shift in the face of 20× molar excess non-specific competitor DNA; however, in the face of 20× molar excess specific competitor, Rv0494 protein preferentially binds to the more abundant IR800-labeled competitor DNA. Rv0494 protein also showed specific binding to the 9-mer DNA consensus motif, though at a higher protein concentration than the 17-mer motif. We note that none of the 9-mer-Rv0494 interactions were associated with detectable changes in proximal gene expression, indicating that such binding events can nonetheless be validated by alternate means. Altogether, these data indicate that consensus DNA binding motifs derived from ChIP-seq can be validated by alternate experimental methods, and demonstrate a correlation between ChIP peak quality score and protein-DNA affinity.

DISCUSSION

Robert Koch described the cause of tuberculosis more than a century ago yet MTB remains a pervasive pathogen, infecting 30% of the world's population and causing 2 – 3 deaths every minute. To understand better how MTB adapts within the human host we undertook a systematic characterization of the gene regulatory network. We ectopically induced expression of epitope-tagged copies of nearly every DNA binding protein in MTB (Supplementary Fig. 1). Using this approach we performed ChIP-seq and transcriptional profiling under a uniform condition for 178 TFs. We filtered the binding patterns of experimental samples against a robust negative control peak set, and imposed a stringent significance threshold for inclusion of DNA binding events in downstream analyses. We also associated binding with gene expression changes, incorporating transcriptional data generated under the same experimental conditions. These data provide an in-depth, system-wide view of the DNA binding network in this important bacterial pathogen.

The MTB DNA binding network consists of ~16,000 protein-DNA interactions from 154 genes that passed our stringent filter set (Figure 1). We could not identify consistent attributes to define the 24 proteins that did not bind DNA, and we hypothesize that these proteins require additional signals or modifications to bind the chromosome. We also noted prolific binders – 7 proteins with >500 binding sites each. MEME pattern searching analysis

revealed significant consensus motifs for each of these proteins, which suggests that prolific binding was still dictated by sequence-specific DNA interactions (Supplementary Table 4). The number of binding events per protein could be fit to a power law distribution ($p(k) \sim k^{-1.5}$), with half of the binding coming from 14 proteins and ~90% of the binding from 44 proteins (~25% of all assayed binding proteins, Figure 1c). However, from the perspective of the DNA the chromosome is sparsely bound. More than 85% of the genome bound no TFs, while slightly more than 10% of the genome bound a single TF (Figure 1b). A few loci (~2.5%) were hotspots for binding, and these are prime candidates for combinatorial protein-DNA interactions.

Genome-wide, TF binding was non-random, and we identified significant consensus motifs for 57 TFs (Supplementary Table 4). Furthermore, we observed more than twice as much binding in intergenic regions than would be expected by random chance. Similarly, we found a striking enrichment of TF binding within −150 to +70 nucleotides of annotated start sites (CDS or TSS), with the greatest enrichment in the 0 to −20 region. Altogether, we found approximately one third (5,400 of 15,980) of TF binding sites were within one or more 220 bp promoter windows, resulting in >7,200 TF-promoter interactions (Figure 2a and Supplementary Table 3). More than 1,150 of these binding events were associated with altered gene expression in our experiments, and in the appropriate environmental context, many more of these >7,200 interactions are likely to serve a proximal regulatory function. However, even more binding events (>10,500) were positioned outside of promoters. We observe some instances of promoter-distal binding correlated with proximal gene regulation (Figure 2a), and probably in alternate environmental contexts a greater number of these would act to alter expression of proximal genes. However, it is also likely that many of these promoter-distal binding sites are transcriptionally dormant. Abundant promoter distal binding has been noted before[32,33,43], and in some cases individual proteins that bind DNA prolifically have been shown to regulate transcription at a subset of their loci but not at others[14,16,34]. For instance, in MTB the TF EspR has been labeled both a specific transcription factor[44] and a nucleoid associated protein[14]. Our analysis provides evidence for both ideas. We find that EspR exhibits binding that is both widespread and promoter-proximal, and that only a fraction of binding events directly influence transcription. Furthermore, we observe similar behavior from the majority of TFs in MTB.

To examine further the phenomenon of widespread binding with limited regulation we focused on Rv0494, which binds 77 times and promotes altered expression at only 2 of these loci (Figure 3). We identified consensus motifs associated with both stronger and weaker binding and protein-DNA interactions could be validated by independent experimental approaches (Figure 4). Some Rv0494-dependent expression changes were proximal to strong binding events, however many strong binding events were not associated with any local gene expression changes.

Altogether our analyses both complement and contrast with current models of bacterial transcription. For example, we found numerous strong DNA binding consensus motifs (Supplementary Table 4) and robust enrichment for DNA binding in the window (−150 to +70) relative to transcription start sites (Figure 2a), in agreement with promoter studies in bacteria[45]. However, compared with the model bacterium *E. coli*, the MTB TF-DNA

binding network results were surprising in terms of binding site numbers, locations and effects. Transcription is well-studied in *E. coli*, with substantial information collected and curated at the online repository RegulonDB[46]. This database lists ~2400 *E. coli* TF-DNA binding events, nearly 7× fewer than we observe in MTB. Only 27 individual *E. coli* TFs are known to bind DNA more than 20 times, compared with 69 in MTB. Further, in MTB we find dozens of TFs with widespread binding and few downstream transcriptional changes.

How to reconcile these differences? We have considered the possibility that widespread DNA binding is an artifact of the ChIP approach. However, we have ruled out previously described artifacts such as spurious ChIP enrichment proximal to highly-transcribed loci (Supplementary Fig. 3), applied rigorous control filters (Methods), and our binding data are highly reproducible[17]. The FLAG-tagged TF overexpression and reference conditions that we employed could be sources of artifactual binding, but ChIP under physiological conditions with native antibodies also yield similar binding profiles[14,17]. Further, we have shown that our TF overexpression levels are less than or equal to TF gene expression changes in publically available array studies for over 80% of TFs (Supplementary Fig. 4b and[24]).

Another possibility is that most previous studies, which assess protein-DNA interactions at specific candidate sites, may consistently underestimate the actual extent of binding. Since early groundbreaking work with the Lac operon[47], researchers interested in transcriptional control have focused on individual gene expression changes and thus may have systematically understudied the possibility of transcriptionally dormant binding. In fact, the most common approach to determine TF regulatory targets is transcriptional profiling of a gene disruption mutant, which by definition precludes identification of such binding events. Widespread dormant binding could thus be a phenomenon specific to MTB, however several recent studies in both eukaryotes and prokaryotes used global approaches and reported unexpectedly widespread binding[14,16,48–50], including one study in *E. coli* that was not based on ChIP[43]. In addition, various effects of dormant binding have been reported, including association with chromosome organization, replication and cell division[33,43], altering response kinetics and dynamics at regulation-active loci through transcription factor titration and buffering against noisy input[51–53], suggesting multiple functional contexts for this phenomenon. These observations, in eukaryotes and archaea as well as bacteria, raise the possibility that widespread dormant binding is a common feature of transcriptional systems everywhere that should be considered when developing gene regulatory networks. The implications of these phenomena for MTB biology and for transcriptional control more broadly are largely unexplored, and warrant additional investigation.

Materials & Methods

Construction of Expression Vectors and Strains

Our in-house analysis indicated that there are 214 putative DNA binding genes in the *M. tuberculosis* genome. At the outset of this project we had at our disposal a Gateway Entry Clone library of ~2600 *M. tuberculosis* ORFs in the backbone of pDONR221 (PFGRC/Colorado State University under NIAID contract HHSN266200400091c). In the event that a putative DNA binding gene-of-interest was not included in the extant Entry Clone library,

we created entry clones through PCR amplification of the relevant gene template from H37Rv gDNA, adding the necessary Gateway recombination sequences to the PCR product. In total, 9 genes proved refractory to sub-cloning efforts, and so were triaged from subsequent analyses. Including those genes from the PFGRC entry clone library and our sub-cloning efforts the final putative DNA binding clone library contains 206 genes. We inserted each of these genes in to an *E. coli*-mycobacterial episomal shuttle vector modified to contain an anhydrotetracycline (ATc)- inducible promoter[54] and a Gateway cloning recombination cassette (kind gift of Eric Rubin). We further modified this vector to contain an N- or C-terminal FLAG epitope tag – amino acid sequence: n-DYKDDDDK-c. For the present work the C-terminal FLAG tagged version was used for all DNA binding experiments, with the exception of experiments utilizing Rv3133c/DosR, which contained the N-terminal FLAG tag. *M. tuberculosis* H37Rv strains containing these ATc-inducible, FLAG-tagged, expression vectors are available from BEI resources (nr-46512, www.beiresources.org).

Culturing Conditions

M. tuberculosis strain H37Rv was cultured in Middlebrook 7H9 with the ADC supplement (Difco) and 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 to maintain the plasmid. 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) 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. 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 (1:55 dilution) M2 anti-FLAG antibody (Sigma, F1804). 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.

ChIP-seq Peak Control Dataset

To determine significance thresholds for peak inclusion in our data set we generated a ChIP-seq control compendium consisting of 10 different sequencing data sets. Because no single control type captures all known or potential ChIP artifacts or biases we included an array of control types, including: wildtype H37Rv chromatin immunoprecipitated with and without anti-FLAG antibody, chromatin samples from uninduced expression-vector bearing cells immunopreciptated with and without anti-FLAG antibody, as well as chromatin samples from induced non-TF genes immunoprecipitated with anti-FLAG antibody.

Illumina Library Prep Sequencing

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

Read Alignment & Peak Calling

Peak calling was carried out using an in-house algorithm outlined in Supplementary Fig. 10 and available for download at http://networks.systemsbiology.net/mtb. Short reads were aligned to the H37Rv reference genome using Bowtie 0.12.7 with default parameters, resulting in 98% of reads being successfully aligned. Read pileups were converted to wiggle tracks for forward, reverse, and cumulative strands, and then searched for local extrema. We then estimated half width at half height for each local maximum (a *de facto* “peak”), and using nonlinear least squares optimization we found the optimal Gaussian or Gumbel model distribution that best fit the aligned reads. We assigned 0–1 scores based on relative height, width and drift from starting local maximum of each fitted peak. We then merged the forward, reverse, and cumulative results in to a “combo peak” and re-scored that triplet with the addition of score values for separation and relative heights of the forward and reverse strand peak centerpoints. The final score for a single ChIP peak was the product of [ScoreF * ScoreR * ScoreC * Sep * EqHts] on a 0–1 scale, with 1 being a “perfect” score.

Assigning Significance Scores to Called Peaks

Each sequencing data set was subjected to the “read alignment and peak calling” algorithm. To determine significance scores of experimental data, we collapsed the peaks (with respective scores) from the 10 control experiments described above into a single data set

containing 2027 scored “final” peaks as negative controls (Supplementary Table 5). For each scored peak in an experimental data set we measured the probability of identifying a comparably high scoring matched peak type in the control data set. Thus, an experimental peak with p = 0 indicates that no peaks in the negative control set had an equivalently robust score. Similarly, a peak with a p value of 0.01 has a peak quality score better than 99% of all peaks identified in the negative control set. A table with scoring metrics and significance scores for all DNA binding proteins assayed, all peaks, is provided in Supplementary Table 1.

RNA Isolation

RNA was isolated as described previously[55]. 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[27]. Briefly, 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[56]. 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=702). Altered gene expression was considered significant if it produced an empirical Bayes method p <0.01. Raw microarray data are available at the gene expression omnibus (GEO) in series GSE59086. Additional details can be found in[24].

Promoter window size analysis

Receiver operation curves (ROCs) were used for assessing the accuracy of promoter window sizes to associate binding with target regulation. Upstream promoter window sizes were tested every 10 nucleotides from −10 to −200 upstream of designated start sites and at varying nucleotide lengths to −1500 upstream. Similarly, window sizes were tested every 10 nucleotides from +10 to +200 downstream. The set of ChIP-seq binding events with target regulation was formed by instances within a given window size that a particular TF has a significant overlap of proximal gene targets and differentially expressed genes (as determined by Rustad et al[24]. The overlap was computed using hypergeometric enrichment

p-values. The ROC curves were formed by considering the overlap of each possible pairwise combination of TFs and measuring the sensitivity and specificity of the overlap, where sensitivity represents the fraction of differentially expressed target genes that had a binding peak within the promoter window, and specificity represents the fraction of non-differentially expressed target genes that did not have a binding peak within the promoter window. The R open-source package pROC was used to calculate AUC values of tests performed at each window size[57]. The optimal window size was determined by the largest AUC in the upstream and downstream regions and resulted in a −150:+70 window. As a result, genes targeted by a particular TF were identified by having a significant ChIP-Seq binding peak in the −150:+70 window of their start site or by being part of an operon with a binding site in the −150:+70 region of an upstream gene in the operon.

Identifying consensus motifs from ChIP-seq data

For consensus motif determination we searched for conserved DNA signatures within +/- 50 nucleotides of ChIP-seq peak centers using MEME[36]. Peaks were further subdivided into subsets that were only within or outside of our defined promoter windows (thus there are three subsets for each TF – “all” significant peaks, those “in” promoters, and those “out”). We performed each motif search twice for each grouping – one unconstrained and one constrained to detect only palindromes. For each motif, we computed the significance of the distribution of its locations relative to the corresponding peak centers, relative to a uniform null distribution, using the Komolgorov-Smirnov (KS) test. We also scanned each motif in an unbiased manner across the entire genome using FIMO[58] and computed whether these scanned locations were significantly located within +/- 50 nt of the corresponding ChIP-seq peak locations relative to randomly sampled locations throughout the genome. Motifs with MEME E-values (E <=1) and peak location (p <=0.05) were considered significant.

Recombinant Rv0494 Protein Purification

The Rv0494 CDS was subcloned in to the pET28b expression vector (Novagen/EMD Millipore). The Rv0494 locus was PCR amplified from purified H37Rv gDNA, adding an XhoI restriction endonuclease site to the 3′ end of the cassette. The primer specific to the 5′ end of the gene cassette contained an NdeI RE site as well as the recognition motif for the HRV 3C protease. After ligation pET28-Rv0494Ab inducibly expressed Rv0494 with an N-terminal 6x-HIS tag upstream of the HRV 3C cut site and native Rv0494 sequence. For recombinant protein production we transformed BL21(DE3) *E. coli* with pET28-Rv0494Ab. Cultures were grown to an OD600 of 0.5 in Terrific Broth before treatment with 100 μM IPTG shaking overnight at 18°C. Following sonication, recombinant protein was recovered from crude lysates by FPLC metal affinity chromatography and size exclusion chromatography. To remove the 6x-HIS tag from the final protein product, recombinant Rv0494 was subjected to HRV 3C protease (Novagen) digestion (rotating overnight at 4°C). Following cleavage, Rv0494 solutions were again passed over a metal affinity column to remove the liberated epitope tag and the HIS-tagged protease. Final purification was effected through size exclusion chromatography. Protein aliquots were snap frozen in storage buffer (150 mM NaCl, 20 mM Tris-HCl – pH 7.5, 5% glycerol) and kept at −80°C for subsequent applications. The purified 26kDa Rv0494 protein contains two non-native amino acids at the N-terminus.

Universal Electrophoretic Mobility Shift Assays

Similar to the technique described in[42], for universal electrophoretic mobility shift assays (uEMSAs), 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 the Rv3094c-Rv3095 intergenic region in the *M. tuberculosis* genome, or the consensus motif sequences flanked by GC-matched randomized nucleotides. 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 IR680 or IR800 infrared dye covalently coupled to the 5′ end. The IR680 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 recombinant Rv0494 protein was removed from storage buffer (150 mM NaCl, 20 mM Tris-HCl – pH 7.5, 5% glycerol) and exchanged to sterile-filtered reaction buffer (10 mM Tris-HCl – pH 8.0, 10 mM NaCl, 1 mM DTT, 1 mM EDTA, 5 ng/μL BSA) using a 10kDa-cutoff spin column (Amicon). In the present study, Rv0494 was present at 0.1 μM for the Rv3094c-Rv3095 and 17-mer consensus motif uEMSA experiments. Rv0494 was present at 2.0 μM for the 9-mer consensus motif uEMSA experiments. 50 nM specific, IR680-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 = 1 μM). 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project has been funded with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Dept. of Health and Human Services under contract HHSN272200800059C and grant U19 AI106761, as well as The Camille Dreyfus Teacher-Scholar Award Program (NDP) and the NIH Center for Systems Biology /2P50GM076547 (NSB, NDP). KJM acknowledges NIH Training Grant T32AI007509. SM acknowledges the NSF Graduate Research Fellowship DGE-1144245. We thank Antonio Frandi and Daniel Wozniak as well as members of the Sherman and Baliga labs for helpful input, and the anonymous reviewers for helping to produce a stronger manuscript.

References

1. WHO. Geneva: 2013.

2. Balázsi GHAP, Shi L, Gennaro ML. The temporal response of the *Mycobacterium tuberculosis* gene regulatory network during growth arrest. Mol Syst Biol. 2008; 4:225. [PubMed: 18985025]
3. Rohde KH, Veiga DF, Caldwell S, Balazsi G, Russell DG. Linking the transcriptional profiles and the physiological states of Mycobacterium tuberculosis during an extended intracellular infection. PLoS Pathog. 2012; 8:e1002769. doi:10.1371/journal.ppat.1002769 PPATHOGENS-D-11-02225 [pii]. [PubMed: 22737072]
4. Beste DJ, et al. (13)C metabolic flux analysis identifies an unusual route for pyruvate dissimilation in mycobacteria which requires isocitrate lyase and carbon dioxide fixation. PLoS Pathog. 2011; 7:e1002091. doi:10.1371/journal.ppat.1002091 10-PLPA-RA-4128 [pii]. [PubMed: 21814509]
5. Jamshidi N, Palsson BO. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol. 2007; 1:26. doi:1752-0509-1-26 [pii] 10.1186/1752-0509-1-26. [PubMed: 17555602]
6. Peterson EJ, et al. A high-resolution network model for global gene regulation in Mycobacterium tuberculosis. Nucleic Acids Res. 2014 doi:gku777 [pii] 10.1093/nar/gku777.
7. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2010; 107:17845–17850. doi:1005139107 [pii] 10.1073/pnas.1005139107. [PubMed: 20876091]
8. Park HD, et al. Rv3133c/dosR is a transcription factor that mediates the hypoxic response of *Mycobacterium tuberculosis*. Mol Microbiol. 2003; 48:833–843. [PubMed: 12694625]
9. Kendall SL, et al. Cholesterol utilization in mycobacteria is controlled by two TetR-type transcriptional regulators: kstR and kstR2. Microbiology. 2010; 156:1362–1371. doi:mic. 0.034538-0 [pii] 10.1099/mic.0.034538-0. [PubMed: 20167624]
10. Kendall SL, et al. A highly conserved transcriptional repressor controls a large regulon involved in lipid degradation in Mycobacterium smegmatis and Mycobacterium tuberculosis. Mol Microbiol. 2007; 65:684–699. doi:MMI5827 [pii] 10.1111/j.1365-2958.2007.05827.x. [PubMed: 17635188]
11. Lun DS, Sherrid A, Weiner B, Sherman DR, Galagan JE. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. Genome Biol. 2009; 10:R142.10.1186/gb-2009-10-12-r142 [PubMed: 20028542]
12. Sala C, et al. Genome-wide regulon and crystal structure of BlaI (Rv1846c) from Mycobacterium tuberculosis. Mol Microbiol. 2009; 71:1102–1116.10.1111/j.1365-2958.2008.06583.x [PubMed: 19154333]
13. Gordon BR, 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; 107:5154–5159.10.1073/pnas.0913551107 [PubMed: 20133735]
14. Blasco B, et al. Virulence regulator EspR of Mycobacterium tuberculosis is a nucleoid-associated protein. PLoS Pathog. 2012; 8:e1002621.10.1371/journal.ppat.1002621 [PubMed: 22479184]
15. Smollett KL, et al. Global analysis of the regulon of the transcriptional repressor LexA, a key component of SOS response in Mycobacterium tuberculosis. J Biol Chem. 2012; 287:22004–22014. doi:10.1074/jbc.M112.357715 M112.357715 [pii]. [PubMed: 22528497]
16. Kahramanoglou C, et al. Genomic mapping of cAMP receptor protein (CRP Mt) in Mycobacterium tuberculosis: relation to transcriptional start sites and the role of CRPMt as a transcription factor. Nucleic Acids Res. 2014; 42:8320–8329. doi:10.1093/nar/gku548 gku548 [pii]. [PubMed: 24957601]
17. Galagan JE, et al. The Mycobacterium tuberculosis regulatory network and hypoxia. Nature. 2013; 499:178–183. doi:10.1038/nature12337 nature12337 [pii]. [PubMed: 23823726]
18. Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. Nature. 2000; 406:378–382.10.1038/35019019 [PubMed: 10935628]
19. Milo R, et al. Network motifs: simple building blocks of complex networks. Science. 2002; 298:824–827. doi:10.1126/science.298.5594.824 298/5594/824 [pii]. [PubMed: 12399590]
20. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet. 2002; 31:64–68. doi:10.1038/ng881 ng881 [pii]. [PubMed: 11967538]
21. Lew JM, Kapopoulou A, Jones LM, Cole ST. TubercuList–10 years after. Tuberculosis (Edinb). 2011; 91:1–7.10.1016/j.tube.2010.09.008 [PubMed: 20980199]

22. Galagan JE, et al. TB database 2010: overview and update. Tuberculosis (Edinb). 2010; 90:225–235. doi:S1472-9792(10)00041-7 [pii]10.1016/j.tube.2010.03.010. [PubMed: 20488753]
23. Gillespie JJ, et al. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect Immun. 2011; 79:4286–4298.10.1128/IAI.00207-11 [PubMed: 21896772]
24. Rustad TR, et al. Mapping and manipulating the MTB transcriptome using a transcription factor overexpression derived regulatory network. Genome Biology. 2014 In Press.
25. Teytelman L, Thurtle DM, Rine J, van Oudenaarden A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc Natl Acad Sci U S A. 2013; 110:18602–18607.10.1073/pnas.1316064110 [PubMed: 24173036]
26. Park D, Lee Y, Bhupindersingh G, Iyer VR. Widespread misinterpretable ChIP-seq bias in yeast. PloS One. 2013; 8:e83506.10.1371/journal.pone.0083506 [PubMed: 24349523]
27. Rustad TR, Harrell MI, Liao R, Sherman DR. The enduring hypoxic response of Mycobacterium tuberculosis. PloS One. 2008; 3:e1502.10.1371/journal.pone.0001502 [PubMed: 18231589]
28. Fitzgerald DM, Bonocora RP, Wade JT. Comprehensive Mapping of the Escherichia coli Flagellar Regulatory Network. PLoS genetics. 2014; 10:e1004649. doi:10.1371/journal.pgen.1004649 PGENETICS-D-14-00988 [pii]. [PubMed: 25275371]
29. Jones CJ, et al. ChIP-Seq and RNA-Seq reveal an AmrZ-mediated mechanism for cyclic di-GMP synthesis and biofilm development by Pseudomonas aeruginosa. PLoS Pathog. 2014; 10:e1003984. doi:10.1371/journal.ppat.1003984 PPATHOGENS-D-13-02008 [pii]. [PubMed: 24603766]
30. Cortes T, et al. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis. Cell reports. 2013; 5:1121–1131.10.1016/j.celrep.2013.10.031 [PubMed: 24268774]
31. Schuessler DL, Parish T. The promoter of Rv0560c is induced by salicylate and structurally-related compounds in Mycobacterium tuberculosis. PloS One. 2012; 7:e34471.10.1371/journal.pone.0034471 [PubMed: 22485172]
32. Vora T, Hottes AK, Tavazoie S. Protein occupancy landscape of a bacterial genome. Mol Cell. 2009; 35:247–253. doi:10.1016/j.molcel.2009.06.035 S1097-2765(09)00479-1 [pii]. [PubMed: 19647521]
33. Browning DF, Grainger DC, Busby SJ. Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. Curr Opin Microbiol. 2010; 13:773–780. doi:10.1016/j.mib.2010.09.013 S1369-5274(10)00141-4 [pii]. [PubMed: 20951079]
34. Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Microbiol. 2010; 8:185–195. doi:10.1038/nrmicro2261 nrmicro2261 [pii]. [PubMed: 20140026]
35. Buxton RS, et al. Long range transcriptional control of virulence critical genes in Mycobacterium tuberculosis by nucleoid-associated proteins? Virulence. 2012; 3:408–410. doi:10.1128/JB.00142-12 20918 [pii]10.4161/viru.20918. [PubMed: 22722242]
36. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994; 2:28–36. [PubMed: 7584402]
37. Anand S, et al. Equilibrium binding and kinetic characterization of putative tetracycline repressor family transcription regulator Fad35R from Mycobacterium tuberculosis. FEBS J. 2012; 279:3214–3228.10.1111/j.1742-4658.2012.08707.x [PubMed: 22805491]
38. Balhana RJ, et al. bkaR is a TetR-type repressor that controls an operon associated with branched-chain keto-acid metabolism in Mycobacteria. FEMS microbiology letters. 2013; 345:132–140.10.1111/1574-6968.12196 [PubMed: 23763300]
39. Maciag A, et al. Global analysis of the Mycobacterium tuberculosis Zur (FurB) regulon. J Bacteriol. 2007; 189:730–740.10.1128/JB.01190-06 [PubMed: 17098899]
40. Vindal V, Suma K, Ranjan A. GntR family of regulators in Mycobacterium smegmatis: a sequence and structure based characterization. BMC genomics. 2007; 8:289.10.1186/1471-2164-8-289 [PubMed: 17714599]

41. Biswas RK, et al. Identification and characterization of Rv0494: a fatty acid-responsive protein of the GntR/FadR family from Mycobacterium tuberculosis. Microbiology. 2013; 159:913–923.10.1099/mic.0.066654-0 [PubMed: 23475950]
42. Jullien N, Herman JP. LUEGO: a cost and time saving gel shift procedure. Biotechniques. 2011; 51:267–269.10.2144/000113751 [PubMed: 21988693]
43. Ishihama A. Prokaryotic genome regulation: a revolutionary paradigm. Proceedings of the Japan Academy Series B, Physical and biological sciences. 2012; 88:485–508. doi:DN/JST.JSTAGE/pjab/88.485 [pii].
44. Raghavan S, Manzanillo P, Chan K, Dovey C, Cox JS. Secreted transcription factor controls Mycobacterium tuberculosis virulence. Nature. 2008; 454:717–721. doi:10.1038/nature07219 nature07219 [pii]. [PubMed: 18685700]
45. Collado-Vides J, Magasanik B, Gralla JD. Control site location and transcriptional regulation in Escherichia coli. Microbiol Rev. 1991; 55:371–394. [PubMed: 1943993]
46. Salgado H, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. 2013; 41:D203–213. doi:10.1093/nar/gks1201 gks1201 [pii]. [PubMed: 23203884]
47. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. Journal of molecular biology. 1961; 3:318–356. [PubMed: 13718526]
48. Wade JT, Struhl K, Busby SJ, Grainger DC. Genomic analysis of protein-DNA interactions in bacteria: insights into transcription and chromosome organization. Mol Microbiol. 2007; 65:21–26. [PubMed: 17581117]
49. MacQuarrie KL, Fong AP, Morse RH, Tapscott SJ. Genome-wide transcription factor binding: beyond direct target regulation. Trends Genet. 2011; 27:141–148. doi:10.1016/j.tig.2011.01.001 S0168-9525(11)00002-3 [pii]. [PubMed: 21295369]
50. Park DM, Akhtar MS, Ansari AZ, Landick R, Kiley PJ. The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. PLoS genetics. 2013; 9:e1003839. doi:10.1371/journal.pgen.1003839 PGENETICS-D-13-01251 [pii]. [PubMed: 24146625]
51. Brewster RC, et al. The transcription factor titration effect dictates level of gene expression. Cell. 2014; 156:1312–1323.10.1016/j.cell.2014.02.022 [PubMed: 24612990]
52. Lee TH, Maheshri N. A regulatory role for repeated decoy transcription factor binding sites in target gene expression. Mol Syst Biol. 2012; 8:576.10.1038/msb.2012.7 [PubMed: 22453733]
53. Burger A, Walczak AM, Wolynes PG. Abduction and asylum in the lives of transcription factors. Proc Natl Acad Sci U S A. 2010; 107:4016–4021. doi:10.1073/pnas.0915138107 0915138107 [pii]. [PubMed: 20160109]
54. Ehrt S, et al. Controlling gene expression in mycobacteria with anhydrotetracycline and Tet repressor. Nucleic Acids Res. 2005; 33:e21. [PubMed: 15687379]
55. Minch K, Rustad T, Sherman DR. Mycobacterium tuberculosis Growth following Aerobic Expression of the DosR Regulon. PloS One. 2012; 7:e35935. doi:10.1371/journal.pone.0035935 PONE-D-11-23514 [pii]. [PubMed: 22558276]
56. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003; 19:185–193. [PubMed: 12538238]
57. Robin X, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12:77. doi:10.1186/1471-2105-12-77 1471-2105-12-77 [pii]. [PubMed: 21414208]
58. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; 27:1017–1018. doi:10.1093/bioinformatics/btr064 btr064 [pii]. [PubMed: 21330290]
59. Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19:1639–1645. doi:10.1101/gr.092759.109 gr.092759.109 [pii]. [PubMed: 19541911]

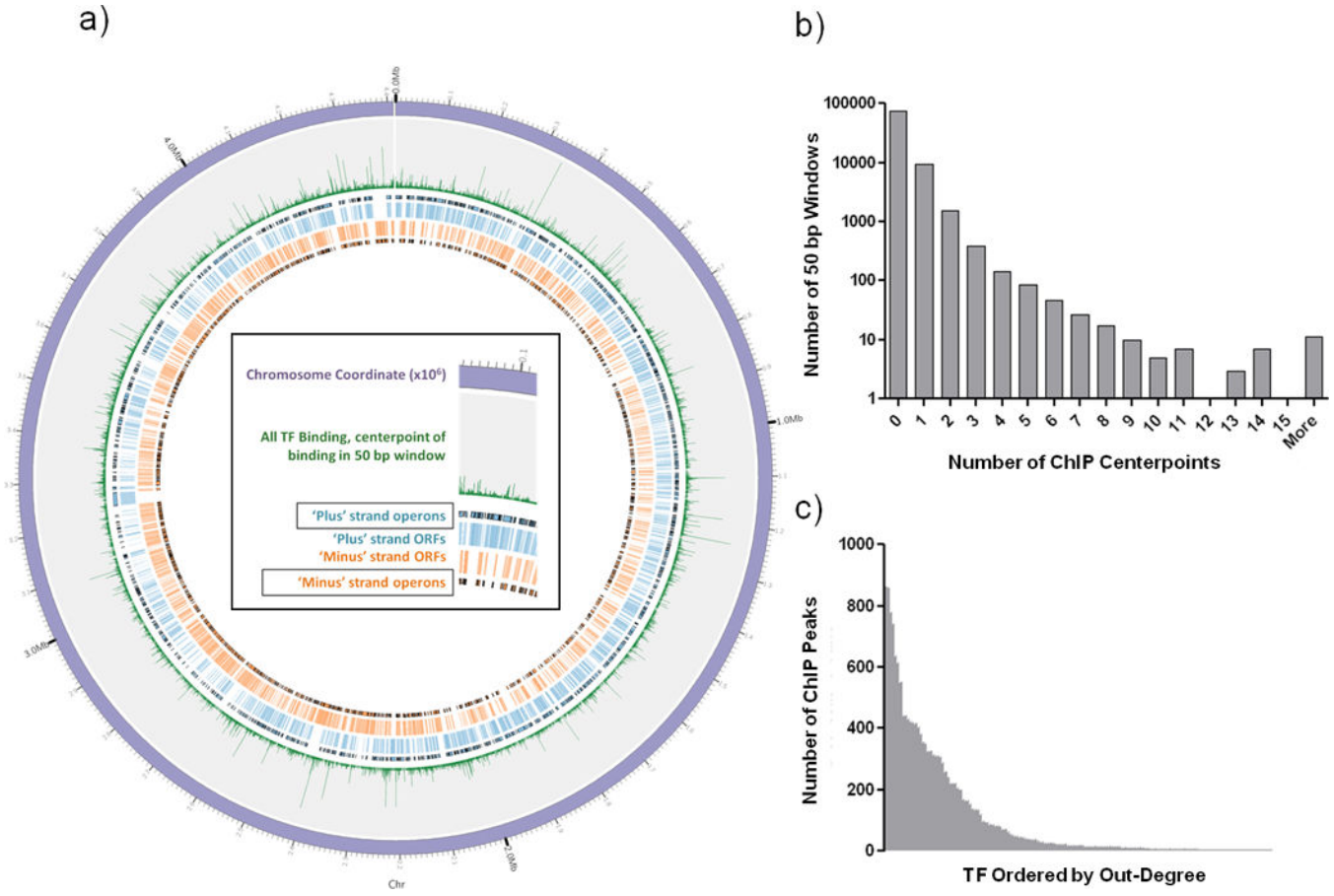

Figure 1. A global view of DNA binding

A) TF binding sites identified by ChIP-seq plotted with Circos[59]. Sense (blue) and antisense (orange) CDS and operon boundaries illustrated with black edges. The 4.4 Mb H37Rv chromosome is divided in to non-overlapping 50bp windows, and green spikes represent the total number of TF binding events within each window. **B)** Histogram of number of TF binding events per 50 bp window. **C)** Number of ChIP binding events (out-degree) for each of the 156 DNA binding proteins with at least one binding site.

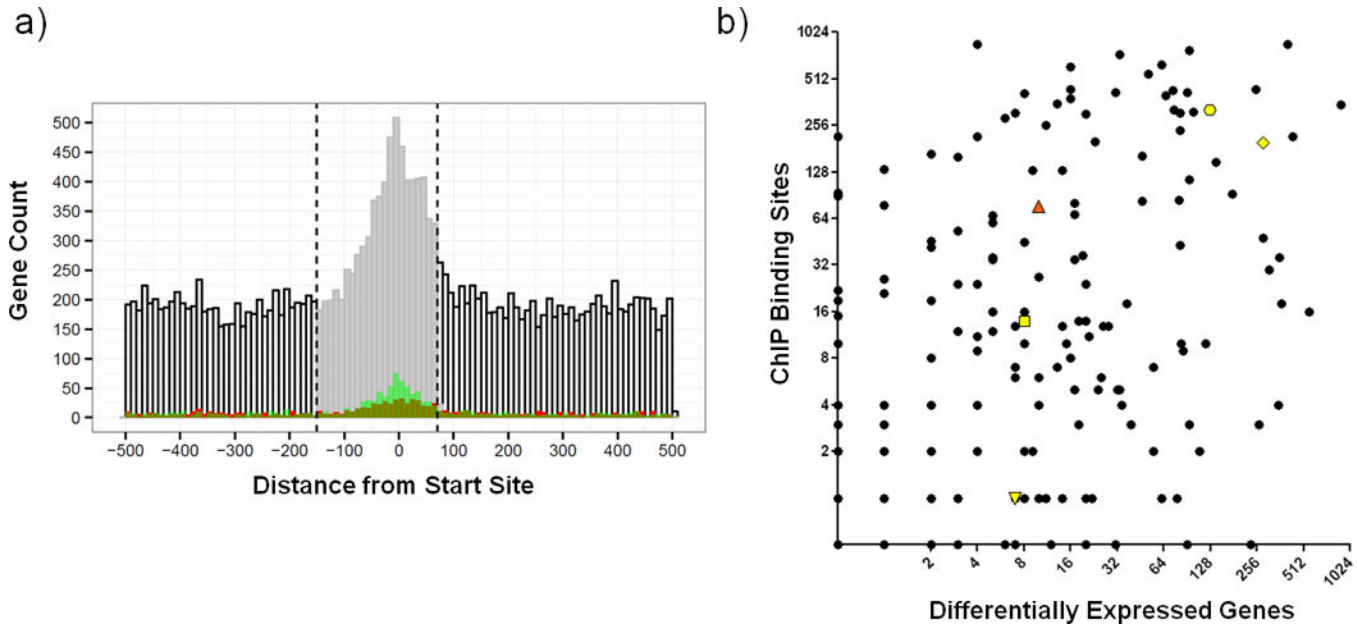

Figure 2. Global view of DNA binding and transcriptional regulation in MTB

A) Plot of binding distribution in 1kb nucleotide window (−500 to +500) surrounding CDS or transcription start sites. ROC/AUC analysis indicated that the optimal promoter window size is −150 to + 70 nucleotides (indicated by vertical dashed lines and shading of histogram). Binding events correlated with 1.5-fold induction or repression in the over-expression dataset[24] are depicted in red and green, respectively. **B)** The relationship between the number of binding events detected vs. the number of transcriptional changes associated with induction of each TF in this study. Rv0494 (orange triangle) is an example of a TF with prolific binding but limited differential gene expression. Additional genes discussed in text highlighted in yellow with symbols as follows: Rv1657/ArgR (downward triangle), Rv1846v/BlaI (square), Rv3133c/DosR (octagon), Rv3849/EspR (diamond).

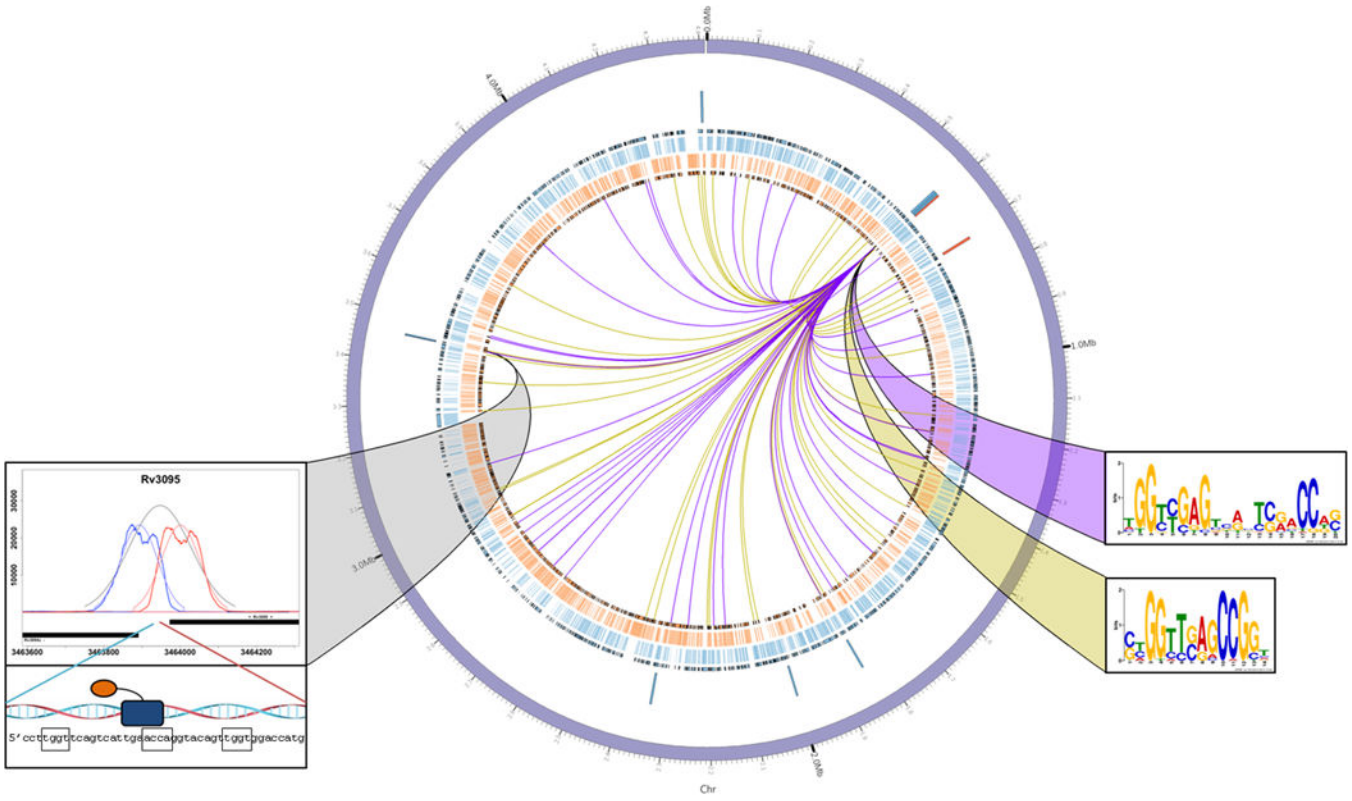

Figure 3. Rv0494 as an example of widespread binding and local gene regulation

Rv0494 binding and regulation of gene expression was plotted using Circos[59], with sense (blue) and antisense (orange) CDS and operon boundaries illustrated with black edges. The Rv0494 gene locus is positioned at ~2 o'clock, and ChIP binding sites are denoted at the terminus of each edge radiating out from Rv0494 (peaks of $p<0.001$ indicated by purple lines, $0.001<p<0.01$ indicated by yellow lines). Rv0494 consensus motifs corresponding to peak significance thresholds are indicated by the color-matched ribbons. Genes that exhibit significant differential regulation (>2-fold change with empirical Bayes method $p<0.01$ from 5 biological replicate microarrays) upon induction of Rv0494 are indicated by blue (gene repression) and red (gene induction) bars. The Rv0494 ChIP binding site between regulated genes Rv3094c-Rv3095 is shown connected by a gray ribbon.

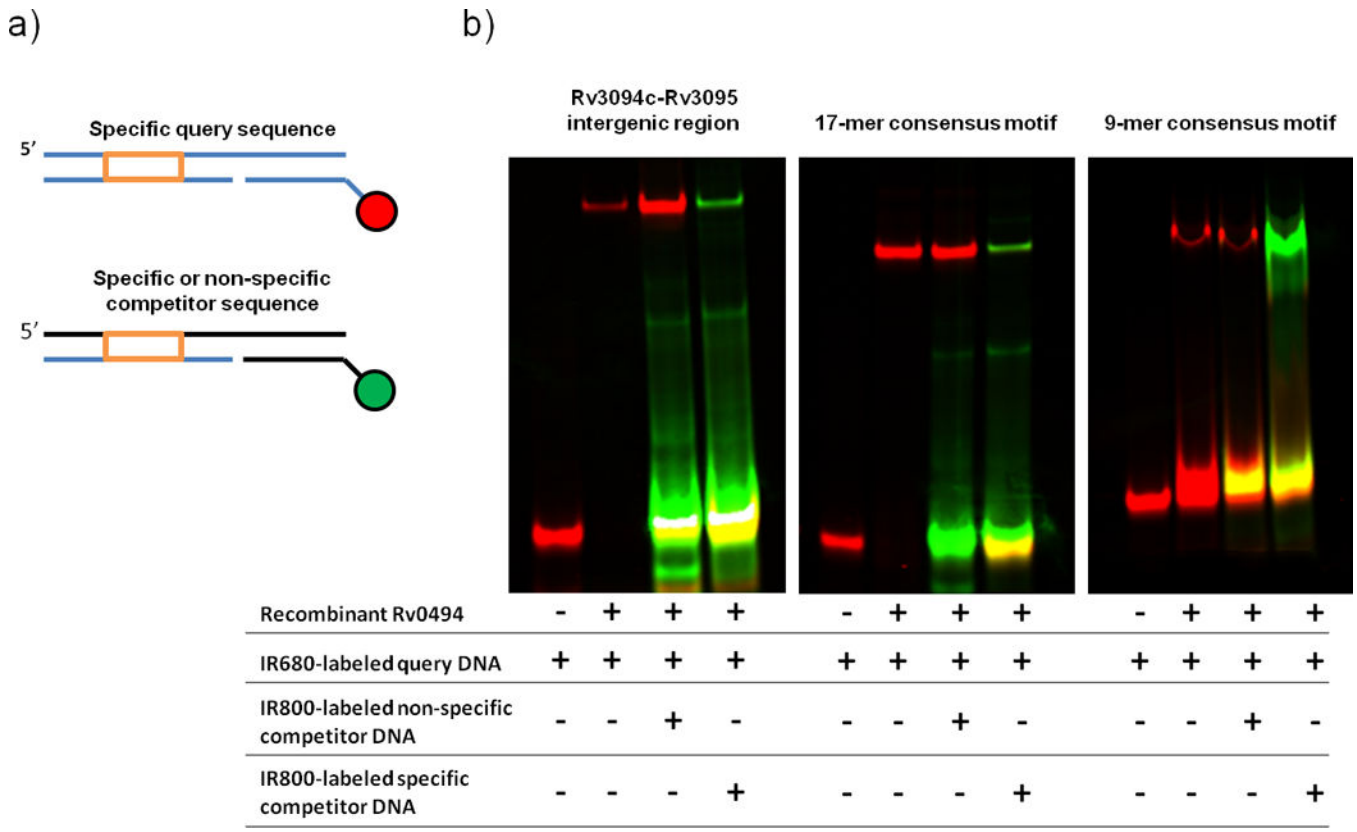

Figure 4. Universal electrophoretic mobility shift assay (uEMSA) validation of Rv0494 binding.

A) Schematic of DNAs used in uEMSA experiments. Three DNAs are annealed to form a single dsDNA product: a specific query sequence (orange box) is annealed in a 3-piece dsDNA fragment to a unique 12-mer sequence covalently coupled to a reporter dye. In these experiments, the specific query DNA was labeled with IR680 (red) and specific or non-specific competitor DNAs were labeled with IR800 (green). **B)** Purified recombinant Rv0494 binds specifically to ChIP-identified wildtype sequence (left panel), the 17-mer consensus motif (middle panel), and the 9-mer consensus motif (right panel). In the absence of protein, dye-coupled DNA does not shift (lane 1); however, the protein-DNA complex runs at a higher molecular weight (lane 2). This protein-DNA complex persists in the face of 20x molar excess green-labeled non-specific competitor DNA (lane 3), but can be outcompeted by the addition of 20x molar excess green-labeled specific competitor DNA (lane 4).