doi:10.1038/nature12337

Contents

1. Supplementary Text and References

To accommodate the limit on references in the main text, below, in some cases, we have included the same text as appears in the main manuscript but with references included.

1.1. Associating Binding with Regulation

As with previous reports[1], we cannot assign regulatory roles to all detected binding sites (Figure S13). Some of the binding we observe likely represent false positives, as noted above, or is not functional, as has been suggested in other organisms[2]. Moreover, as estimated by the FDR calculations, some assignments to binding sites likely reflect false positives arising from the large number of binding sites, as well as indirect regulatory effects.

However, our method may also 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[3]. 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 in 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[4] (as recently reported for the MTB factor EspR[5-7]), which would not necessarily be detected by our method. Fourth, regulation might be obscured by combinatorial interactions or the need to recruit additional factors to the binding site. Finally, binding may serve other functional roles including the maintenance of regulator concentration[1], the recruitment of other factors[8], modulation of chromatin structure[9], or alterations in the affinity of nearby sites as has been shown for DosR[10].

Additional details on the validation of ChIP-Seq binding is provided below in Section 2.4.

1.2. Topology of the MTB regulatory network model

Consistent with known examples in MTB[11,12] and *Corynebacterium glutamicum*[13], most factors are predicted to have both inducing and repressing roles. The TB regulatory network model also displays topological features seen in regulatory networks for other organisms[14-16]. The TB regulatory network model has topological features seen in networks for other organisms. The number of genes with which individual TFs interact can be roughly fit to a power law distribution ($p(k) \sim k^{-2}$) where a few TFs, or "hubs", interact with many genes, while most TFs 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. A total of 28 (56%) of the mapped regulators display auto-binding (compared to 28% of regulators in *E. coli* of which 85% are negative autoregulators[14,17]), and we also identify other motifs commonly found in biological networks, including over 900 predicted 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[14,18-22].

1.3. Rv0081 and Rv3597 (Lsr2) are interacting hubs.

Surprisingly Rv0081 forms the largest hub identified for the TFs reported. Rv0081 is part of the initial hypoxic response[23,24], but has been little studied. Rv0081 binds at 560 regions with 880 predicted binding locations, and its overexpression differentially regulates >400 genes equally split between activation and repression. Rv0081 also displays a statistically significant overlap ($p<1e-12$) with sites from another hub, Lsr2 (Rv3597), binds in the Lsr2 promoter, and is predicted to repress Lsr2. Lsr2 is an MTB analog of the H-NS nucleoid binding protein[25-27] (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 associated with high AT regions as previously reported by ChIP-chip[25]. Lsr2 alters chromatin[28] structure through DNA looping and thus likely modulates binding of other factors. 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).

1.4. A core subnetwork linking hypoxia, 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. Of particular interest is a subnetwork involving responses to altered oxygen status and lipid availability. 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[29]. KstR (Rv3574) controls a large regulon mediating cholesterol degradation as well as lipid and energy metabolism[30,31]. KstR was also identified as part of the EHR[29], 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 a 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[29]. 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, Rv0767c and PhoP (Rv0757). Rv2034 is an EHR regulator, and is predicted to activate DosR, thus providing possible positive feedback from the enduring to the initial hypoxic response (during revision, the link between Rv2034 and DosR was confirmed in an independent publication[32]). PhoP (Rv0757) mediates a range of responses including up-regulating hypoxia adaptation genes and DosR[33-35] 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 genetic link and supporting the conclusion that regulation of hypoxia adaptation by PhoP is indirect through this connection[33]. PhoP also mediates pH adaptation and our data confirm direct binding between PhoP and the aprABC locus required for this[36]. 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[37,38]. In addition to PhoP, both Rv0081 and Lsr2 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.

1.5. KstR de-repression and cellular lipid changes during hypoxia.

Consistent with predictions of the regulatory network during hypoxia we found strong induction of genes associated with lipid catabolism and cholesterol degradation, including the regulator KstR (Figure S18). KstR induction by hypoxia is predicted by the core regulatory network. However, KstR is a repressor[30] yet KstR-repressed cholesterol degradation genes are among those induced. KstR de-repression occurs during growth on cholesterol[31]. However no cholesterol or other exogenous lipids are present in our medium, suggesting additional factors or mechanisms that actively inhibit KstR-mediated transcriptional repression during hypoxia.

We explored the possibility that fatty acids endogenous to MTB or their metabolites might relieve KstR regulon repression. For this purpose we used the highly conserved KstR ortholog of *M. smegmatis*; when added individually to the growth medium, we found that fatty acids with carbon chain length from propionate (C3) to octanoate (C8), but not palmitate (C16) or oleate (C18), inhibited KstR repression nearly as well as cholesterol (Figure S23). This result led to the hypothesis that short or intermediate chain length fatty acids might be generated during hypoxia, resulting in the de-repression of KstR and the induced expression of cholesterol degradation genes. This possibility is being tested.

1.6. Accumulation and utilization of TAGs during hypoxia and re-aeration.

Triacylglycerides (TAGs) accumulation during hypoxia and in TB patient sputum samples, and their utilization upon re-aeration, has been reported[39-44]. We also observe accumulation of TAGs during hypoxia and rapid depletion during re-aeration. Our data provide a detailed systems view associated with these changes. Triacylglyceride synthase 1 activity is required for TAG accumulation[39,45] and we find that the primary TAG synthase gene, *tgs1*[46], is very highly expressed during hypoxia along with transient induction of *tgs4*. In addition, our data also reveal declines in expression during hypoxia of fatty acid and DAG synthase genes responsible for the reactions upstream of triacylglyceride synthase. Moreover, DAGs decrease dramatically during hypoxia and glycerol-3-P (G3P), a precursor of DAG, also decreases by day 7. Together, these data suggest a scenario in which metabolites upstream of DAG decrease in production, and TAG accumulation results from conversion of existing DAGs to TAGs via triacylglyceride synthase.

However, we also observe changes potentially related to TAG utilization. Of the 24 lipase genes predicted to cleave acyl groups from TAG[40], 18 show repression during hypoxia (data not shown but available at TBDB.org). During re-aeration, in contrast, the expression of most lipase genes return to baseline, and at the same time we observe a striking decrease of TAGs to nearly undetectable levels. Concurrently, we observe increases in the expression of genes involved in fatty acid synthase I (FAS1), energy metabolism and β-oxidation. We also observe further induction of methylcitrate cycle genes along with accumulation of methylcitrate during re-aeration. The utilization of TAGs is important for reactivation from dormancy in *M. bovis*[41]*,* and our data suggest a similar mobilization and utilization of stored TAGs in MTB.

Three lipase genes increase expression significantly and specifically during hypoxia, including *lipY* that has been implicated in mobilization of TAG stores during MTB starvation[72]. Together with the accumulation of long chain free fatty acid species that are most common in TAGs[47], this suggests some degradation of TAGs during hypoxia even as total cellular stores are increasing, though the data cannot exclude the possibility that LipY protein is stored during hypoxia in an inactive form for rapid mobilization of TAGs during re-aeration. The degradation, in turn, of liberated odd-chain fatty acids might also be a potential source of propionate.

The regulatory network identifies several regulatory links potentially relevant to the changes in TAG accumulation and utilization. Induction of *tgs1* by DosR is well established[23,39,45], and we identify this link. The network also identifies oxygen-responsive regulators of *tgs2* (Rv0081, Rv0324) and *tgs4* (DosR, Rv0324) and our models predict positive regulation of these genes in hypoxia by these TFs. Further, three of four lipase genes (Rv3176, Rv1169c, and Rv3097c) induced during hypoxia are influenced by regulators in the core network, and in these three cases we are able to predict their expression profiles using our gene expression models. The regulatory network thus connects alterations in lipid biosynthesis and utilization, and reflects an underlying hypoxia-driven program.

1.7. Alterations in the methylmalonyl pathway to branched chain lipids.

The methylmalonyl pathway converts propionate into methylmalonyl-CoA (MMCoA). MMCoA in turn can be metabolized into the TCA cycle through the action of a vitamin B12 dependent enzyme, or shunted directly into the production of methyl-branched lipids. The production of methyl-branched lipids can serve as sink for reducing equivalents and propionate. In our system, however, this sink does not appear to be activated during hypoxia, possibly exacerbating accumulation of propionate. During re-aeration, in contrast, modest methyl-branched lipids biosynthesis may result from reductants and intermediates produced by the degradation of TAG stores.

1.8. Alterations in amino acid metabolism and protein synthesis.

We measure 17 amino acids commonly found in proteins. Of these, 14 amino acids showed a trend of decreasing intracellular levels during hypoxia and increasing levels during re-aeration. In seven, this trend reached significance ($p<0.05$) in two replicates, while the remaining seven reach significance in one set of replicates (see Supplementary Material). Three amino acids, in contrast, showed increases during hypoxia (Supplementary Material). Surprisingly, nearly all amino acids also accumulate extracellularly throughout hypoxia. This is not likely due to lysis: extracellular phosphorylated sugars are a signature of cell lysis, and although measured intracellularly, they are not detected in our cell lysates. One consistent explanation for these data is the export of amino acids during hypoxia. We were unable to measure most amino acid catabolites, however, in two cases we detect amino acid break down intermediates indicating that degradation also occurs.

Global proteomic profiling suggests that changes in protein metabolism may also contribute to changes in amino acids levels. Discovery based proteomic profiling of 1000 of the most abundant cellular proteins (Methods) revealed significant changes in 128 (13%) proteins during hypoxia, with ~6 proteins decreasing for each one increasing in concentration (Table S2). These trends are also seen in targeted profiling with multiple reaction monitoring (MRM) – Table S3). Conversely, during re-aeration, 103 (10%) proteins display significant changes relative to the last day of hypoxia with roughly twice as many decreasing proteins (Table S2 and Figure S24). The changes during hypoxia may reflect, in part, decreased translation in non-replicating cells. However, hypoxia induces the transcription factor ClgR (Rv2745c) and the Clp protease it regulates[48], along with other proteases (data available at TBDB.org), suggesting possible increased proteolysis.

Independent of the relative contributions of transcription, export or proteolysis, these data indicate a hypoxia-induced rebalancing of amino acid and protein pools, in which amino acids increase extracellularly and decrease intracellularly. Released amino acids, in turn, may serve as an energy, carbon, or nitrogen source during hypoxia-induced dormancy.

Three amino acids do not decrease during hypoxia. Asparagine is present in the culture medium, and as a consequence it is increased in hypoxia. Similarly, aspartate increases throughout hypoxia and may reflect deamination of asparagine. Glycine also increases during hypoxia, likely from shunting of glyoxylate from malate to glycine[49]. Consistent with this, malate synthase is not activated during hypoxia, but we see activation of shunt through, and accumulation of, 3-phospho-D-glycerate[49]. The majority of genes predicted to be involved in the metabolism or transport of amino acids are down-regulated or show no change during hypoxia. These genes typically return to baseline during re-aeration. Exceptions include certain genes predicted to be involved in the degradation of alanine, serine, glutamate, and valine.

1.9. Concluding Remarks

This report presents the first step 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 50 MTB transcription factors. Although this initial network 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 about 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 transcriptional regulatory network, and to extend this network to encompass other regulation including signaling and non-coding RNAs[50].

The results presented here identify compelling questions for further investigation. As with previous reports, we identify more binding sites than previously expected for even well studied TFs, including binding outside the proximal promoter region. A similar finding was recently reported for EspR in MTB[5] and many similar examples are known for other bacteria[51-62]. The functions, if any, of the majority of the reproducible TF binding sites are not known. Our results suggest regulatory functions for some, although given the large number of sites tested a fraction of those predicted to have function are likely false positives (pFDR=0.15). Experiments are ongoing to validate predicted functional sites and assess 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. In all, our network data suggest that transcriptional regulation in prokaryotes is likely more complex than generally considered. Similarly, our profiling data suggest interpretations for changes in gene expression, metabolites, and proteins that require future targeted experimentation for verification. These data also provide a consistent and comparable data set that can be used in the development of systems level modeling algorithms. 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. Moreover, the profiling data provides a biological context for certain of the regulatory links we observe. In particular, the regulatory network reveals a link between hypoxia adaptation and lipid metabolism that suggests that certain alterations in lipid metabolism are hard-wired as a response to hypoxia. This is exemplified by changes in cholesterol metabolism during hypoxia in the absence of external cholesterol. 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.

2. Methods

2.1. Strains

MTB H37Rv was used for all experiments with the single exception of one experiment performed in *M. smegmatis* and described in Figure S23. The specific MTB strain was acquired from the Colorado State University TB Vaccine Testing and Research Materials Contract. This MTB 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, a library consisting of the majority of MTB regulators under the control of an inducible promoter was generated (one clone per TF) and are available on request.

2.2. ChIP-Seq

To systematically map transcription factor binding sites, we performed ChIP-Seq[63-65] using FLAG-tagged transcription factors episomally expressed under control of a mycobacterial tetracycline-inducible promoter[66-68] (Methods and Figure S1), a method previously validated for ChIP-Seq in other systems[69,70]. The 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.

Cloning. An ATc-inducible episomal vector containing a Gateway Recombination™ (Invitrogen) cassette (kind gift of Eric Rubin) was modified to contain an in-frame N- or C-terminal FLAG epitope tag (plasmid Destination Tet. N-terminal Flag Tag: pDTNF; or plasmid Destination Tet. C-terminal Flag Tag: pDTCF). A transcription factor-of-interest was selected from a Gateway entry clone library (supplied by the PFGRC, contracted by the NIAID) and recombined in to this vector to create N- or C-terminally epitope-tagged expression vectors (plasmid EXpression N-terminal Flag Tag: pEXNF; or plasmid EXpression C-terminal Flag Tag: pEXCF). In the event that the desired transcription factor was not included in the entry clone library, the ORF was sub-cloned from the H37Rv genome using gene-specific oligos containing Gateway recombination sequences at the 5' ends. Fidelity of all clones was confirmed by sequencing (data not shown). Unless otherwise indicated, expression of all transcription factors was induced with 100 ng ATc per mL of culture.

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 50 million reads per sample. All library preparation and sequencing was performed using standard Illumina protocols.

Sequence Analysis. An overview of our analysis pipeline is shown in Figure S2. Sequence reads were mapped to version two of the M. tuberculosis genome using MAQ[71]. The *pileup* command from the SAMTools[72] 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. Since 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.

These regions were then compared against two wild-type experiments in order to remove enriched regions caused by sequencing artifacts. Enriched regions were identified from the wild-type experiments as described above, and those regions identified in each immunoprecipitation experiment that were found to overlap with those called in a wild-type experiment were filtered. A region was considered overlapping if it overlapped a wild-type region by the larger of either 30 nucleotides or 10 percent of the region length. Overlapping regions with height of less than 5-fold relative to the median compared to the corresponding region in either of the wild-type experiments were filtered from further analyses.

We observed that certain regions showed statistical enrichment in nearly all ChIP-Seq experiments regardless of the ChIPed transcription factor. On closer inspection, it was found that most of these regions displayed strong binding by the histone-like protein Lsr2[25]. We expect these regions reflect non-specific binding. To remove these regions from further analysis, we filtered binding peaks for TFs that overlapped with strong Lsr2 binding sites. Specifically, each of the regions passing the background significance filter described above and found to overlap an Lsr2-bound region (by the larger of 10% of the region length or 30nt) were compared in height to the Lsr2 ChIP-seq lane. ChIP-enriched regions with a height relative to the median height less than 2-fold that of the overlapping Lsr2 region were filtered from further analysis.

Finally, we used a modified version of the CSDeconv blind deconvolution algorithm[73] 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[74]. The enriched regions were searched for this motif using FIMO[75], 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.

2.3. Transcriptomics of TF Induction.

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 the transcriptomic section of this manuscript. 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. RNA samples were profiled using both custom Nimblegen microarrays and multiplex RT-PCR, as described in more detail in the supplemental methods. A comparison of the results from both methods reveals a high level of agreement between these two different measures of gene expression (see supplemental figures).

2.4. ChIP-Seq Validation

We focused on validating the physiological nature of the binding that we observed. In particular, we sought to confirm three things:

- That the binding we observed did not reflect non-random attachment of protein to DNA as a result of induced expression or artificial tagging.
- That the binding sites we observed could be bound at physiological concentrations of transcription factor.
- That the binding we observed could be observed under different cellular conditions

2.4.1. Binding does not reflect random protein-DNA attachments due to induction or tagging

We describe in the main text a number of experiments to confirm that the sites we detect represent sequence specific binding. In particular:

- The binding sites we detect are highly reproducible in both enrichment and location (Figure 1 and Figure S4).
- A binding motif can be discerned for nearly all binding sites (Supplementary File: Transcription Factory Summary).

- Genes associated with weak KstR binding are non-randomly associated with expected gene functions, consistent with the known regulatory role of KstR[30,31] (Figure S10).
- Binding sites increase in enrichment with increasing induction levels of TF, in both normoxia and hypoxia, as would be expected for natural binding sites (Figure 1, Figure S6, and Figure S7).
- We have confirmed direct binding to selected sites using EMSA (Figure S29).

To further assess any possible impact of protein overexpression on the state of the cell that could lead to nonspecific binding, we examined binding data for three non-transcription factor genes that were overexpressed and ChIPed using our system. The genes were the conserved hypothetical protein Rv0020c; Rv3164c, incorrectly annotated as a transcription factor in our records but now thought to serve as a methanol dehydrogenase; and Rv0560c, a gene encoding benzoquninone methyltransferase. As expected, in every case no binding sites were identified, supporting the conclusion that the binding sites we detect are not simply a consequence of general protein overexpression or the use of antibiotics in our system.

Finally, during revision of our manuscript, a report by Blasco and colleagues was published describing the mapping of binding sites for EspR[5]. This report utilized antibodies to the WT EspR protein (and thus induced from its native promoter). We used this report as an opportunity to further validate our system by performing ChIP-Seq of EspR in our inducible flag-tagged system at 100ng/ml ATc (our standard protocol) and comparing the results to that from the WT native Ab. The results are shown in Figure S26. As can be seen, there is a remarkably high level of agreement between these two methods. Our method recapitulates the distribution of binding location and heights, the motif, and the specific pattern of binding at individual loci. Our inducible system thus recapitulates the results of a ChIP-Seq performed via a physiological stimulus.

Importantly, the WT native antibody system on EspR also recapitulates all the features of binding that we observe in nearly all TF we have ChIPed. Namely that not all binding is restricted to the region directly upstream of the start codon for genes (Figure S26). Blasco and colleagues have confirmed, using atomic force microscopy, that these binding sites play a functional role in long range interactions[6]. Binding sites can be occupied at physiological levels of TF

2.4.2. Binding sites can be occupied at physiological levels of TF

To assess the degree to which the biding sites we observed could be bound at physiological levels of transcription factor, we first measured the levels of TF mRNA for three different transcription factors (Rv0081, DosR, and KstR) when expressed at several different levels using our inducible promoter system and when expressed from their native promoters by physiological stimuli. (Figure S8) The selection of these three TFs for further analysis here reflects an overarching physiological theme of our study and comes from their central role in the response to hypoxia and the linkage this response with lipid metabolism. We also evaluated the transcriptional response of 52 genes previously documented to be members of the DosR regulon when the DosR TF was expressed using our inducible promoter system at the levels we used for ChIP-Seq (Figure S9).

For DosR, our standard ChIP-Seq experimental condition (100 ng ATc/ml) induces dosR to physiological levels attained by WT MTB in response to hypoxia (Figure S8A). At lower levels of ATc, DosR is induced to sub-physiological levels. Moreover, as shown in Figure S9, induction of DosR at the 100ng/ml ATc level that we used for our ChIP-Seq experiments results in nearly identical induction levels of previously documented DosR regulon genes as compared to their level of induction to hypoxia. Induction at the 100ng/ml level most closely approximates induction by hypoxia as compared to the lower levels of ATc induction that we investigated. Thus, not only is DosR being induced to physiological levels in our ChIP-Seq experimental system, but it is also activating its regulon to physiological levels.

In addition, ectopic induction results in the same physiological DosR regulon response in both hypoxic or normoxic conditions (Figure S9). This supports that conclusion that ectopic induction at the 100ng/ml ATc levels used for our ChIP-Seq studies identifies the same DosR binding sites with the same transcriptional regulatory effect in both hypoxia and normoxia.

For Rv0081, , our standard ChIP-Seq experimental condition (100 ng ATc/ml) induces this TF to ~3x fold greater levels then compared to levels attained by WT MTB in response to hypoxia (Figure S8B). Hypoxic induction of Rv0081 leads to transcript levels that correspond to those achieved between 1 and 10 ng/ml ATc in our system. Importantly, between the 1 and 10ng/ml ATc levels that correspond to physiological hypoxic levels of Rv0081, we can identify over 400 binding sites (Figure S7). Thus ~80% of the binding sites we identify at 100ng/ml ATc are still identifiable at levels of induction that correspond to those attained by the physiological inducer, hypoxia. Although the remaining binding sites at 100ng/ml fall below our detection limit at 10ng/ml, their affinity – as assessed by peak height - does not fall off drastically from those sites that are identified at physiological levels (Figure S7). Together, these data support the conclusion that a significant majority of Rv0081 binding sites are bound at physiological levels.

For KstR, we compared the induction levels in our standard ChIP-Seq system to that of cholesterol. Induction by 1 mg/ml cholesterol results in KstR mRNA levels that are near the 0.1 ng/ml ATc levels (when normalized by SigA). The 100ng/ml ATc levels result in roughly 4x fold over expression (Figure S8C). To assess detection of binding sites when KstR is induced by cholesterol, we inserted a copy of flag-tagged KstR along with its native promoter into the WT MTB chromosome at the attB integration site. We then induced with cholesterol and performed ChIP-Seq. As expected from the transcript levels, we see binding that roughly equivalent to ChIP-Seq from the inducible promoter at the 0.1-1ng/ml ATc levels. And the binding in cholesterol is correlated with the binding in the inducible promoter system (Figure S27).

From these new data we can characterize the identified KstR binding sites into two categories. First, several novel peaks can be identified in ChIP-Seq from either the WT cholesterol system or our inducible system at the 1ng/ml ATc level. Our data thus confirms that these sites can be bound at physiological levels of KstR. Second, a set of peaks are identified that have binding heights equivalent to known peaks but are only identifiable at higher levels of induction. Both these novel peaks and known peaks are missed at lower levels of induction. The similarity in peak height in this second category suggests that the novel peaks in this category can be bound at levels of KstR that are physiological. Their lack of detection at lower levels of induction thus likely reflects a lower sensitivity of the method at these lower levels.

Finally, During the revision of our manuscript a report by Gao et al. (2012)[32] was published that characterized a small number of Rv2034 binding sites. This manuscript brought to our attention an earlier report by the same group[76]. Rv2034 is an EHR regulator that we mapped, and our data reproduce the results of the detailed biochemical analyses of these published reports. In particular:

- Gao et al. (2012) demonstrated using EMSA that Rv2034 binds to its own promoter more strongly than to promoters of other genes they tested. Our ChIP-Seq data confirm that autobinding is the strongest peak for Rv2034.
- We identified the binding of Rv2034 both upstream of the DosR operon as well as near the DosR gene, as reported by Gao et al using EMSA. Gao et al. (2012) also confirm that Rv2034 is a functional regulator of DosR. Thus, as described in detail below, the report by Gao et al. confirms the feedback from the EHR to the initial hypoxic response that we report (we have modified our text to note this confirmation).
- Moreover, the binding of Rv2034 to DosR is one of the weaker binding sites for Rv2034 (1.4% of the maximal binding peak). The data by Gao et al. (2012) thus confirms the biological relevance of this weak binding site.
- In the earlier report[76], Gao and colleagues reported that Rv2034 regulates GroEL2 and binds upstream of this gene using EMSA. Our data confirms this. Importantly, the binding site for Rv2034 resides 746bp upstream of the start codon of GroEL2.
- We identify a motif that consists of the same core motif as reported by Gao et al. (2012).
- In the earlier report[76], Gao and colleagues reported that Rv2034 regulates PhoP and binds to its promoter in EMSA experiments. We identify enrichment in our ChIP-Seq data that suggests binding of Rv2034 upstream of PhoP. Interestingly, the enrichment we observe is at our threshold for calling a binding site. This published result thus further validates the existence of functional weak binding sites.
- Importantly, the reports by Goa and colleagues only discuss a handful of sites for Rv2034. Our manuscript represents a more complete description of the binding potential of this TF and 50 other TFs.

2.4.3. Binding sites can be observed under different cellular conditions

The ultimate goal or our project is to map all transcription factors in the MTB genome. For this it was necessary to select a uniform reference condition that was both practical and not specific to any subset of transcription factors. Moreover, as our data are intended as a community resource, we sought to choose a condition that would reflect experiments performed by other groups. For these reasons we elected to use a standard normoxic culture condition widely used for MTB.

However, to test the degree to which the binding sites we observed could be detected under different physiological conditions, we investigated whether the same binding sites would be observed for 11 different TFs when ChIP-Seq was performed under hypoxic conditions as compared to normoxic conditions. We included several TFs known to be induced by hypoxia, KstR, Lsr2, and EspR. We have compared these data to the ChIP-Seq of the same TFs in normoxia, and the results are shown in Figure S5.

As shown in Figure S5, for all TFs there is substantial concordance between the binding enrichment seen in normoxia and that seen in hypoxia. For the large majority of the regions where we find evidence for binding in normoxia for these TFs, we also see evidence for binding in hypoxia. Moreover, we do not observe any binding sites in hypoxia for which we do not observe evidence for binding in normoxia. In addition, the relative affinities between different binding regions, as judged by read coverage, are broadly conserved between experiments performed during hypoxia and during normoxia. We show this for each TF in Figure S5 as both a scatter plot of enriched regions between replicates as well as through examples of raw coverage for selected regions in each of the conditions.

We note that the enrichment of coverage for binding sites is generally lower in hypoxia as compared to normoxia. This is not surprising, since the majority of the TFs that were selected for hypoxic ChIP-Seq are themselves induced by hypoxia. In our system, this means that the amount of untagged native TF will increase relative to the amount of tagged induced TF. At each binding site, therefore, there will be more competition between tagged and untagged TFs leading to decreased enrichment. Because the same number of sequencing reads is always generated, less enrichment and binding peaks will necessarily mean higher background coverage. This is born out in our results. In every comparison, background coverage is higher in hypoxia than normoxia.

Part of the modest variation we detect in binding between normoxic and hypoxic conditions also likely arises as a consequence of changes in affinity of some TFs or binding sites in the different conditions. As noted in our previous submission, we do not expect that all binding sites will be occupied identically in different conditions. Differences in the expression of the TF itself, post-translational modifications of the TF, differential expression of co-factors, and changes in DNA conformation and accessibility are all expected to play important modulatory roles. Through the use of an inducible promoter, we limit the impact of differences in expression between conditions. In this case, based on our data, the same binding sites appear bound in two different conditions, even though we observed some variation in the apparent binding affinity.

In over 4000 predicted binding sites for 10 TFs we observe only a handful of sites that show marked differences in binding between hypoxia and normoxia. The three binding sites showing the most substantial differences between conditions are notable. These are highlighted in red in the scatter plots in Figure S5. All three are present in both conditions, but show greater binding in hypoxia relative to normoxia. One of these was reported in our previous submission – the autobinding site for Rv0081. The two additional sites are also autobinding sites for two other TFs – one for Rv2034 and the other for Rv3249c. All three of these TFs are differentially expressed in hypoxia. Thus, the three binding sites that show the most significant changes in affinity between hypoxia and normoxia are all autobinding sites in the promoters of TFs that are differentially expressed in hypoxia. And in all three cases, affinity increased in hypoxia. We are actively following up on this intriguing observation.

2.4.4. Potential Issues

Despite the validation presented above, some potential issues arise from the use of our experimental system. Ectopic expression using an inducible promoter system may lead to binding sites that, while sequence specific, may not arise under physiological conditions. These false positives may arise from alterations in DNA structure or interactions with other proteins not normally co-expressed with a given TF. Moreover, the normal physiological levels for most TFs are not known and thus we cannot rule out false positives for these factors as a result of excessive over-expression. Finally, although our data suggests few false negatives, we cannot rule out binding sites that would only be detectable in specific conditions or when additional molecular factors are expressed.

2.5. 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[29]. 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. One experiment was used for cross platform analyses. The second was used for higher sampling frequency lipidomics. The third was done to increase confidence in the metabolomics and proteomic data sets. 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[48], 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 (Figure S28).

2.6. RNA Isolation

The hypoxia and re-aeration time course expression data for Nimblegen is a combination of 3 separate experiments – 2, 6 and 7. The number of replicates used per experiment is shown in Figure S17.

2.7. Transcriptomics using NimbleGen Microarrays

Total RNA was assayed using a Roche NimbleGen custom 12-plex microarray consisting of 105K probes tiled every ~100 bp over both strands of the genome and 30K random probes used as a measure of background (NCBI GEO platform GPL14824). Probes were positioned to capture recently described small and non-coding RNAs[77] as well as uncharacterized intergenic regions and antisense RNA. The microarray was implemented as described previously[78] following the manufacturer's protocols for hybridization, washing, and scanning.

Background subtracted intensities from probes covering each ORF were collapsed and normalized using RMA on all arrays[79,80]. The expression values for each day in hypoxia/re-aeration were obtained by averaging over the RMA values of all available replicates from corresponding experiments. The fold changes are log2 ratios of each day to day 0. Transcription factor overexpressor and hypoxic time course arrays were normalized separately.

2.8. Transcriptomics using Multiplexed RT-PCR

Total RNA from cells grown in the hypoxia time course model was isolated and assayed using a multiplexed RT-PCR protocol. The protocol assays ~2200 genes selected as the first genes of predicted operons. It consists of three parts:

(1) First strand cDNA synthesis and Controlled Multiplex Pre-Amplification of cDNAs

(2) Preparation of Primer and Probe Sets

(3) Individual Real Time PCR (Taqman) quantification of Amplified cDNAs in 384-well format using LightCycler480.

First strand cDNA synthesis: Based upon the provided total RNA concentration (µg/µl), samples were diluted to 10 ng/µl. 50ng of each RNA sample (5µl) was taken into two separate cDNA synthesis reactions (RT+ and RT-) to control for DNA contamination. An additional water control was also added to the RT+ samples. To each sample, 0.5µl Exo-resistant Random Primer (Fermentas S0181), 1µl 10mM dNTPs (Fermentas R0193) and 3.5µl of Nuclease Free Water (Ambion AM9938) were added for a total of 10µl. This mix was incubated for 3 minutes at 70ºC in a thermal cycler, and then placed on ice.

During this incubation, two cocktails were prepared: one containing reverse transcriptase (RT+) and one without as a control (RT-). Each RT+ cocktail contained 4µl 5X Maxima RT Buffer (Fermentas EP0741), 0.5µl Ribolock RNase-Inhibitor (Fermentas EO0382), 0.5µl Maxima RT enzyme (Fermentas EP0741), and 3.0µl Nuclease Free Water for a total of 10µl. The RT- cocktail was exactly the same except water was substituted for the RT enzyme. These cocktails were scaled up for multiple samples and 10µl aliquots were added to each RT+ or RT- sample prepared above. The samples were mixed gently and then spun at 1200 RPM for 2 minutes. They were then incubated at 50 ºC for 1 hour, 95^{o}C for 2 minutes to inactivate, and then held at 4^{o}C. The samples could be taken directly for amplification or stored at 4^{o}C overnight. Maxima Reverse Transcriptase (RT) is a novel reverse transcription enzyme developed by Thermo Scientific through *in vitro* evolution of M-MuLV RT. The enzyme possesses an RNA and DNA-dependent polymerase activity as well as RNase H activity.

Preparation of RT-Primer Mix for Controlled Multiplex Pre-amplification of cDNAs[81,82]: To prepare the Primer Mixes, equal volumes of outflanking 100µM forward (RTF) and reverse (RTR) primers were mixed together. Because 2179 genes were profiled, genes were split into three mixes containing 718 unique genes. 25 additional genes were added to each mix to control for variation across amplification mixes.

Since each gene has a forward and reverse primer, 1486 total oligos were added at equal volumes to each Pre-amplification Mix (PA1, PA2, and PA3). The final concentration of the primers in the amplification reaction is about 52 nM.

Each pre-amplification PCR reaction consisted of 3.0µl 10X Advantage 2 Buffer (Clontech 639202), 0.6µl 10mM dNTPs (Fermentas R0193), 23.8µl Primer Mix (discussed above, Biosearch), and 0.6µl Advantage 2 Polymerase (Clontech 639202) for a total of 28µl. The reactions were scaled up for multiple samples (taking into account RT+ and RT-) and then 28µl aliquots of the mixes were pipetted in 96-well plates. 2µl of RT+ or RT- cDNA was then added for a total of 30µl. Additional controls of water and two aliquots of 10^4 gene copy number of H37Rv genomic DNA were also amplified. The ABI 9800 Fast Thermocycler has a capacity of 30ul reactions. In order to generate enough material to profile ~2,200 genes, two reactions were made for each sample. The samples were spun for 2 minutes at 1,200 RPM and then placed in the thermocycler. Samples were denatured at 95^{o}C for 5 minutes. Fifteen cycles were run at 95 ^{o}C for 30 seconds, 60^{o}C for 20 seconds, and 68^{o}C for 1 minute. Finally, samples were held at 4^{o}C. For each sample, the two 30µl reactions were then combined into 60µl. This procedure was repeated for the remaining two Primer Mixes (PA2 and PA3).

Validation of Outflanking and TaqMan Primer Probe Sets: All outflanking primers and TaqMan probe sets had been validated in multiplex PCR pre-amplification for linearity of amplification using all the genes used in each pre-amplification cocktail. We also had validated all individual TaqMan assays from our collection for sensitivity and linearity before we started using them in gene expression profiling here. Complete database with all available validated TaqMan sets can be found at ftp://smd-ftp.stanford.edu/tbdb/rtpcr/taqman_oligos.fa. Sequences and design of PCR primer/probe sets can be also found at *http://genes.stanford.edu/technology.php and http://www.tbdb.org/rtpcrData.shtml.*

Quantitative Real Time RT-PCR (qRT-PCR): Primer and probe qRT-PCR sets for each gene consist of a forward primer (TMF), a FAM/BHQ-labeled probe (TMP), and a reverse primer (TMR). These were ordered from Biosearch at a 100uM concentration and a “TM Mix” (Taqman Mix) was prepared for each gene. 27µl of the forward primer, 27µl of the reverse primer, and 9µl of the probe were mixed in 1737µl of Nuclease Free Water for a total of 1800µl. This resulted in a dilution of the forward and reverse primer to 1500 nM and the probe to 500 nM. We used 2µl of this mix in a 10µl qRT-PCR reaction, resulting in a final reaction concentration of 300 nM for each primer and 100 nM for the probe. The Roche Lightcycler 480 has a 384-well format. To accommodate for this, 200µl aliquots of the TM Mixes were added to 96 well plates for pipetting and storage. Since each 743-gene PA Mix sample will be split between two 384-well plates, two sets of four 96-well plates were assembled per mix. Each set contained 359 unique genes and the 25 repeated genes.

For each gene reaction, a cocktail including 5.0µl of 2X LightCycler 480 Probes Master Mix (Roche 04902343001), 2.93µl Probes Master PCR-grade water (Roche 04902343001), and 0.07µl of pre-amplified cDNA was prepared. This reaction was scaled up for each of the three PA Mixes to make enough for ~800 genes. 8ul aliquots of each cocktail were added to each well of two 384-well qRT-PCR plates. 2ul of each of the TM Mixes discussed above was then added for a total of 10ul and sealed with optically clear adhesive foil. Since there were two plates per PA Mix and three mixes (PA1, PA2, and PA3), there were six 384-well qPCR plates total per sample.

The plates were spun at 1,400 RPM for 2 minutes and then placed in the Roche Lightcycler 480. The Mono Color Hydrolysis Probe detection format for FAM was used. Activation was performed at 95^{o}C for 5 minutes (Ramp Rate of 4.8°C/sec). Then 40 cycles at 95 ^{o}C for 30 seconds (Ramp Rate of 4.8°C/sec) and 60^{o}C for 30 seconds (Ramp Rate of 2.5°C/sec) were run. Finally, a cool down step of 40^{o}C for 30 seconds (Ramp Rate of 4.8°C/sec) was run for 1 cycle. The data were analyzed using Roche's Second Derivative High Confidence algorithm on the Roche LightCycler Software.

Quality Control: After qRT-PCR, RT+ samples should show normal signals, while RT- should show low to no signal, similar to the water controls. If there is a strong signal in RT-, it suggests that there is either contaminating genomic DNA and the RNA samples must be re-DNased or that the RNA is degraded and non-specific amplification is occurring. There should be a difference of about 10 Cts between RT+ and RT- to confirm quality. Two different water controls were run throughout the process: one starting at the cDNA step and one starting at the amplification step. If noise was seen in either water control, the source of contamination or any issues could more easily be found. Additionally, two samples of 10^4 gene copy number H37Rv genomic DNA were amplified in each amplification mix and run on each plate. The 25 repeated genes could be compared across all plates to ensure that all amplification and qRT-PCR reactions were performing similarly. No issues were found in this experiment.

Normalization and Data Analysis: The ability to measure MTB gene expression on a genome-wide scale requires appropriate normalization strategies to control for experimental error introduced during the multistage process required to extract and process the RNA. There are many strategies that can be chosen; these include normalization to sample size, total RNA and the popular practice of measuring an internal reference or housekeeping gene. Improvements in qRT-PCR technology provide opportunity for multiple control genes to be used at once for normalization. However, it is a big challenge to find appropriate "housekeeping" MTB genes to be used as internal reference controls for normalization across a board of various MTB specimens in different metabolic states. Therefore, we adapted a quantile normalization developed for DNA microarray expression analysis, where data-driven methods have become the standard for most experimental designs. Quantile normalization is based on assumption that on average, the distribution of gene transcript levels within the cell remains nearly constant across samples. A quantile is a measure that allows assessing the degree of spread in a data set. Quantile normalization adjusts the overall expression levels so that the distribution for all samples is equal.

In our qRT-PCR experiments, the number of MTB genes assayed in each sample is about 2,200, and a single specimen is profiled on six 384-well PCR plates. To correct for plate effects in raw qPCR data, we included a set of 25 reference genes on every plate and applied a quantile normalization approach where we make the assumption that the distribution of gene expression measures is the same across all plates for the same experimental condition. This assumption could be reasonable because the gene allocation is random and is not related to their expected expression levels or functional properties. After qRT-PCR runs raw data from all 6 plates were assembled into Excel spreadsheets, and normalization was applied to all of the genes assayed for each sample. First, Median Ct value was calculated for all genes for each sample and then Median of the median individual Ct values for all samples calculated. Then, quantile normalization was applied between samples, assuring that each sample has the same distribution of expression values as all of the other samples to be compared. A similar approach has been previously described for microarray normalization[79].

In summary, our quantile normalization method includes the following steps:

1. Deposition of raw Ct data for single RNA sample from all 6 plates into a column of an Excel spreadsheet, where genes are represented by lines and samples by columns; Ct values for 25 reference genes on each plate are assessed for all 6 plates and reduced to their median value, so each gene is represented once.
2. Median Ct is then calculated for each column/specimen.
3. A median of these median Cts from the step 2 is calculated and further used for normalization.
4. dCt for each specimen is the difference between Median of the Median Cts from step 3 and Ct median of each sample from step 2.
5. This individual sample-specific dCt is used to normalize each gene expression value of that specimen.
6. Finally, median Ct value is calculated for a group of specimens representing a particular time point in the experiment and then these median Cts may be compared to each other, or to Cts of specimens representing Time 0 or Day 7 of Hypoxia.
7. Log2 ratios could be calculated using dCts method for each gene and Student Ttest could be applied to identify statistically significant differences with or without Bonferroni corrections.

2.9. 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. In brief, 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 also depicted in Figure S14, 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 compares the model fit to the null hypothesis that the regression variables do not have predictive power[83]. To identify whether this test statistics is larger than would be expected from random chance, it is compared to its distribution under null (the F-distrubution)[83,84], from which p-values are generated for optimal models selected in the training. The Storey method is used to estimate analytical pFDR from these p-values[85-87].
    b. Model selection was guided using AIC[88] and Lilliefors[89] 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 pass this test, the structure with the minimum AIC is selected as optimal. If no model passes 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 we repeat step 5.
6. The model at step 4 is selected as the final optimal model if adding an additional regulator in step 5 did not improve prediction accuracy (Figure S14).

Model Structures

Since the exact relationship between target genes and TFs may vary[17], 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 forming 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 show significantly better performance for most genes while they would present more parameters with more complex models and could overfit. 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.

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.

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 to avoid over-fitting. 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.

By parameterizing the optimal model, expression is predicted as $\underline{\hat{\boldsymbol{y}}} = \hat{\boldsymbol{f}}(\boldsymbol{X})$ and the error is calculated between the predicted and actual expression as $\boldsymbol{L}\left(\underline{\boldsymbol{y}}, \hat{\boldsymbol{f}}(\boldsymbol{X})\right) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{N}}\left(\underline{\boldsymbol{y}}_{\boldsymbol{j}} - \hat{\boldsymbol{f}}\left(\boldsymbol{X}_{\boldsymbol{row=j}}\right)\right)^{\boldsymbol{2}}$, which is sum of squares error or prediction (SSE). F-test p-values are used as a measure of prediction accuracy[83,84,90,91], calculated from SSE. As different TF inductions showed a wide range of variance in gene expressions, the SSE in F-statistics was corrected to aacount for this variance, assuring a ratio of two chi-squared distributions.

This test statistic (as opposed to SSE directly) accounts for different degrees of freedom for different optimal model structures.

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 5-fold cross-validation as below. The data was preprocessed using robust multichip analysis (RMA)[79,80,92].

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 80% 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 20% of the TF induction data set. Prediction accuracy was evaluated as F-test p-values estimated from the sum of squared errors (SSE) between prediction and actual expression values for optimal models selected in the training.
3. Steps 1,2 were repeated 5 times – i.e. 5-fold cross-validation – and the overall accuracy of the model determined by averaging the results of each step 2 across all 5 cross-validated models, as an empirical estimation of the expected p-value (EPV)[93,94].

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. This comparison generates an empirical estimation of FDR. Similar randomization or permutation approaches have been used in other regulatory network modeling efforts[95,96] as well as other methods[97-99] for empirical estimation of the distribution of a test statistic under the null hypothesis. Here, the permutation is such that each TF of the network can be randomly considered as a regressor. 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 5-fold cross-validation was performed – thus the best fitting model structure was selected independently for each random set. The prediction accuracy of the 20 random TF sets form a reference (null) distribution to be compared to the accuracy of binding TFs, resulting in a one-tailed p-value. Also, to correct for large variances, the rank of the true model compared to accuracy of random TFs is also calculated, i.e. indicating where it stands compared to 20 random sets, sorted in ascending order. We also calculate an explicit pFDR estimate from these p-values using Storey's method[85-87].

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, as results were more consistent with companion RT-PCR data.

To generate a single model for each gene that passed the above validation, the best model structure was selected and trained on the entire TF induction expression data set (Figure S14). This step utilized the validated genes from cross-validation, to predict the expression of genes during hypoxia and re-aeration. This step tests the ability of the models, generated from data derived from a baseline aerobic condition, to generalize and predict the expression of genes during a different, hypoxic, condition. Each time point during hypoxia and re-aeration was predicted separately and independent of previous time points. We are thus currently predicting steady-state expression rather than timeseries evolution.

Only genes whose expression changed by more than 2-fold, prior to normalization, were considered.

After predicting the expression of each time point, we calculate the SSE between the model predictions at all time-points and the actual normalized expression data. Similarly, we use this SSE to calculate an F-test p-value as above. We also compare the predictions of the models to random TF's as described above to check late in empirical FDR.

Summary of Modeling Results

TF Induction Cross-validation Results: Out of 3072 genes which have binding with impulse height more than 1%, significant models can be generated for 2755 (89%) with p-value less than 0.1 (pFDR<0.01). Out of these, 953 (36%) had significant predictions with binding TF models with F-test p-value<0.25 and in total, 873 genes/models were validated and predicted better than average random models (pFDR<0.15) (Table S4).

Hypoxia Prediction Results: Out of the genes validated in the cross-valdiation, 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 808 genes tested for generalization to the hypoxic condition. Out of these genes, 651 (80% of changing genes) have significant predictions (p<0.25), of which 533 were also predicted better than average random (FDR<0.19). These results are summarized in Table S4.

2.10. DREM Analysis

Changes in oxygen availability result in expression changes to nearly one-third of all MTB genes. 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[100]. We then assessed the consistency between each path, the expression of each TF that binds genes in the path, and the predicted regulatory role of the TF based on the overexpression transcriptomics.

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.

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 induction, matches the correlation in expression between Rv0081 and these genes during hypoxia and re-aeration. Rv0081 is initially induced during hypoxia, declines in expression throughout hypoxia, and is expressed at a low level during re-aeration. Rv0081 binds 25% to 40% of the genes in each DREM path. For path 1, which displays genes that are highly activated during hypoxia and trend down during re-aeration, 70% of Rv0081 binding is predicted to result in activation. Conversely, for path 3-2-2, which is highly repressed during hypoxia, 85% of Rv0081 binding is predicted to be repressive. Paths with intermediate levels of activation or repression display mixes of activating and repressing Rv0081 binding (Figure 4B). In short, Rv0081 regulatory roles are predictive of the expression of genes in each of the DREM paths. A broad regulatory role of Rv0081 is thus supported by three independent sources of evidence: overexpression of Rv0081 alters the expression of a large number of genes, ChIP-Seq reveals a large number of binding sites (which are also detected when ChIP is performed on Rv0081 during hypoxia – Figure S16), 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.

2.11. Metabolomics

M. tuberculosis cell pellets or their conditioned culture medium were metabolically profiled by Metabolon using three independent platforms: ultrahigh performance liquid chromatography/tandem mass spectrometry (UHPLC/MS/MS) optimized for basic species, UHPLC/MS/MS optimized for acidic species, and gas chromatography/mass spectrometry (GC/MS)[101]. This integrated platform enables the high-throughput collection and relative quantitative analysis of analytical data and identifies a large number and broad spectrum of molecules with a high degree of confidence[102]. Briefly, cells were spun down into pellets, brought up in 10 ml 2:1 chloroform:methanol, incubated at 37°C on a rotating platform for 2h, then dried under a nitrogen stream and frozen at -80°C. Culture medium was filtered through a 0.2 µm Steriflip filter unit twice. Aliquots of each of the cell pellet and medium samples were plated on 7H10 plates for four weeks to confirm sterility. Upon confirmation, samples were sent to Metabolon for further processing.

Extracts of cell pellet samples were removed from the-80°C freezer and placed on ice. A 600 µL volume of 82% MeOH solvent (HPLC-grade methanol in water containing four standards to report on extraction efficiency) was added to each sample then sonicated for 30 seconds to loosen and dissolve the material visible on the walls of the tubes. The samples were briefly vortexed, gently centrifuged at about 800 RPM on a Beckman GS-6R centrifuge at 4°C, and 550 µL of the supernatants from the reconstituted extracts were transferred to a Nunc 96-well deep-well plate. The deep-well plate was then placed on the Hamilton LabStar Liquid Handling Robot and 110 µL aliquots were transferred to four sets of 250 µL autosampler inserts (see below), dried under nitrogen and vacuum-desiccated, then stored at -80°C until needed.

Samples of the culture medium were removed from the -80°C freezer and placed on ice to thaw. Osmometry was performed on the samples (Fiske Micro Osmometer Model 210) and a dilution to 1/5 concentration was determined to be needed. A 200 µL aliquot from each sample was transferred to Eppendorf tubes and diluted to 1/5 concentration by the addition of 800 µL of water. A 100 µL aliquot of each diluted sample was transferred to a deep-well plate and 450 µl of 100% HPLC-grade methanol containing four recovery standards was added to precipitate protein from the culture medium. The samples were shaken on a Glen Mills GenoGrinder 2000 at 675 strokes per minute for two minutes. The block was centrifuged at 2000 RPM for 7 minutes on a Beckman centrifuge at 4°C. As with the cell extracts, the resulting supernatant was split into equal aliquots of 110 µl for analysis on the three platforms, with a fourth aliquot reserved for backup. Aliquots, dried under nitrogen and vacuum-desiccated, were stored at -80°C until needed.

Dried aliquots were subsequently reconstituted in either 0.1% formic acid in water (100 µl, acidic conditions) or in 6.5 mM ammonium bicarbonate in water, pH 8 (100 µl, basic conditions) for the two UHPLC/MS/MS analyses or derivatized to a final volume of 50 µl for GC/MS analysis using equal parts bistrimethyl-silyl-trifluoroacetamide and solvent mixture acetonitrile:dichloromethane:cyclohexane (5:4:1) with 5% triethylamine at 60°C for one hour. In addition, three types of controls were analyzed in concert with the experimental samples: samples generated from pooled experimental samples (for culture medium) or a pool of human plasma that has been extensively characterized by Metabolon (in place of pooled cell extracts) served as technical replicates throughout the data set, extracted water samples served as process blanks, and a cocktail of standards spiked into every analyzed sample allowed instrument performance monitoring. Experimental samples and controls were randomized across the platform run.

For UHPLC/MS/MS analysis, aliquots were separated using a Waters Acquity UPLC (Waters, Millford, MA) and analyzed using an LTQ mass spectrometer (Thermo Fisher Scientific, Inc., Waltham, MA), which consisted of an electrospray ionization source and linear ion-trap mass analyzer. The MS instrument scanned 99-1000 m/z and alternated between MS and MS/MS scans using dynamic exclusion with approximately 6 scans per second. Derivatized samples for GC/MS were separated on a 5% phenyldimethyl silicone column with helium as the carrier gas and a temperature ramp from 60°C to

340°C and then analyzed on a Thermo-Finnigan Trace DSQ MS (Thermo Fisher Scientific, Inc.) operated at unit mass resolving power with electron impact ionization and a 50-750 atomic mass unit scan range. Metabolites were identified by automated comparison of the ion features in the experimental samples to a reference library of chemical standard entries that included retention time, molecular weight (m/z), preferred adducts, and in-source fragments as well as associated MS spectra and were curated by visual inspection for quality control using software developed at Metabolon[103].

2.12. Lipidomics

To measure lipids we utilized an untargeted lipidomics method using liquid chromatography high-mass accuracy mass spectrometry. Briefly, cell pellets were subjected to a chloroform:methanol (2:1) extraction, and lipid extracts were dried down and then subjected to a water wash by re-suspending total lipid in 6ml of chloroform:methanol (2:1) and 1ml of water (Fisher, HPLC grade). After centrifugation at 2000rpm for 15min, the water layer was removed and discarded to rid of contaminating glycerol extracted from the culture medium. Lipid extracts were weighed and 10μg of total lipid from each sample was injected onto a Varian Monochrom 3μm diol column (2x150mm) and separated using an Agilent 1200 series HPLC with a starting mobile phase of hexanes:isopropanol (70:30) and an eluting mobile phase consisting of isopropanol:methanol (70:30). High-mass accuracy mass spectra were analyzed by an Agilent 6520 Accurate-Mass Q-ToF detector. We acquired data in both positive- and negative-ion modes, but focused on analyzing the positive-ion mode data as it included more lipid classes. The negative ion-mode data was used to corroborate the results. The raw data were analyzed by Agilent Mass Hunter Qualitative Analysis software (version B.03.01), and the data files exported as mzdata files. The raw spectra in these files were then subjected to ion extraction, integration and alignment by an R-based XCMS ion finding algorithm with similar parameters to those described in Layre et al. 2011 in order to deduce the absolute peak area of each ion (including isotopes and different adducts), and then the extracted m/z values were annotated with names of lipids using in-house scripts to match raw m/z values within 10ppm of the theoretical values in MycoMass database (Layre et al. 2011). m/z values and corresponding name assignments were double-checked using the MtbLipidDB (Sartain et al.). The assignments of lipids were confirmed by deducing the structures of representative species from each lipid class by MS/MS in both the positive and negative mode. The annotated species were sorted by lipid class and graphed in Graphpad Prism software. The data shown in the main text are the positive-ion mode data from Experiment 6, and were representative of results seen in Experiment 2 and 5. .

2.13. Proteomics

Sample Preparation for Mass Spectrometry Analysis

Bacterial cultures, received in 6 M guanidinium chloride from the SBRI were pelleted at 16,000 x *g* for 10 minutes at room temperature and the supernatant removed. A solution of chloroform:methanol (3:1 v/v) was added to each sample to remove lipids and centrifuged at 16,000 x g for 20 minutes. The supernatant was removed and the pellets were solubilized in 150 μL of 8M urea at 80°C for 10 minutes. The concentration of urea was diluted to 2M with HPLC grade water. Sequencing grade trypsin (Promega, WI, USA) was added in a 25:1 protein:enzyme ratio in two aliquots. The second aliquot of trypsin was added after 3.5 hours, and digestion allowed to proceed for an additional 16 hours. The reducing agent tris(2-carboxyethyl)phosphine (TCEP) was added to each sample to obtain a final concentration of 10 mM, and incubated for 30 minutes at room temperature. Samples were acidified with concentrated hydrochloric acid to obtain a final concentration of 0.5N. Tryptic digests were desalted using a 3M Empore 96-well plate (Fisher Scientific) containing C18 sorbent as the extraction medium. After a wash step containing 5% ACN in water (v/v), peptides were eluted using a solution of 48% ACN in water (v/v). The final eluant was evaporated to dryness in a 96-well plate prior to LC-MS analysis. Dried extracts were reconstituted in 10 µL of resolubilization solution (95:5 water:ACN, 0.2% formic acid, v/v/v) containing five internal standards at 100 ng/mL.

2.13.1. MRM Assay Development

A total of 593 synthetic peptides were generated for targeted proteomics (JPT peptide technologies GmbH, Berlin, Germany). Each peptide standard was solubilized in 50:50 DMSO:water, then further diluted using LC-MS mobile phase (95:5 water:ACN, 0.2% formic acid, v/v/v). Individual synthetic peptide standards were pooled together to produce a mixed standard at a final concentration of 200 pmol/mL each.

Targeted proteomics using multiple reaction monitoring (MRM) was performed using an AB SCIEX 5500 Q-trap hybrid triple quadrupole-linear ion trap mass spectrometer (AB SCIEX, Concord, Ontario, Canada). Declustering potential (DP) and collision energy (CE) were optimized using in-house developed software. Theoretical MRM ion transitions were generated for each peptide and exported to Analyst software (AB SCIEX, version 1.5.1). Potentials and CE for maximum ion transmission and sensitivity were optimized by performing MRM with enhanced product ion (EPI) scan in information dependent acquisition (IDA) mode. Peptide sequences were confirmed using the Mascot search engine (Matrix Science, London, UK). Acquired data from MRM-EPI runs were exported to MRM program manager, and the two most sensitive ion transitions were selected for the final MRM acquisition method.

2.13.2. Mass Spectrometry Using Targeted MRM

Targeted LC-MS peptide analysis was performed using a Waters Nanoacquity HPLC system (Milford, MA, USA) coupled to an AB SCIEX 5500 Q-trap mass spectrometer equipped with a

Turbo V source. A total of 1044 peptide MRM transitions were monitored using scheduled MRM, including five internal standards. A five microliter aliquot of each protein digest was injected onto the LC-MS system. Chromatographic separation was achieved using a Biobasic C18 HPLC column (Thermo Fisher Scientific), 320 µm i.d. x 15 cm, with a solvent system consisting of 0.2% formic acid in water (v/v, solvent A) and 0.2% formic acid in acetonitrile (v/v, solvent B). Peptide digests were eluted from the HPLC column using gradient elution from 7.5 to 30% B over 30 minutes, with a flow rate set to 5 µL/min.

Data Normalization for Targeted-based Proteomics

Transition peak area data was normalized on a per time point basis. First, a list was created of transitions observed with an intensity value above the limit of quantification in all the samples, on a per time point basis. For each time point, the median intensity over all the transitions was then obtained. Each sample was then normalized by using the median intensity obtained in the previous step (i.e. all the intensities of that sample are divided by the time point median intensity, then log-transformed).

2.13.3. Statistical Analysis – Transition-Level Analysis

A one-factor ANOVA model was used for the analysis and is defined as follows:

$$I_{ij} = M + T_i + \varepsilon_{ij}$$

where *I* is the peptide intensity, *M* is the overall average intensity, *T* is the time, and *ε* is random error. The vessel of origin factor was excluded from this model because the per time point normalization procedure affected the longitudinal property of the data. FDR (false detection rate) and q-value were calculated, based on the p-values obtained from the ANOVA model, using Storey's method to make multiple testing adjustments (implemented in MATLAB). 'Post hoc' contrast analyses were conducted using Tukey's hsd method to calculate p-values associated with each pair wise comparison.

2.13.4. Statistical Analysis – Protein-Level Analysis

The following ANOVA model was used to take into account multiple transitions mapping to the same protein:

$$I_{ikl} = M + T_i + P_k + \varepsilon_{ikl}$$

where *I* is the measured protein intensity, *M* is the overall average intensity, *T* is the time, P is the 'transition' factor, and *ε* is random error. FDR as well as 'Post hoc' contrast analysis were conducted as described above. FDR and q-value were calculated, based on the p-values obtained from the ANOVA model, using Storey's method to make multiple testing adjustments (implemented in MATLAB). 'Post hoc' contrast analyses were conducted using Tukey's hsd method to calculate p-values associated with each pair wise comparison.

2.13.5. Global LC-MS/MS Ion Profiling and Peptide Sequencing

LC-MS/MS analyses for peptide sequencing were performed using a LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA) equipped with a nanoelectrospray interface. Tryptic digests of SG9 protein extracts were injected once using a Waters Nanoacquity HPLC system. A five microliter aliquot of each protein digest was loaded onto a capillary C18 trapping column (2 mm x 180 µm i.d.) for 5 min at 6 µL/min using a loading solvent (99:1 water:acetonitrile, 0.2% formic acid, v/v/v). Chromatographic separation was achieved using a Waters BEH300 C18 column, 150 µm i.d. x 100 mm, using the same solvent system as described above in the targeted MRM section. The instrument was operated in positive ion mode and scanned from m/z 400 to 1600, resolution set to 60,000, with a target value of 1.0 x 10^6. For tandem MS/MS sequencing, the LTQ ion trap was scanned from m/z 50 to 2000, with a target value of 1.0 x 10^4. One full survey scan in the Orbitrap was followed by data dependent MS/MS acquisition on the six most intense precursor ions. Peptide digests were eluted from the HPLC column using gradient elution from 7.5 to 30% B over 120 minutes, with a flow rate set to 0.6 µL/min.

Data Normalization for Discovery-based Proteomics

Peptides observed in less than three samples with intensity count above 25,000 were excluded from data normalization and statistical analysis. All raw intensity values were log (base e) transformed with negative values replaced by 0. The sum of the intensities for each sample was then calculated. In this study, 13 samples fell into the intensity-sum range of [60000, 72000] and were used to create an average sample (i.e. Reference sample), against which the actual samples were then normalized. The normalization factors were chosen in such a way that the median of the log ratios between the actual and the Reference samples over all the peptides was adjusted to 0.

2.13.6. Statistical Analysis – Peptide-Level Analysis

A two-factor ANOVA model was used for the data analysis and is defined as follows:

$$I_{ijk} = M + T_i + V_j + \varepsilon_{ijk}$$

where *I* is the peptide intensity, *M* is the overall average intensity, *T* is the time, *V* is the vessel of origin, and *ε* is random error. FDR (false detection rate) and q-value are calculated, based on the *p*-values obtained from the ANOVA model, using Storey's method to make multiple testing adjustments (implemented in MATLAB). 'Post hoc' contrast analyses were conducted using Tukey's hsd method to calculate p-values associated with each pair wise comparison.

2.13.7. Statistical Analysis – Protein Level Analysis

The following ANOVA model was used to take into account multiple peptides mapping to the same protein:

$$I_{ijkl} = M + T_i + V_j + P_k + \varepsilon_{ijkl}$$

where *I* is the measured protein intensity, *M* is the overall average intensity, *T* is the time, *V* is the vessel of origin, *P* is the individual 'peptide' factor, and *ε* is random error. FDR (false detection rate) and q-value are calculated, based on the *p*-values obtained from the ANOVA model, using Storey's method to make multiple testing adjustments (implemented in MATLAB). 'Post hoc' contrast analyses were conducted using Tukey's hsd method to calculate p-values associated with each pair wise comparison.

3. Supplementary Figures

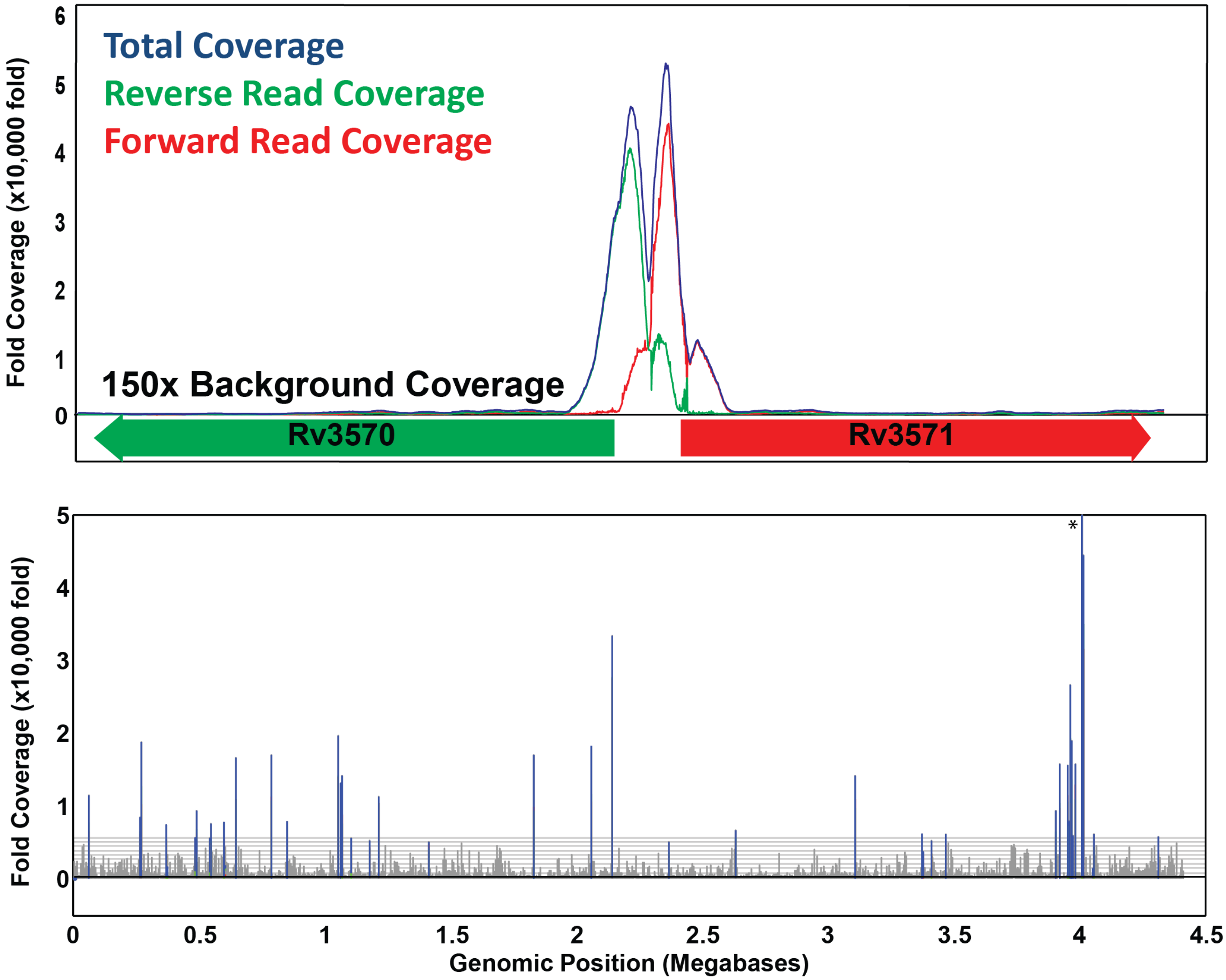


Figure S1: Example MTB ChIP-Seq Peak and Genome Wide Binding – The top panel displays the fold read coverage for a single binding region with two known binding sites for KstR (Rv3574). The ChIP-Seq coverage visually resolves both binding sites and predicts that the site closest to Rv3571 is weaker affinity. The total coverage is shown in blue and the forward and reverse coverage is shown in red and green respectively. Typical genome-wide background coverage is 150x, and we have observed peaks with coverage as high at 1.2 million fold. The binding event also displays the expected shift in position between the forward and reverse reads (c.f. Valouev 2008). The bottom panel displays the genome wide fold coverage for the same ChIP experiment from KstR. Peaks above a coverage threshold are shown in blue. The peak shown in the top panel is marked with a star in the bottom panel. The horizontal grey lines indicate increments of the standard deviation of background coverage.

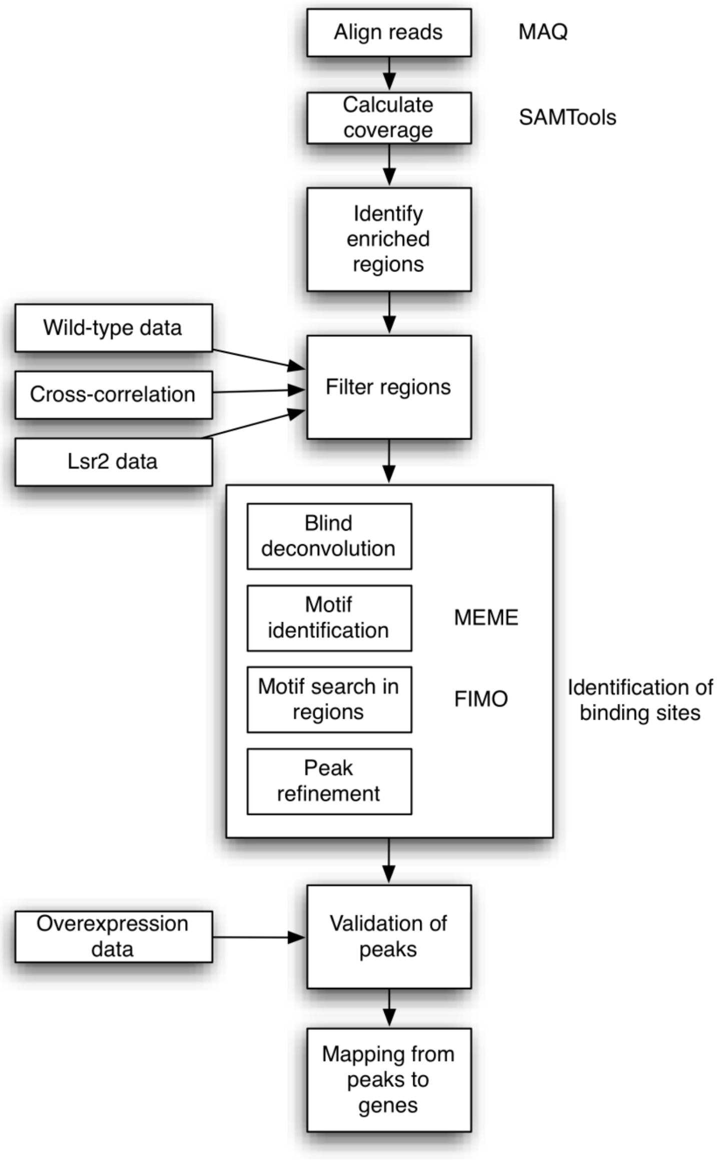


Figure S2: Overview of Analysis Pipeline. Reads are mapped to the H37Rv genome, scored against a lognormal distribution, and filtered to remove computational and experimental artifacts. Blind deconvolution and motif finding are used to identify binding events at single nucleotide resolution. See Methods and Supplementary Material for more details.

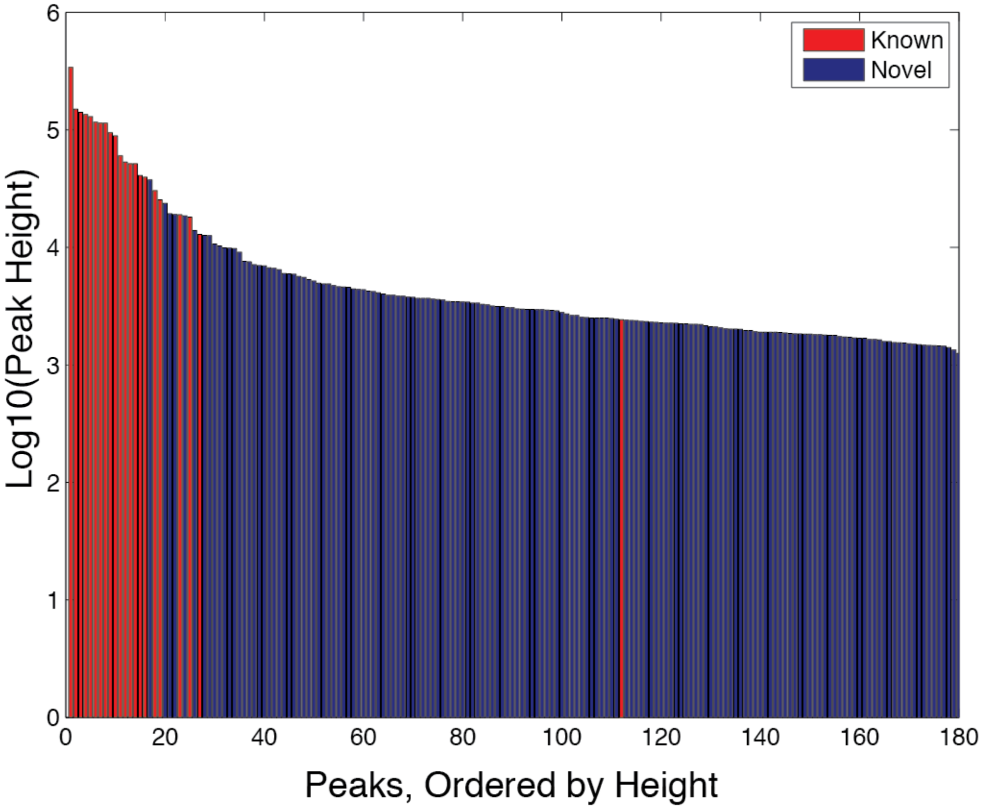


Figure S3: Distribution of Peak Heights and Identification of all known Binding Sites for DosR . We identify all known binding sites for DosR. Binding site heights are plotted as bars and are ordered by peak height. Red bars indicated previously identified binding sites. Blue indicate newly identified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

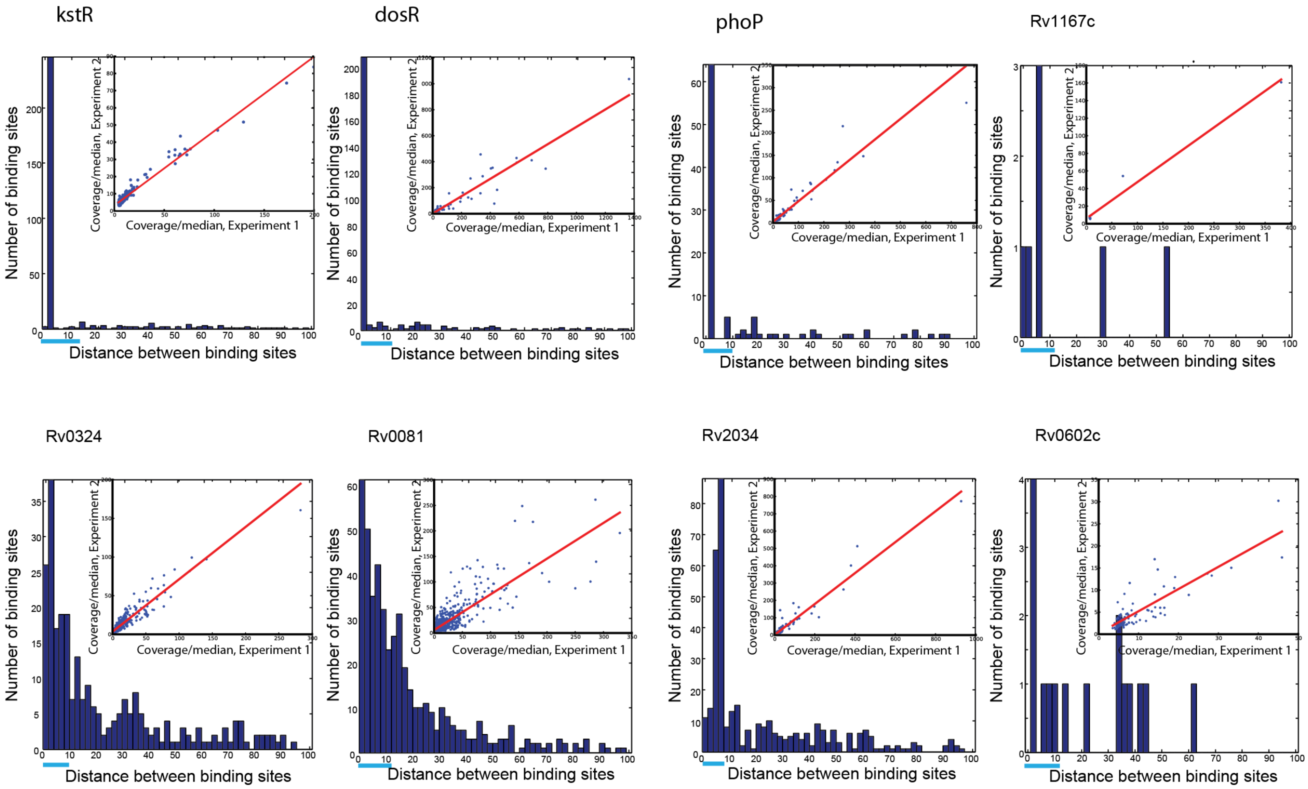
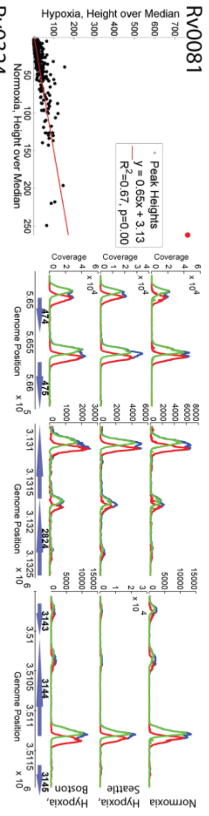
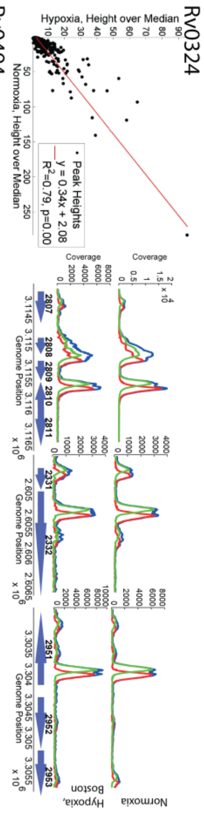
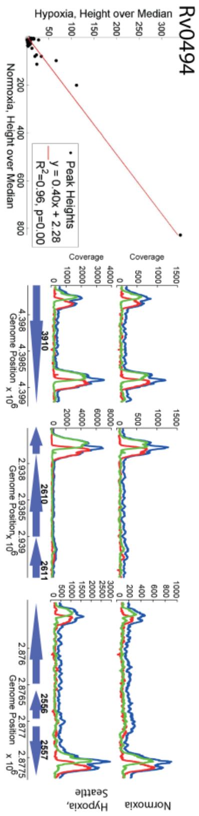
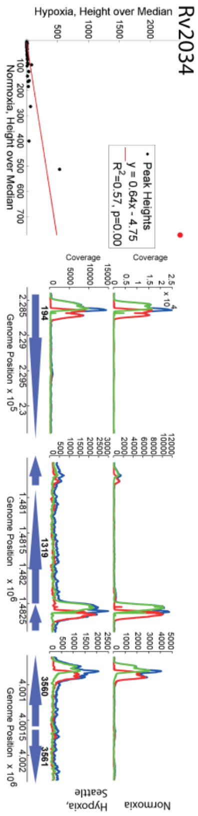


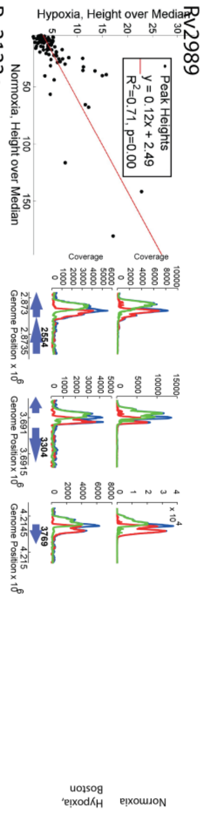
Figure S4: ChIP Binding Shows High Reproducibility in Peak Height and Location. The bar plot shows the distance between corresponding binding sites in two replicates for the same transcription factor. The blue line under the X-axis indicates the width of the predicted motif. The scatter plot shows the correlation of coverage in replicates. For more details, see Figure 1B in the main text.

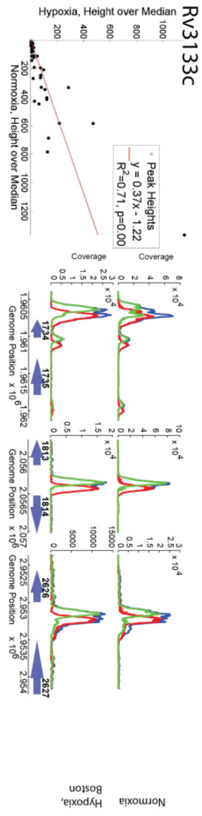

Rv0081
Peak Heights
y = 0.65x + 3.13
R²=0.67, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
474
475
2824
3143
3144
3145
Normoxia
Hypoxia, Seattle
Hypoxia, Boston

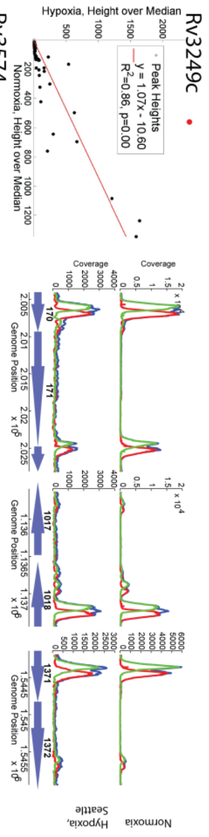

Rv0324
Peak Heights
y = 0.34x + 2.08
R²=0.79, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
2807
2808 2809 2810
2811
2331
2332
2951
2952
2953
Normoxia
Hypoxia, Boston

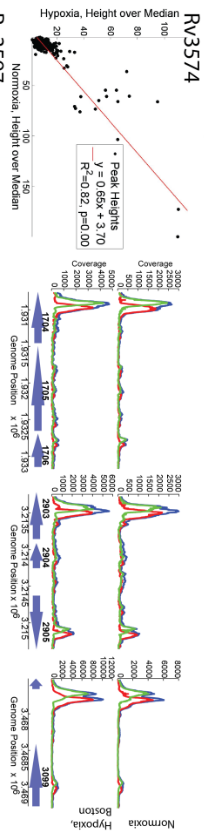

Rv0494
Peak Heights
y = 0.40x + 2.28
R²=0.96, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
3910
2610
2611
2556
2557
Normoxia
Hypoxia, Seattle

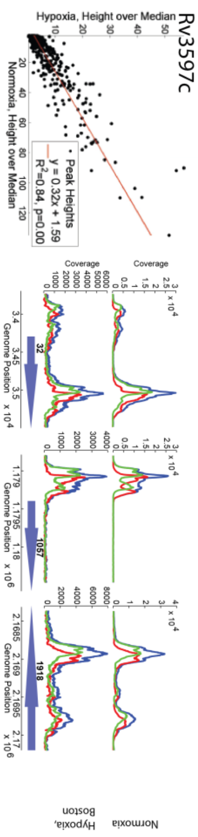

Rv2034
Peak Heights
y = 0.64x - 4.75
R²=0.57, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
194
1319
3560
3561
Normoxia
Hypoxia, Seattle

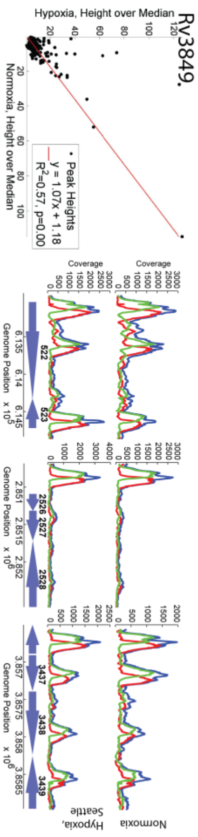

Rv2989
Peak Heights
y = 0.12x + 2.49
R²=0.71, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
2554
3304
3769
Normoxia
Hypoxia, Boston


Rv3133c
Peak Heights
y = 0.37x - 1.22
R²=0.71, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
1734
1735
1813
1814
2626
2627
Normoxia
Hypoxia, Boston


Rv3249c
Peak Heights
y = 1.07x - 10.60
R²=0.86, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
170
171
1017
1018
1371
1372
Normoxia
Hypoxia, Seattle


Rv3574
Peak Heights
y = 0.65x + 3.70
R²=0.82, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
1704
1705
1706
2903
2904
2905
3099
Normoxia
Hypoxia, Boston


Rv3597c
Peak Heights
y = 0.32x + 1.59
R²=0.84, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
32
1067
1918
Normoxia
Hypoxia, Boston


Rv3849
Peak Heights
y = 1.07x + 1.18
R²=0.57, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
522
523
2526 2527
2528
3437
3438
3439
Normoxia
Hypoxia, Seattle

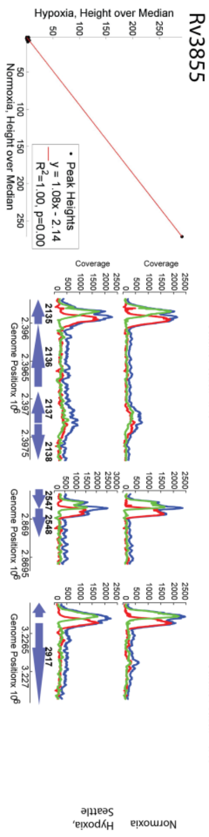

Rv3855
Peak Heights
y = 1.08x - 2.14
R²=1.00, p=0.00
Hypoxia, Height over Median
Normoxia, Height over Median
Coverage
Genome Position
2135
2136
2137
2138
2547 2548
2917
Normoxia
Hypoxia, Seattle

Figure S5: Comparison of Chip-Seq enrichment between normoxia and hypoxia. For each TF, a scatter plot is shown the displays the correlation between read coverage for all regions identified in either hypoxia or normoxia. Coverage in normoxia is shown on the X-axis and coverage in hypoxia is shown on the Y-axis. To the right of each scatter plot are images of raw coverage for example binding sites. For each example, the top plot shows coverage in normoxia and the bottom plot shows coverage in hypoxia for the same region. In the case of Rv0081, example peaks for two different replicates of hypoxia performed at different times and in different labs are provided. As shown, the pattern of enrichment is nearly identical in the different experiments and conditions. The three binding sites that show the most substantial difference in affinity between hypoxia and normoxia are colored in red in the scatter plots, and described in more detail in the text. All experiments were performed using the inducible promoter system. The experiments in normoxia were performed as described in the main methods. For the experiments in hypoxia, cells were grown for 6 hours in aerobic conditions with ATc ,followed by 24 hours in 0.2% O_2 also with ATc.

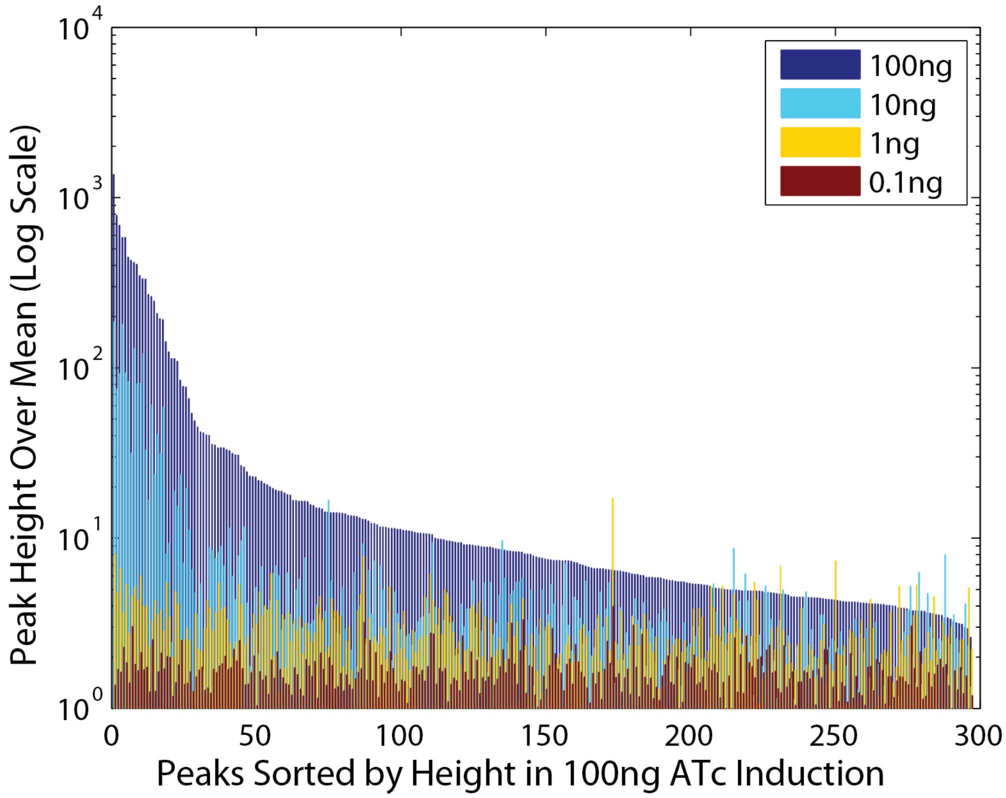


Figure S6: DosR ChIP-Seq at different expression levels using the inducible promoter system. DosR was induced at 4 different levels of ATc and ChIP-Seq performed. Shown are plots of peaks identified at each of the induction levels. Corresponding peaks are plotted at the same position on the horizontal axis.

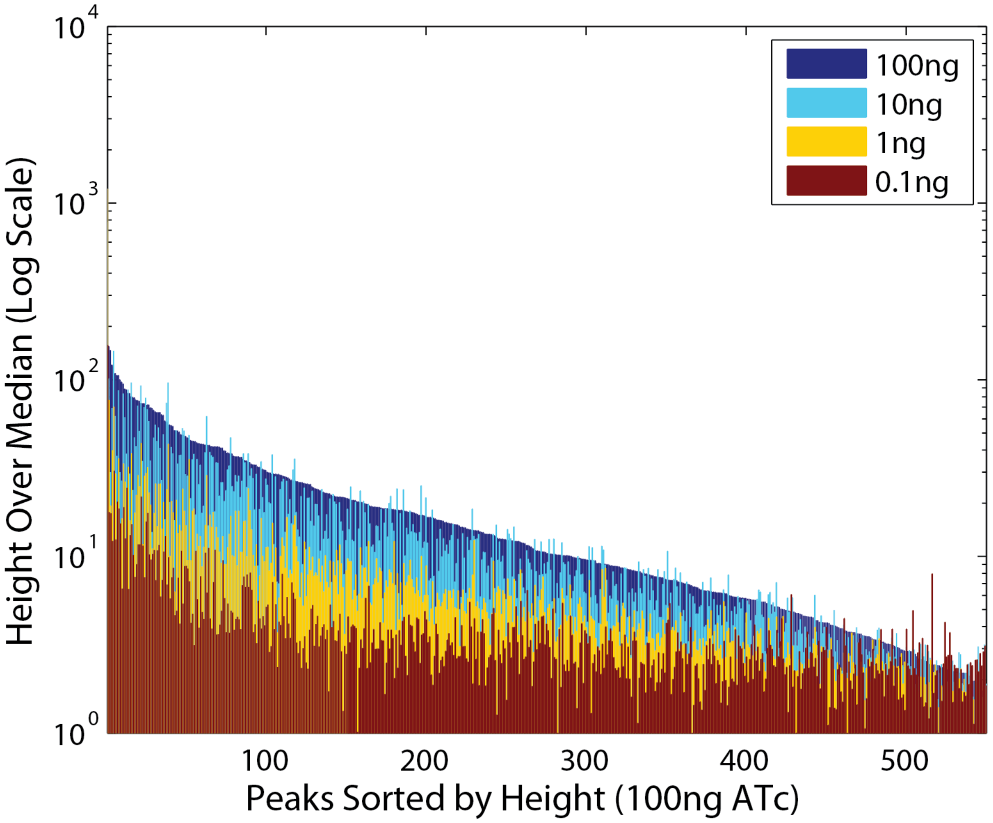


Figure S7: Rv0081 ChIP-Seq at different expression levels using the inducible promoter system. Rv0081 was induced at 4 different levels of ATc hypoxic conditions, and ChIP-Seq performed. Shown are plots of peaks identified at each of the induction levels. Corresponding peaks are plotted at the same position on the horizontal axis.

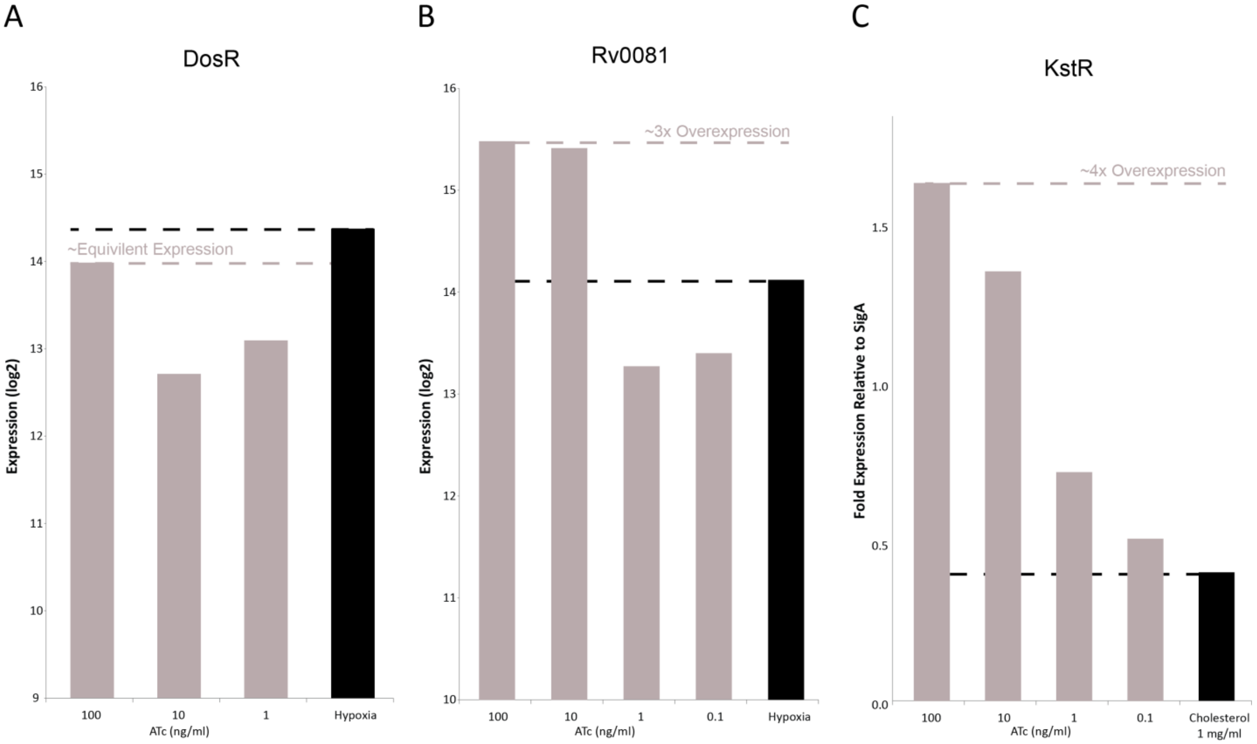


Figure S8: Transcript expression levels in our inducible promoter system compared to physiological stimuli. For three different transcription factors, we measured the relative transcript levels of three different transcription factors in our inducible promoter system at several different levels of induction. We compared these measurements to expression when the TFs were induced from their native promoter with a known physiological stimulus. **(A)** and **(B)** Expression levels of DosR and Rv0081 after two hours of hypoxic stress compared to induction with a range of anhydrotetracycline (ATc) in the inducible system. Expression levels were assayed using our custom Nimblegen array and RNA normalized using methods described in the text. Hypoxia induced DosR to levels comparable to 100 ng/mL ATc. Rv0081 is induced ~3-fold more by 100 ng/ml ATc than its level of induction by hypoxia. Hypoxia expression levels were comparable to levels induced by between 1 and 10 ng/ml ATc. **(C)** Expression levels of KstR induced by cholesterol compared to induction with a range of tetracycline in the inducible system. Wild type M. tuberculosis was grown in 7H9 ADC Tween medium to mid log phase and induced by cholesterol (1 mg/ml as previously described[31]). Expression levels were determined relative to the amount of sigA transcript by quantitative RT-PCR. KstR is induced ~4 fold more than cholesterol at the 100ng/ml ATc level.

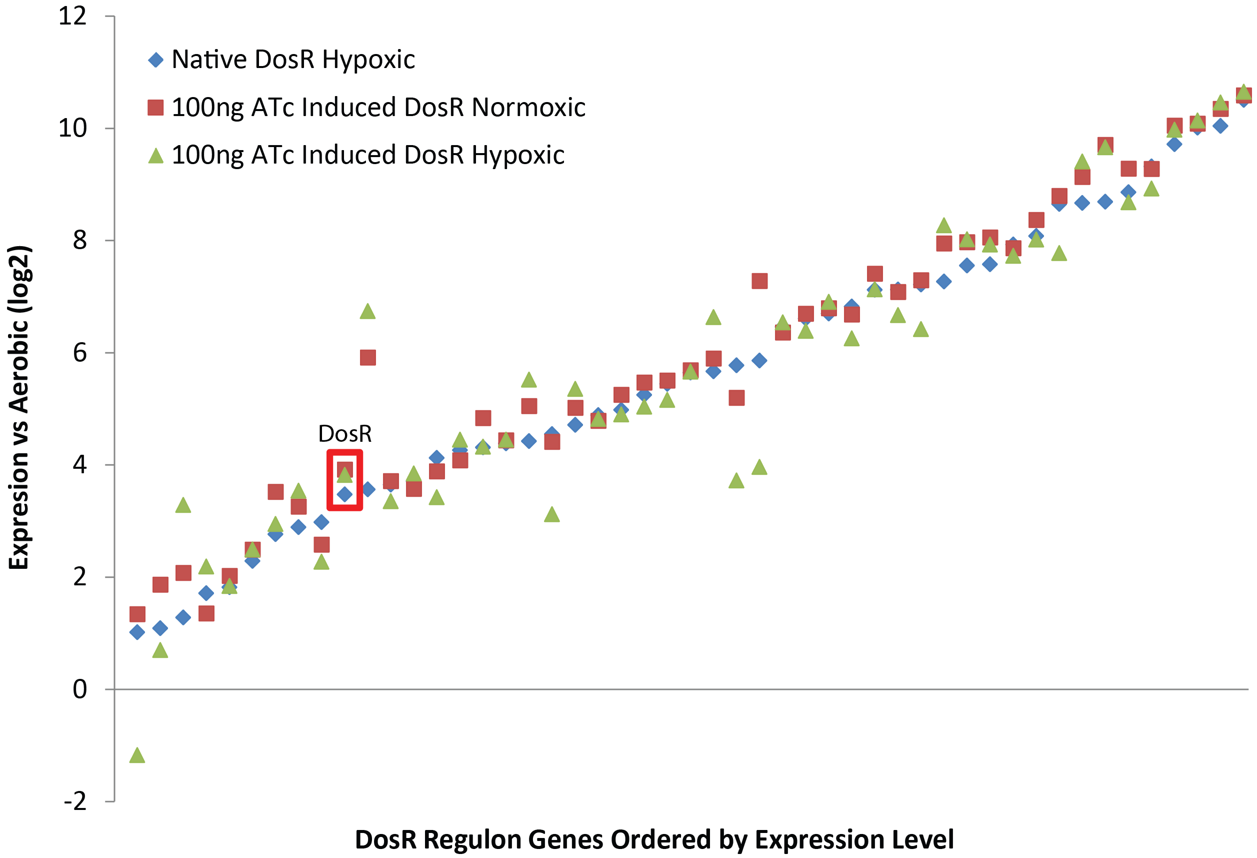


Figure S9: Expression levels of 52 previously documented DosR regulon genes when ectopically induced DosR is stimulated by 100 ng/ml ATC in hypoxia or normoxia compared to the expression when WT DosR is induced by hypoxia. Expression of DosR in our inducible promoter system using the standard 100ng/ml ATc level in either hypoxia or normoxia results in activation of the DosR regulon genes to nearly identical levels as when WT DosR is induced by hypoxia, its physiological stimulus. The known DosR regulon genes[23] are ordered along the X-axis in order of expression level in the WT hypoxia experiment. Expression levels were measured using our Nimblegen array and calculate relative to the expression of each gene in WT aerobic conditions:

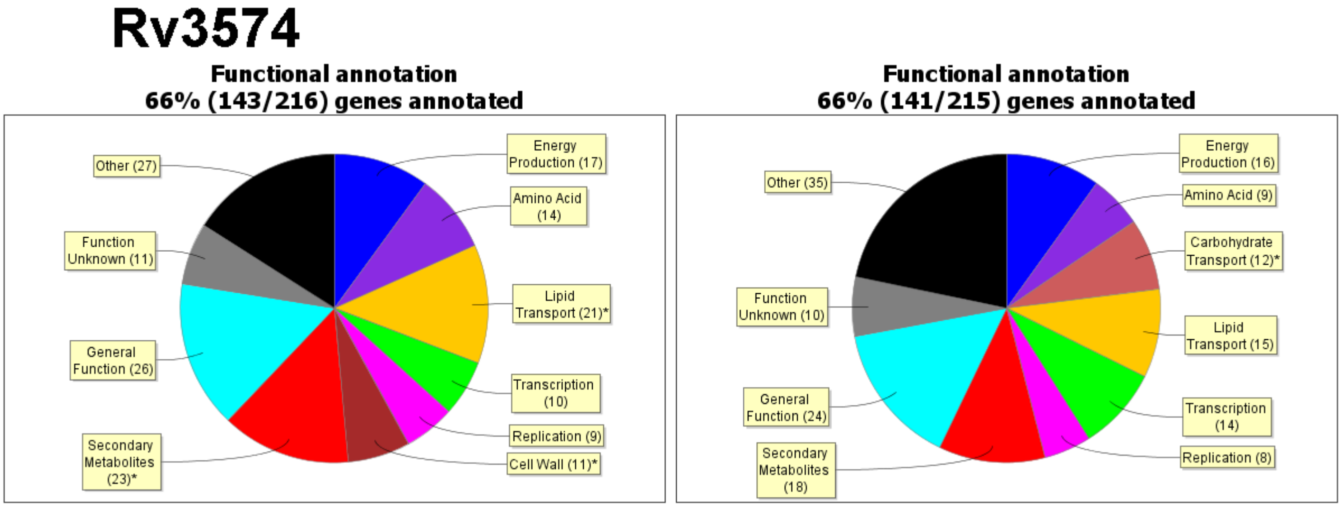


Figure S10: KstR Binding Shows Category Enrichment. The top 50% of strong peaks show similar category enrichment (top left) and the bottom 50% of weakest peaks (top right).

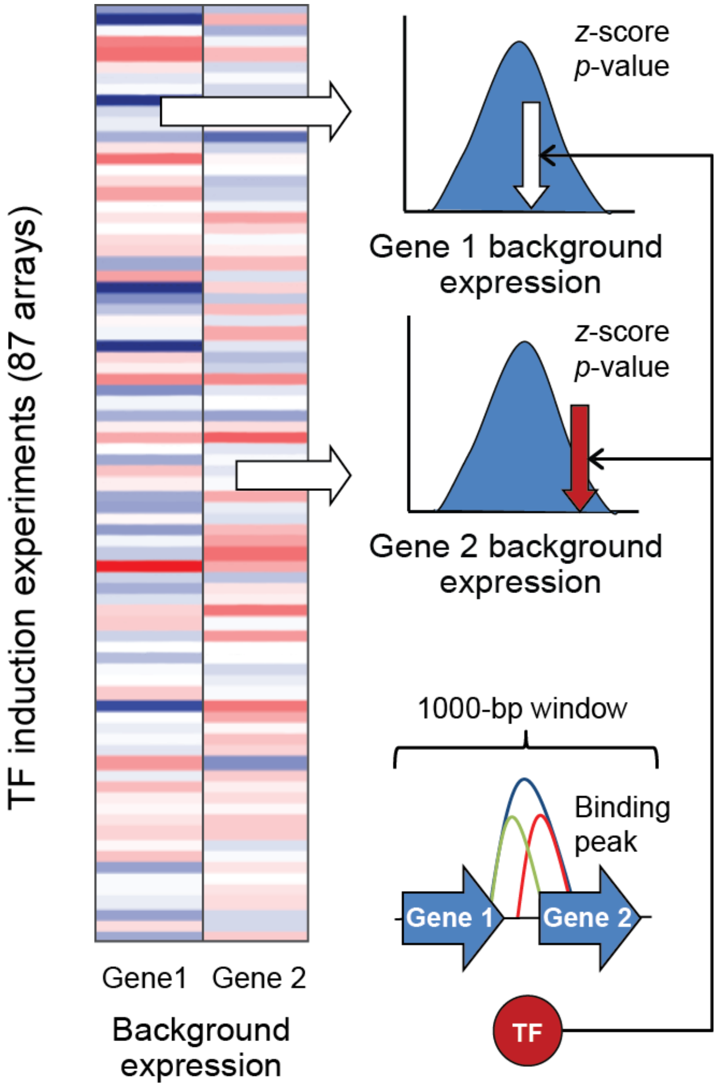


Figure S11: Associating ChIP-Seq Binding with Regulation. For each site, all target genes within 1Kb are examined to determine if induction of the corresponding TF significantly alters target gene expression (different window size results in Figure S13). Target gene expression after TF induction is assigned a z-score (positive=activation, negative=repression) and two-tailed p-value based on a background distribution determined from control experiments. If any target has p-value<0.05 after multiple testing, the peak is assigned a potential regulatory role.

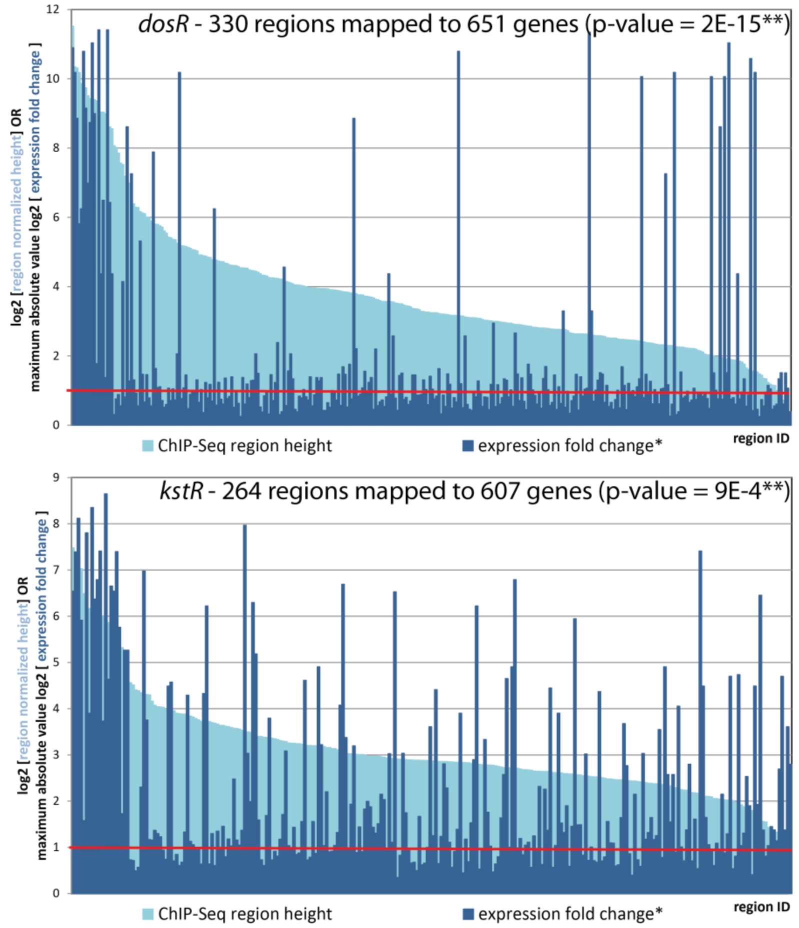


Figure S12: Enrichment of Differential Expression in KO Strains Relative to WT in KstR and DosR Binding Sites. Light blue bars show ChIP-Seq binding heights which are sorted from highest to lowest peaks from left to right along the x-axis. At each peak location, dark blue bars show the highest absolute expression value from the knock-out expression data for the target gene(s) associated with that peak. The p-value associated with each figure represents the probability of the observed overlap between the genes differentially expressed greater than 2-fold (red horizontal line) in the knockout experiments and the ChIP-Seq peaks. (Top) Three DosR datasets are used: overexpression microarray from current project, overexpression RT-PCR from current project, knockout from Park et al[23]. (Bottom) Four KstR expression datasets are used: overexpression microarray from current project, knockout RNA-Seq (unpublished), knockout microarray from Nesbitt et al., and knockout microarray by Kendall et al.[30] **The p-value is obtained from Fisher's exact test: probability of getting the observed number of genes with significant expression fold change and mapped to a ChIP-Seq region compared to a total number of genes mapped to ChIP-Seq regions and total number of genes in the genome that show significant expression fold change.

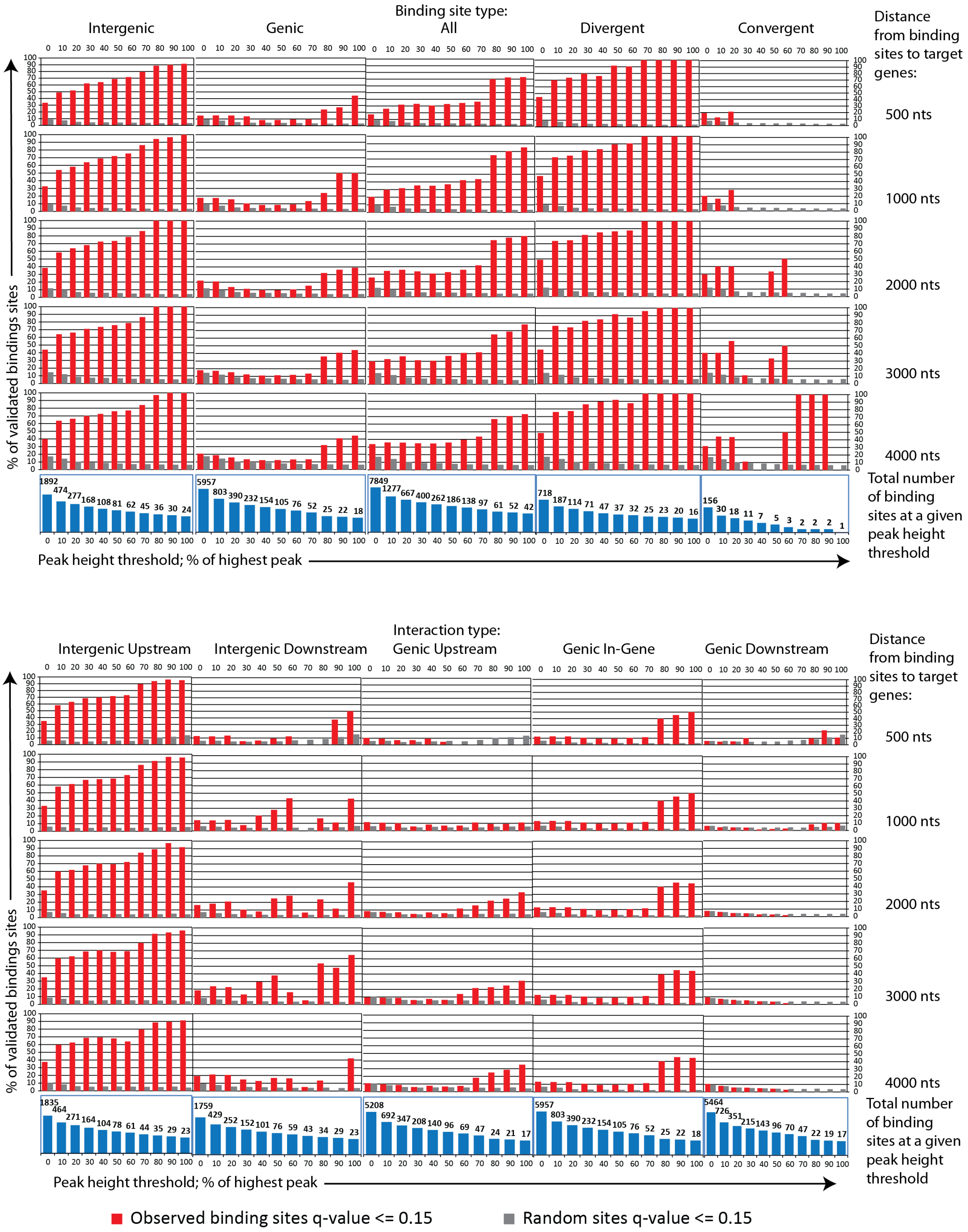

Binding site type:
Intergenic
Genic
All
Divergent
Convergent
Distance from binding sites to target genes:
500 nts
1000 nts
2000 nts
3000 nts
4000 nts
% of validated bindings sites
1892 474 277 168 108 81 62 45 36 30 24
5957 803 390 232 154 105 76 52 25 22 18
7849 1277 667 400 262 186 138 97 61 52 42
718 187 114 71 47 37 32 25 23 20 16
156 30 18 11 7 5 3 2 2 2 1
Total number of binding sites at a given peak height threshold
0 10 20 30 40 50 60 70 80 90 100
Peak height threshold; % of highest peak
Interaction type:
Intergenic Upstream
Intergenic Downstream
Genic Upstream
Genic In-Gene
Genic Downstream
Distance from binding sites to target genes:
500 nts
1000 nts
2000 nts
3000 nts
4000 nts
% of validated bindings sites
1835 464 271 164 104 78 61 44 35 29 23
1759 429 252 152 101 76 59 43 34 29 23
5208 692 347 208 140 96 69 47 24 21 17
5957 803 390 232 154 105 76 52 25 22 18
5464 726 351 215 143 96 70 47 22 19 17
Total number of binding sites at a given peak height threshold
0 10 20 30 40 50 60 70 80 90 100
Peak height threshold; % of highest peak
Observed binding sites q-value <= 0.15
Random sites q-value <= 0.15

Figure S13: Summary of Overall Assignment of Regulation to Binding Events. Peaks are assigned potential regulation as described in Figure 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 eleven overlapping groups by their relative coverage (0-100%, 10-100%, ..., 90-100%, 100%) and test these 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. Red bars correspond to false discovery rate of 15%. Grey bars show the estimated level of validation we expect at random for the same significance level. 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.

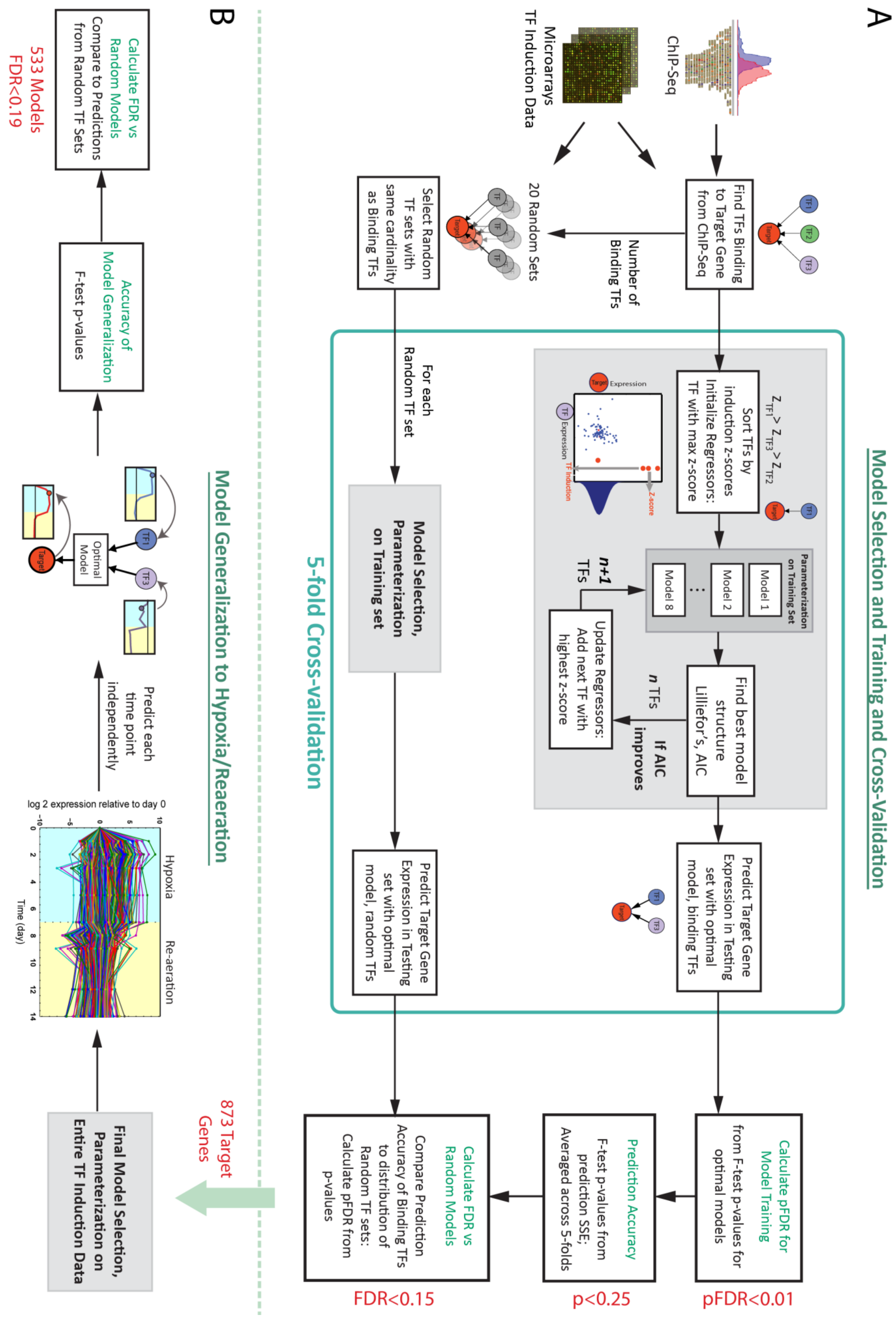

A
ChIP-Seq
Microarrays
TF Induction Data
Find TFs Binding to Target Gene from ChIP-Seq
TF1
TF2
TF3
Target
Number of Binding TFs
20 Random Sets
Select Random TF sets with same cardinality as Binding TFs
Model Selection and Training and Cross-Validation
$Z_{TF1} > Z_{TF3} > Z_{TF2}$
Sort TFs by induction z-scores
Initialize Regressors: TF with max z-score
Expression
TF Expression
TF Induction
z-score
Parameterization on Training Set
Model 1
Model 2
...
Model 8
n+1 TFs
Find best model structure
Lilliefor's, AIC
n TFs
If AIC improves
Update Regressors: Add next TF with highest z-score
Predict Target Gene Expression in Testing set with optimal model, binding TFs
5-fold Cross-validation
For each Random TF set
Model Selection, Parameterization on Training set
Predict Target Gene Expression in Testing set with optimal model, random TFs
Calculate pFDR for Model Training from F-test p-values for optimal models
pFDR<0.01
Prediction Accuracy
F-test p-values from prediction SSE; Averaged across 5-folds
p<0.25
Calculate FDR vs Random Models
Compare Prediction Accuracy of Binding TFs to distribution of Random TF sets: Calculate pFDR from p-values
FDR<0.15
873 Target Genes
B
Model Generalization to Hypoxia/Reaeration
Final Model Selection, Parameterization on Entire TF Induction Data
Hypoxia
Re-aeration
log 2 expression relative to day 0
10
5
0
−5
−10
0
2
4
6
8
10
12
14
Time (day)
Predict each time point independently
Optimal Model
TF1
TF3
Target
Accuracy of Model Generalization
F-test p-values
Calculate FDR vs Random Models
Compare to Predictions from Random TF Sets
533 Models
FDR<0.19

Figure S14: Schematic of approach used for gene expression modeling and validation. The approach consists of two parts. **(A)** Model selection and training was performed using ChIP-Seq binding data and TF induction microarray data all generated in normoxic conditions. For each gene, 8 different model structures were considered. The relationship between TFs and target genes were parameterized based on subsets of the overexpression data and tested on the remaining using cross-validation. Individual model fits were validated with a Lilliefor's normality test on residuals and overfitting was corrected with the use of AIC. The best model structure was selected for each gene. We calculated two FDR estimates for each gene model: an analytical pFDR based on F-test p-values using the method of Storey, and an empirical FDR based on comparison to random TFs. **(B)** To assess the ability of models validated in the first step to generalize to another independent data set, models were tested on their ability to predict the expression of target genes during the time course of hypoxia and re-aeration. Each time point during hypoxia and re-aeration was predicted separately and independent of previous time points. We again calculated both an analytical and empirical FDR. A total of 533 models predicted the hypoxia time course with significant accuracy and better than random TFs with an FDR<0.19.

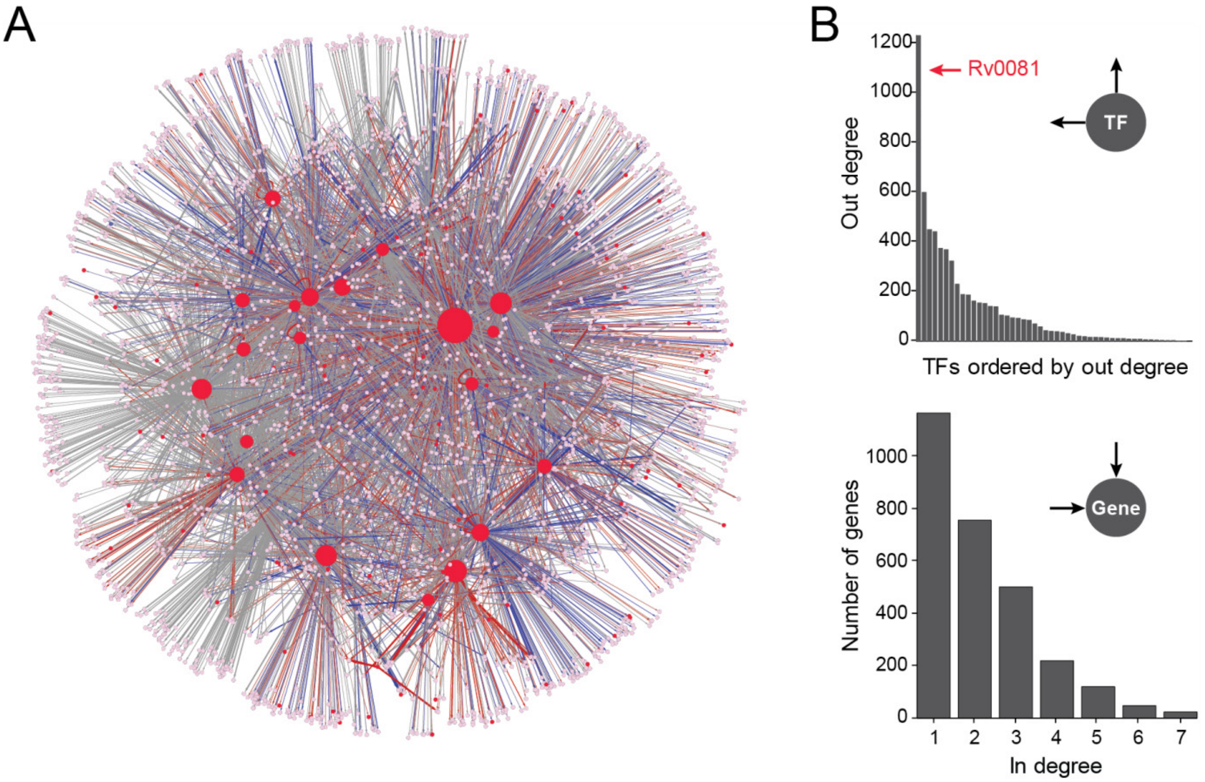


Figure S15: *M. tuberculosis* Regulatory Network. **(A)** Transcriptional regulatory network based on ChIP-Seq binding and TF induction expression data for 50 transcription factors. The network encompasses 2704 genes, including 141 transcription factors, and 5387 TF-gene interactions based on 9865 binding 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 the main text 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. Only binding peaks greater than 1% the height of the maximum peak for each TF were utilized. **(B)** Out-degree and in-degree for all TFs. Out-degree for each TF is plotted in the top figure ordered by out-degree. A distribution of in-degree for all target genes is shown in the bottom figure.

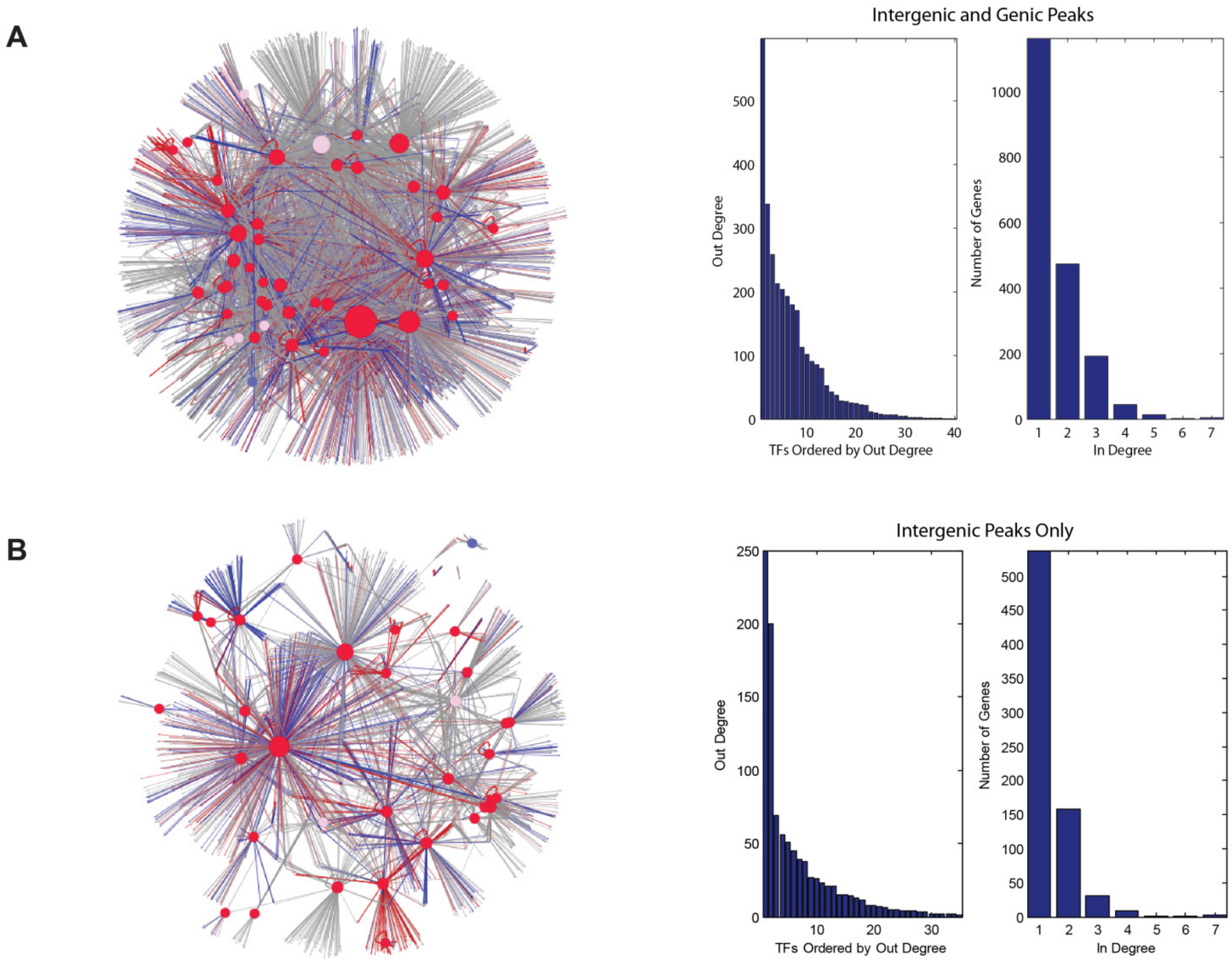


Figure S16: Regulatory Network Models using Different Criteria for Including TF-Gene Links. (A) Network including links between a TF and target gene if the TF binds in the upstream intergenic region or in the target gene itself independent of possible predicted regulatory role. **(B)** Network only including links if the TF binds in the upstream intergenic region independent of possible predicted regulatory role. A dynamic version of these regulatory network maps is available in a cytoscape file in supplementary material (Cytoscape available at http://www.cytoscape.org/).

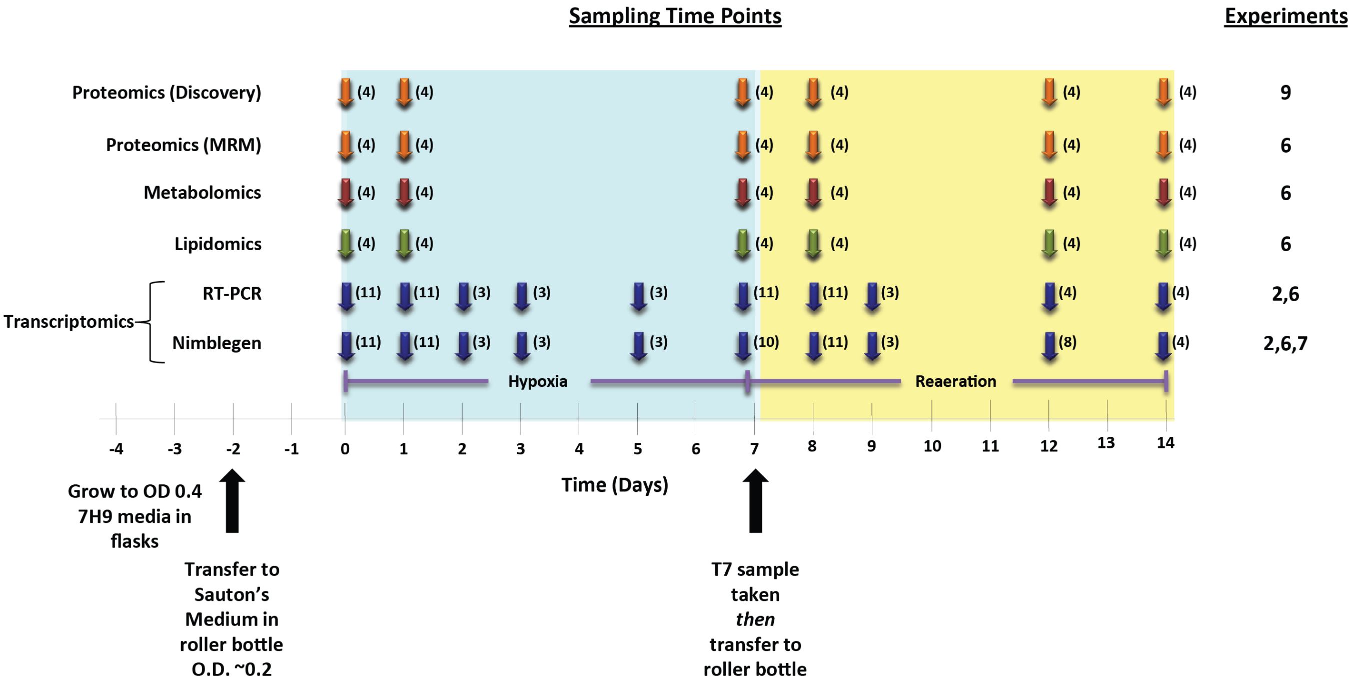


Figure S17: Hypoxia Profiling Sampling Protocol – Schematic of the time points sampled for each of the profiling modalities used. For all experiments, MTB was grown to OD 0.4 in 7H9 medium, transferred to Sauton's medium (see Methods) in a roller bottle and grown for two days prior to experimental sampling. 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. Sample for profiling were taken at the time points indicated for each modality. The number in parenthesis indicates the number of replicates generated at each time point. Proteomics, Metabolomics, and Lipidomics were generated from a single experiment (experiment 6) with 4 biological replicates. Transcriptomics samples were generated from 3 different experiments with multiple biological replicates in each experiment. The breakdown of replicates for each transcriptomic time point from each experiment is shown is described in the Supplemental Methods.

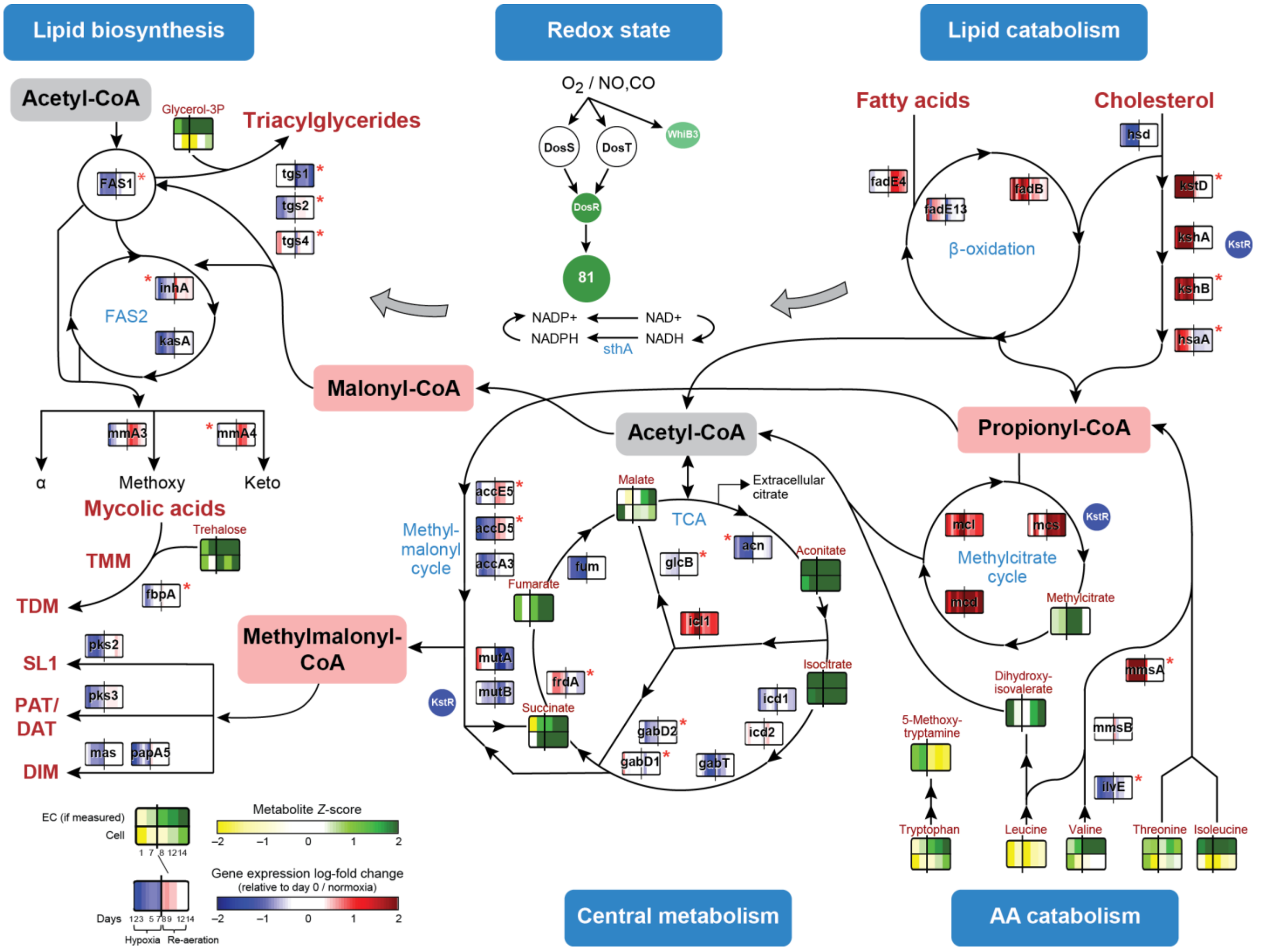


Figure S18: ***M. tuberculosis*** **Metabolic Overview Integrated with Global Profiling Data and Regulation.** Overview of MTB metabolic network focused on lipid, amino acid, and central metabolism and redox sensing. Log fold expression data relative to baseline (day zero) during hypoxia and re-aeration shown as red/blue heat maps (see legend). Z-scores for metabolite changes relative to baseline are shown and green/yellow heat maps. Genes that vary significantly and whose pattern of expression can be predicted (see main text) are marked with an asterisk. Full model available as a Cytoscape file in supplementary material (Cytoscape available at http://www.cytoscape.org/).

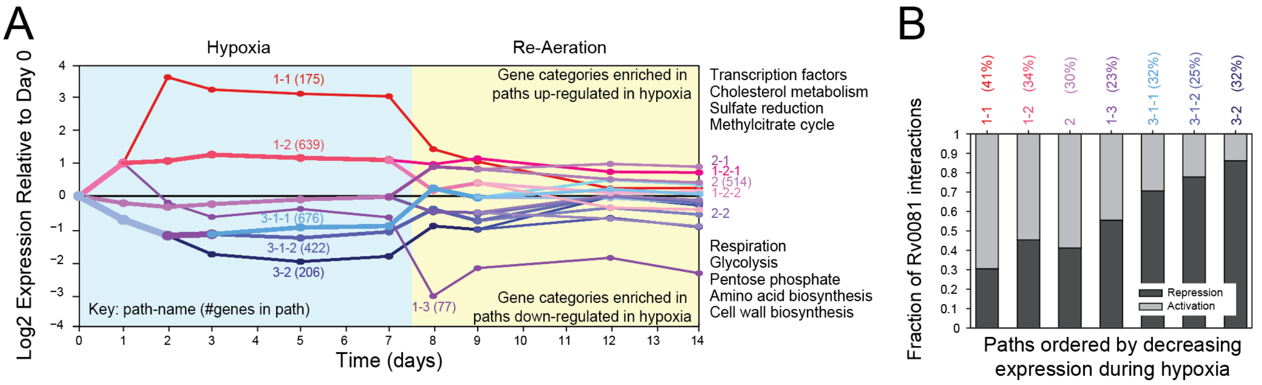


Figure S19: Rv0081 predicts the overall expression of genes in each path during hypoxia. (A) Gene expression during hypoxia and re-aeration was clustered using DREM into paths (sets of genes with similar expression profiles). **(B)** Rv0081 is activated during hypoxia and repressed during re-aeration. Histogram shows fraction of Rv0081 interactions predicted as activating or repressive for each path. X-axis labels indicate the path name and fraction of genes in that path bound by Rv0081. Rv0081 binds 25% to 41% of genes in each path. Activated paths have predominantly activating interactions, and vice versa for down-regulated paths. Regulatory role was not used to cluster genes.

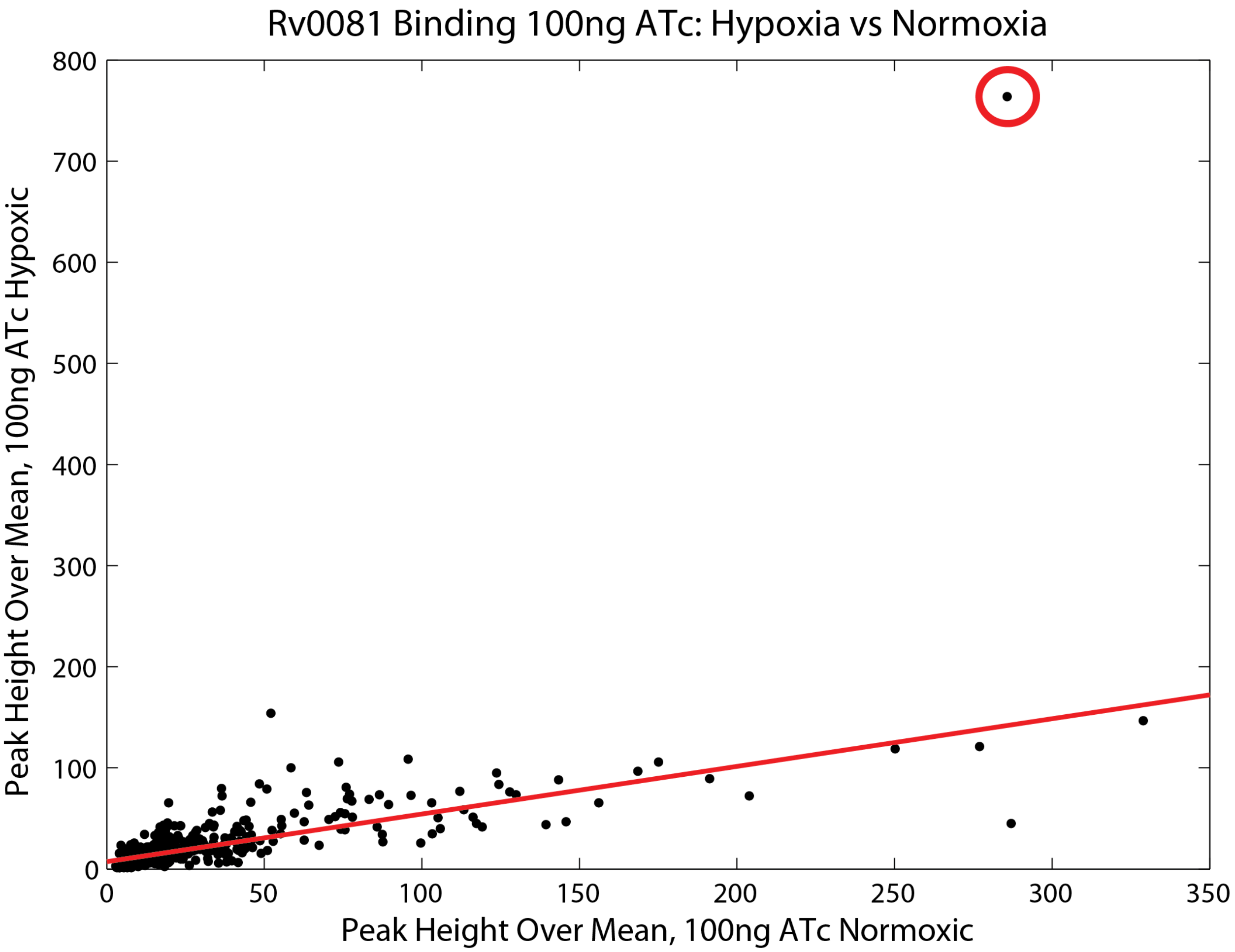


Figure S20: The same binding sites are found for Rv0081 with the inducible promoter system in both normoxia and hypoxia. Region coverage for peaks found with the inducible promoter system in hypoxia is plotted against the coverage for the inducible promoter system in normoxia. The same regions are found in both cases with highly correlated peak heights with one exception: the autobinding peak highlighted by the red circle. Although the autobinding peak was found in both conditions, during hypoxia Rv0081 appears to bind more strongly to its promoter than during normoxia. Red line shows a linear regression with the autobinding peak excluded (R^2=0.68).

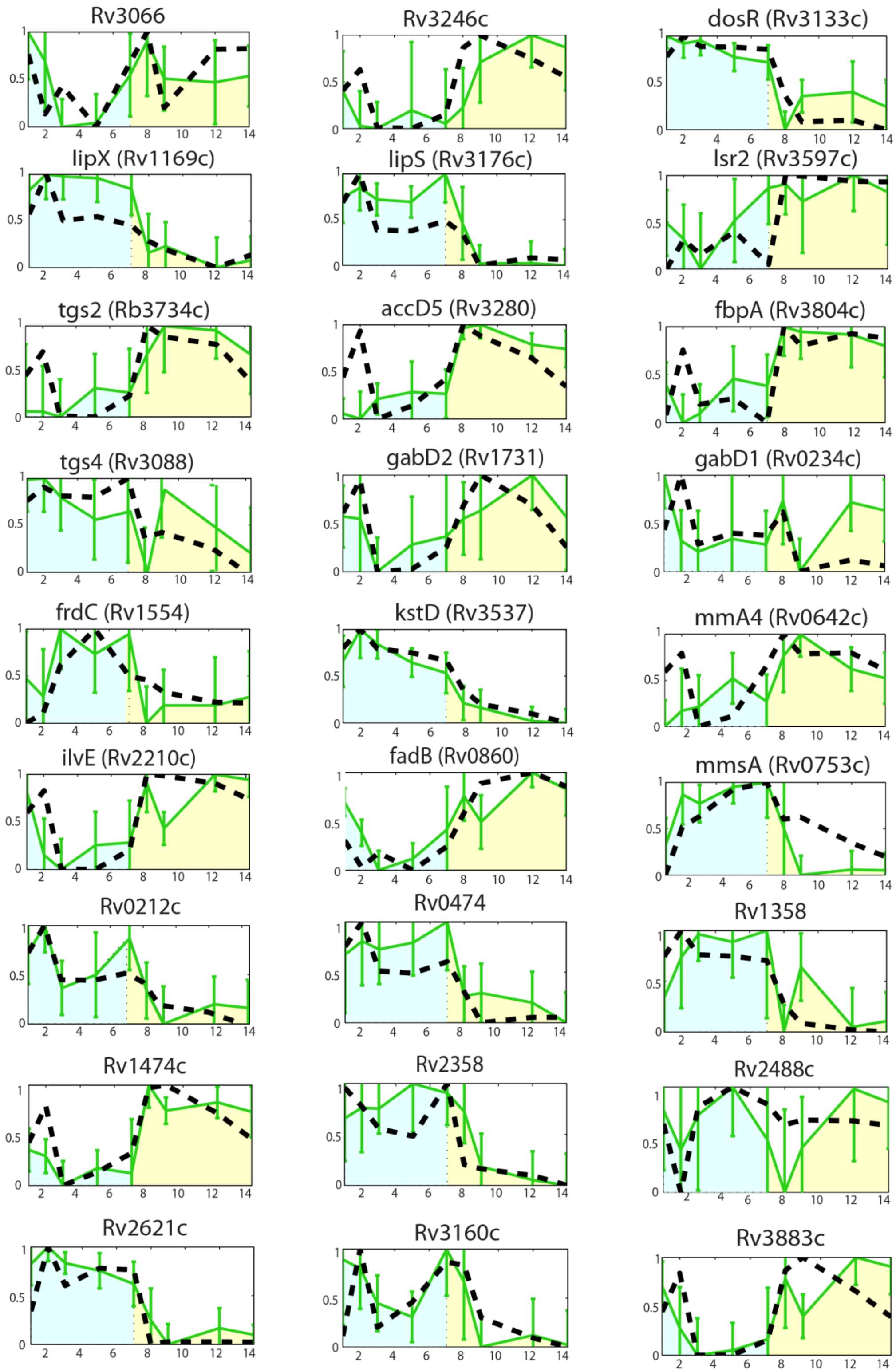


Figure S21: Prediction of Hypoxia and Re-aeration Gene Expression for Specific Genes Mentioned in Text. Prediction of expression patterns during hypoxia and re-aeration time course for additional genes mentioned in the main text. Data generated and plotted as describe in Figure 5C in the main text.

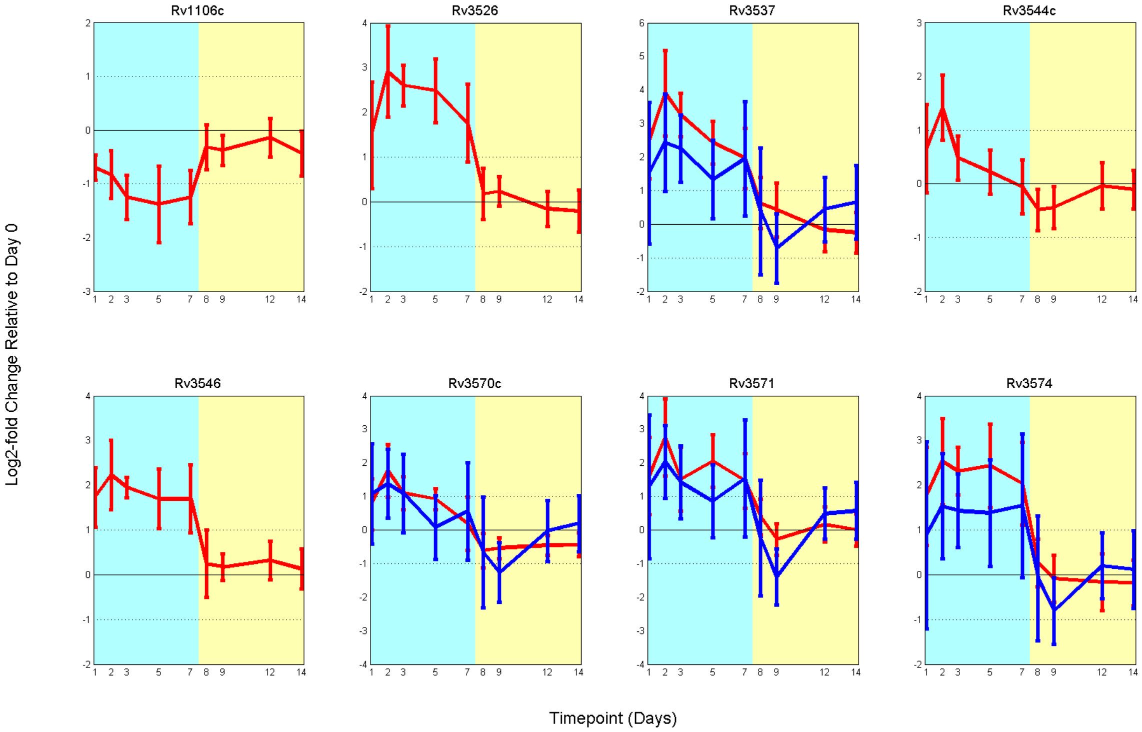


Figure S22: Cholesterol Related Gene Expression – Expression of genes implicated in cholesterol degradation during hypoxia and re-aeration. For each gene, expression as measured by Nimblegen tiling arrays are shown in red with standard deviations shown for each time point. Corresponding RT-PCR measurements, where available (see Methods) shown in blue. Cyan background indicates the hypoxia time points, yellow indicates re-aeration time points. All data plotted as Log2 fold change relative to day zero.

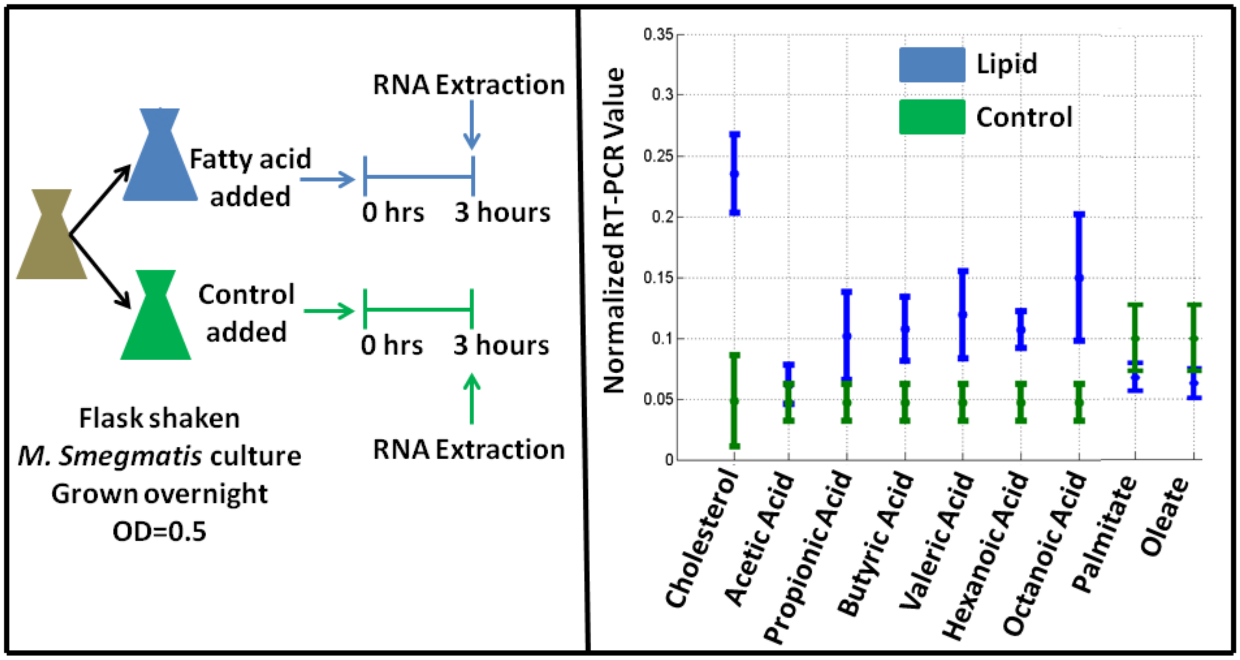


Figure S23: Cholesterol, Propionate, and Small Fatty Acids Relieve Repression by KstR in *M. smegmatis*. *M. smegmatis* was grown overnight in rich media (7H9 supplemented with ADC and 0.05% Tween80). When the culture reached an optical density of 0.5 the following day, it was divided into smaller flasks each of which held 40mL of liquid culture. In the experimental flasks one of 9 compounds was added to the culture: cholesterol (250uM), acetic acid (10mM), propionic acid (10mM), butyric acid (10mM), valeric acid (10mM), hexanoic acid (10mM), octanoic acid (10mM), sodium palmitate (250uM) or sodium oleate (250uM). In the fatty acid control flasks a pH control (HCl) was added, which lowered the pH to the same level created by the fatty acids. For cholesterol control flasks we added a vehicle control, which was 80uL of a 3:1 mixture of ethanol and tyloxapol. After 3 hours RNA was extracted from each flask following standard RNA extraction protocols and RT-PCR was carried out using a QuantiFast SYBR Green PCR kit from Qiagen. Primers were designed for KstR (MSMEG_6042)) that amplified a 133 bp region 161 bp after the first nucleotide of the start codon and for SigA (MSMEG_2758) that amplified a 130 bp region 818 bp after the first nucleotide of the start codon. The left panel provides an overview of the experimental design. The right panel shows the normalized RT-PCR value for KstR: $\mathbf{2}^{(Ct_{SigA}-Ct_{KstR})}$ for each lipid condition.

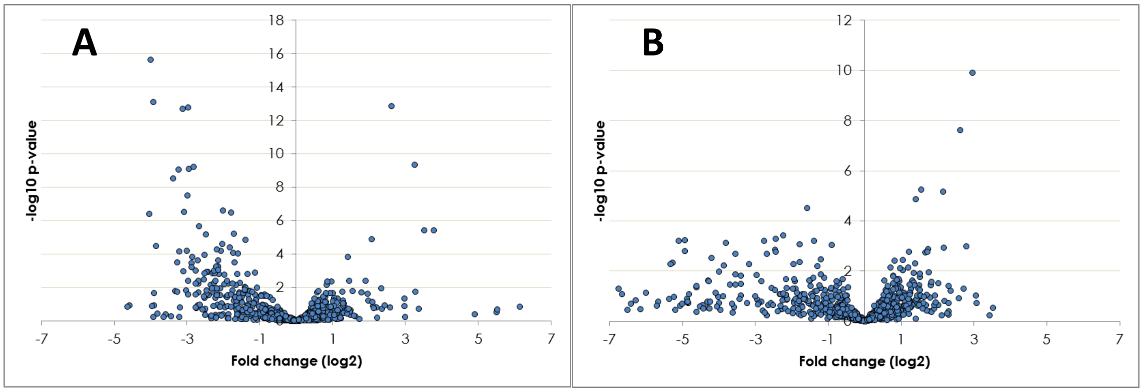


Figure S24: Changes in Protein Levels during Hypoxia and Re-aeration. Volcano plots showing statistical significance (y-axis) in relation to fold-change (x-axis). Each data point represents a distinct protein, for a total of 1,000 proteins. (A) Day 7 versus Day 0. At a p-value <0.05 and q-value <0.05, 109 proteins are decreased by two-fold or more and 19 proteins are increased. (B) Day 11 versus Day 7. At a p-value <0.05 and q-value < 0.05, 72 proteins are decreased by two-fold or more and 31 proteins are increased.

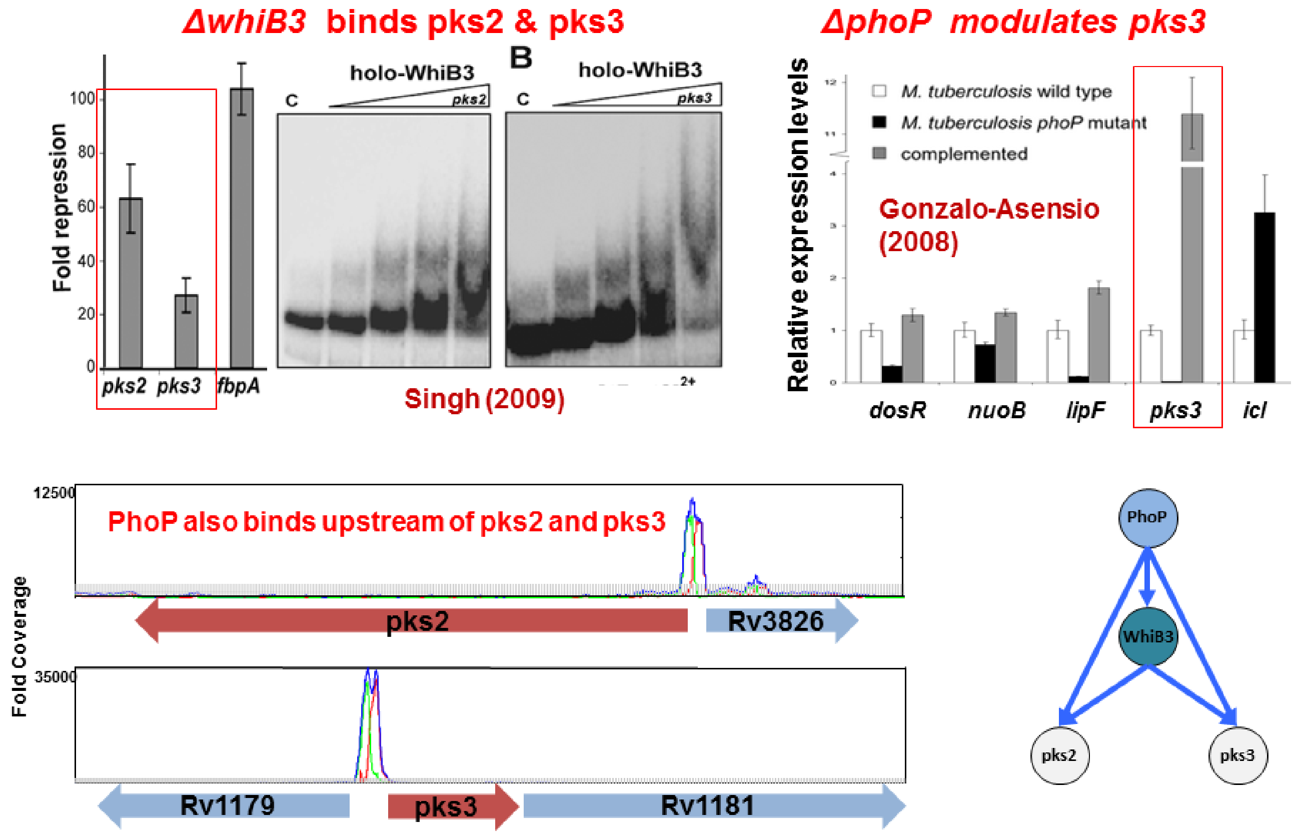


Figure S25: Predicted Feed-Forward Loop Mediating Convergent Regulation of pks2/3 by WhiB3 and PhoP. WhiB3 is known to binding and modulate the expression of pks2 and pks3 (Singh 2009). A PhoP knockout is known to dis-regulate pks2. Our data indicate that PhoP binds in association with both WhiB3 and pks2/pks3, forming a potential feed-forward loop. WhiB3 is upregulated throughout hypoxia and re-aeration and, thus, although WhiB3 is required for the production of PAT/DAT and SL1, our data indicates it is not sufficient. We predict that Rv0081 represses pks3, and during hypoxia this may act to override activation by both WhiB3 and PhoP.

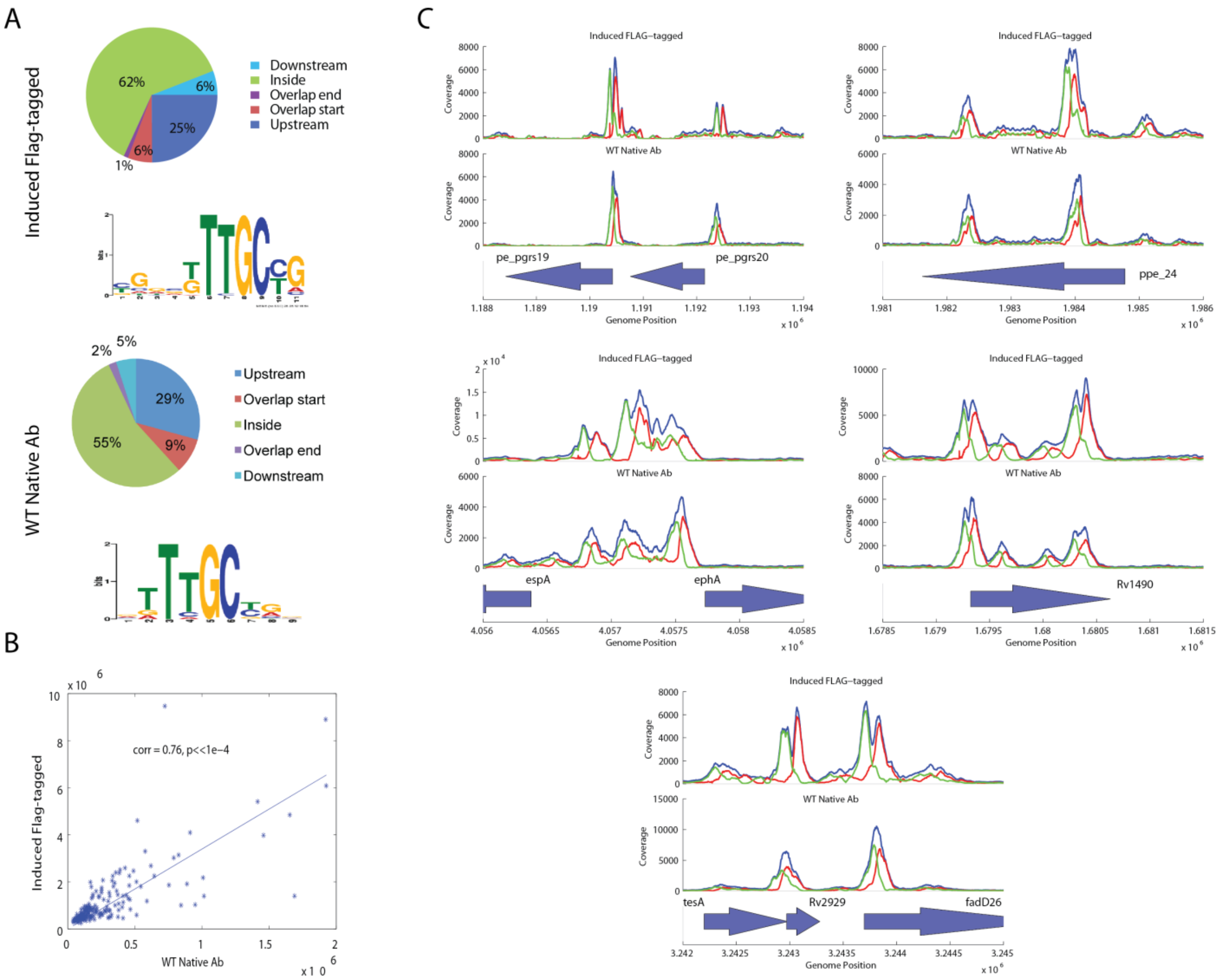


Figure S26: Comparison of EspR ChIP-Seq for WT Native Ab and Induced FLAG-tagged. ChIP-Seq using our induced and FLAG-tagged system shows a high level of agreement with the results of Blasco et al. (2012)[5] using a native antibody to WT EspR. **(A)** The distribution of binding site location and the recovered motifs are highly similar between the two methods. Top part shows the results from induced FLAG-tagged EspR. Bottom part is taken from Blasco et al. (2012). **(B)** Peak height shows high level of correlation between the two techniques. Peaks were called from the induced FLAG-tagged data and the peak heights compared to the coverage from the raw data of Blasco et al. (2012). **(C)** Comparison of specific binding regions shows high level of detailed binding agreement between the two methods. For each plot, the top half shows the results of our induced FLAG-tagged system and the bottom half shows the results from WT native Ab.

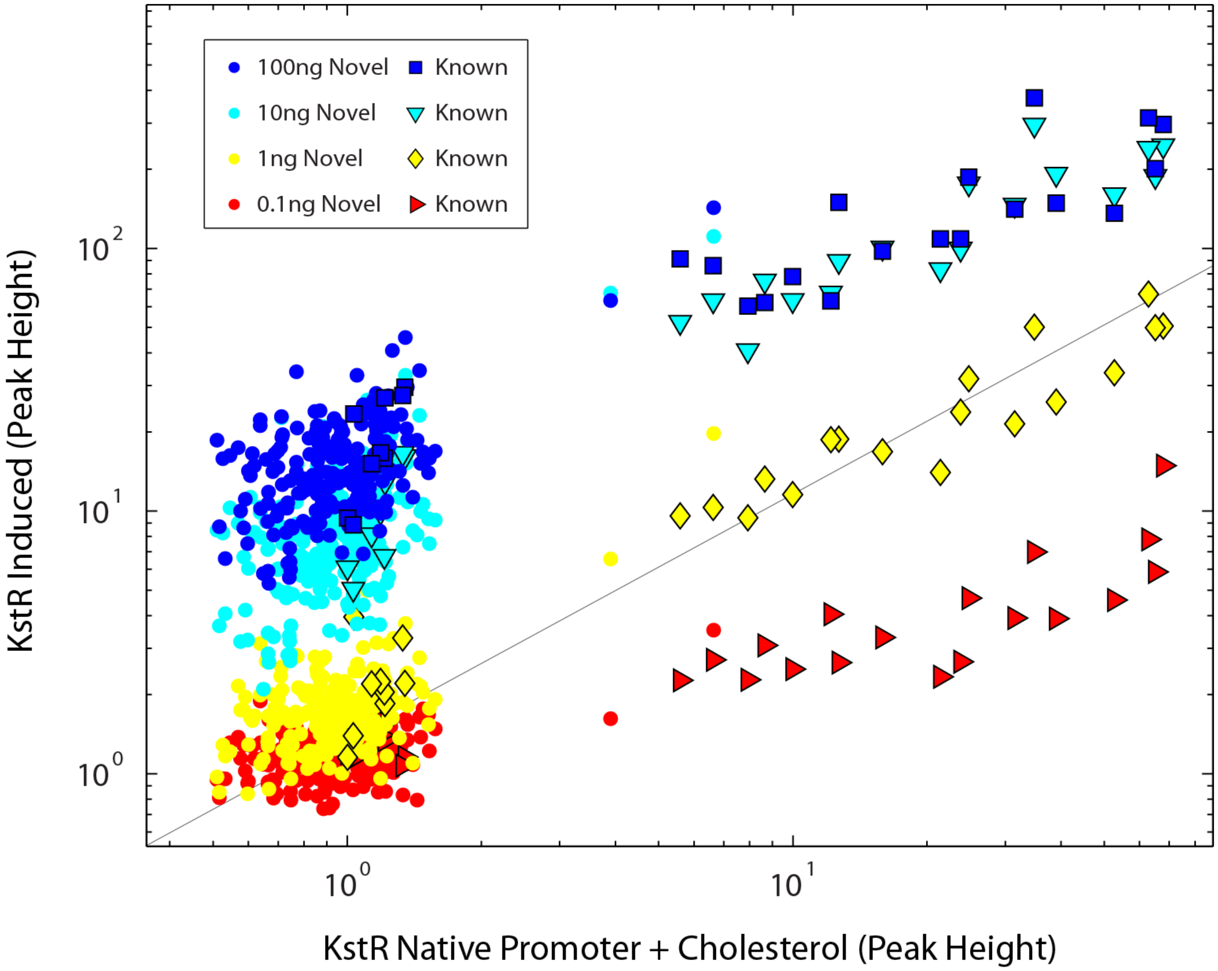


Figure S27: ChIP-Seq peak heights for KstR cholesterol induction versus ATc induction. Plotted are heights for corresponding peaks in ChIP-Seq of KstR at 4 different levels of ATc induction versus induction by cholesterol. The peaks plotted are those found at the 100ng/ml ATc level in the inducible system. Cholesterol induction ChIP-Seq was performed on FLAG tagged KstR with its native promoter integrated into the WT chromosome (i.e. single copy).

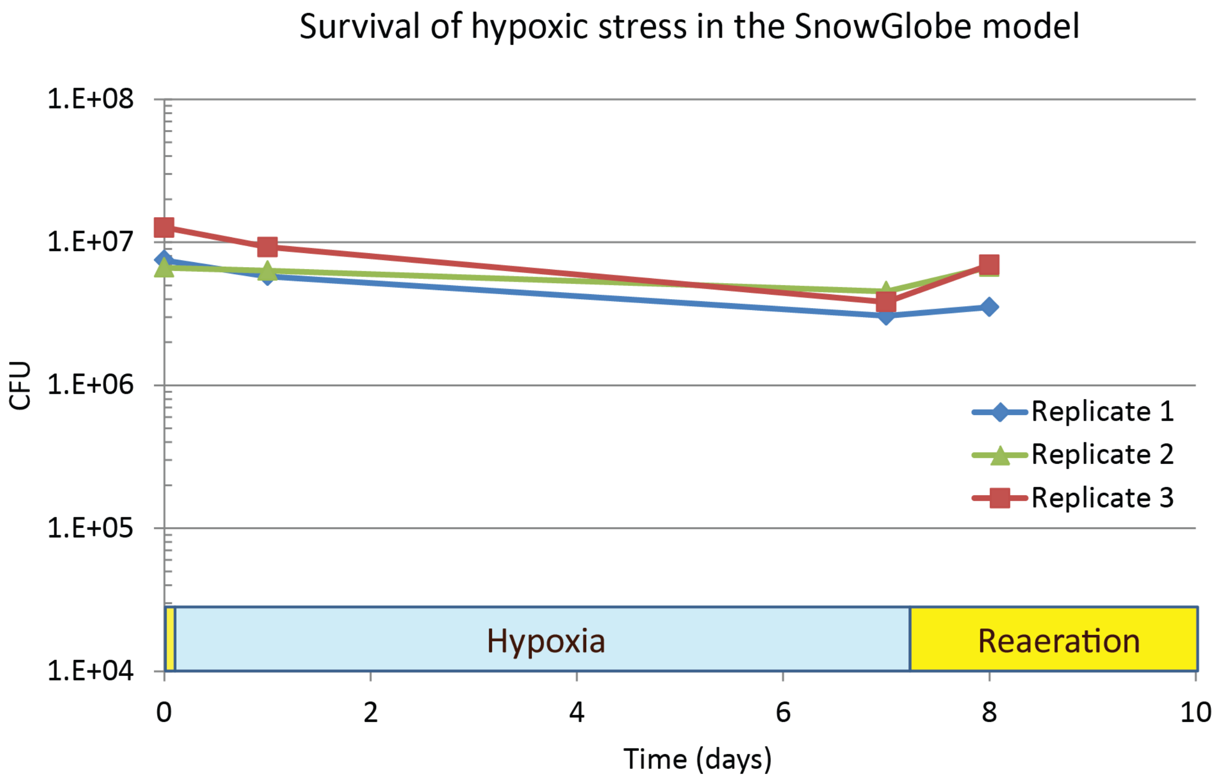


Figure S28: Survival of hypoxic stress in the SnowGlobe model. To gauge the amount of death that occurs over the hypoxic time course we used Tween80 and vigorous agitation to disperse clumps and plated for CFU. In three independent experiments small but noticeable decrease in viability were observed after 7 days of hypoxic stress. Error bars indicate the standard deviation from four biological replicates. These data replicate previously published results in a model of hypoxia in the presence of detergent[48].

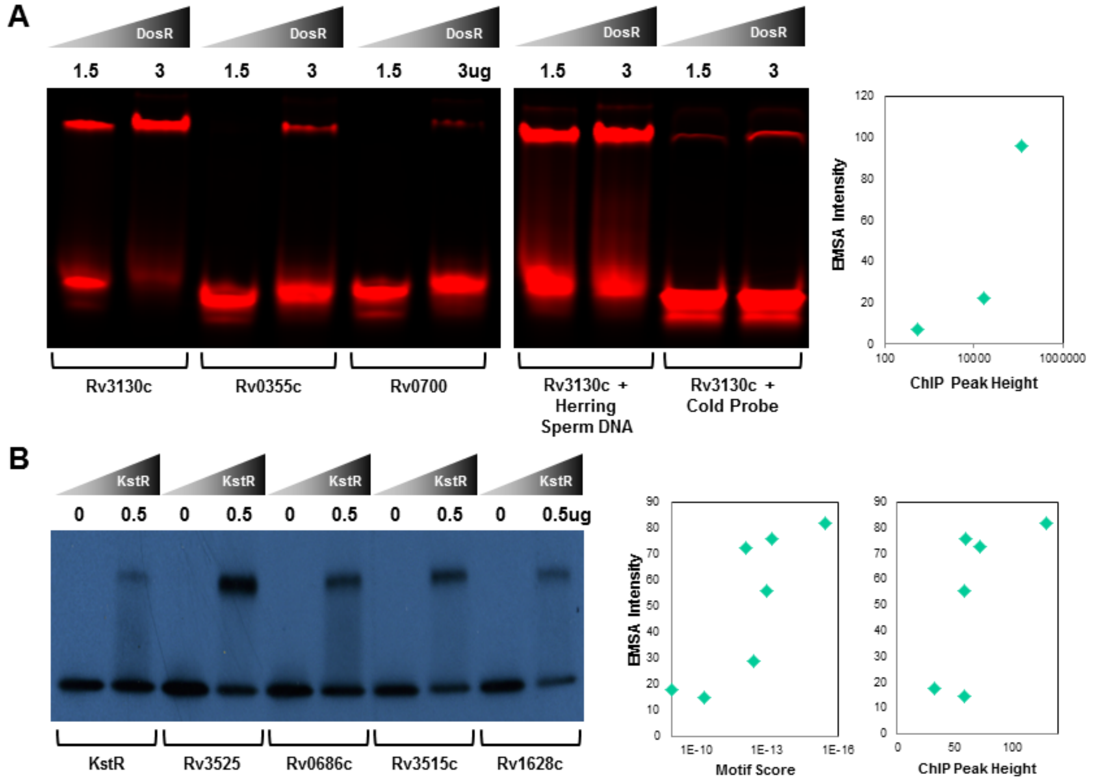


Figure S29: EMSA Validation of Selected Binding Events. (A) EMSA validation for DosR binding sites associated with three different genes. All binding sites were located in the proximal promoters of the genes and encompass a range of ChIP peak heights. In the first three pairs of lanes, 1.5ug and 3ug phosphorylated DosR were incubated with 4 pmol IRDye700 labeled probe in the presence of 1ug poly (dI:dC) for 20 min at 30°C. In the last 2 pairs of lanes, 1.5ug and 3ug phosphorylated DosR were incubated with 4 pmol IRDye700 labeled probe and competed with either 20ug herring sperm DNA or 600 pmol unlabeled (cold) probe. Samples were electrophoresed on a 10% nondenaturing TBE polyacrylamide gel in 0.5x TBE at 80V and 4°C in the dark. Gels were scanned using the Odyssey Classic Infrared Imager (LiCor). Scatter plot shows intensities expressed as kilo integrated intensity per mm2 compared to ChIP peak height (fold coverage on log scale). Phosphorylation and binding conditions were essentially as described previously[104]. Briefly, purified DosR was buffer-exchanged into 40mM Tris-HCl, pH 8.0 and 5mM MgCl2. DosR was phosphorylated by incubating 50ug protein in a 50ul volume for 30min at 30°C with 50mM Lithium Potassium Acetyl Phosphate (Sigma) in 40mM Tris-HCl, pH 8.0 and 5mM MgCl2. After phosphorylation binding was performed in a 20ul reaction with 4pmol of IRDye700 labeled probe for 20min at 30°C in binding buffer (25mM Tris-HCl, pH 8.0, 0.5mM EDTA, 20mM KCl, 6mM MgCl2, 5% glycerol, and 1ug of poly (dI:dC) (Sigma)). 60 nt long oligonucleotides labeled on the 5'end with IRDye700 were purchased from IDT and ds DNA probes were obtained by annealing the respective sense and antisense strands. **(B)** EMSA validation for selected KstR binding sites. Sites were chosen from both genic and intergenic DNA regions, and encompassed a range of ChIP-Seq peak heights and motif qualities. DNA targets were labeled using the DIG-ddUTP end-labeling system (Roche). 0 or 0.5 µg of KstR-6 His tagged protein and 1 ng of labeled target DNA were mixed in a DIG EMSA binding buffer, incubated for 15 minutes at room temperature, and separated in a 7.5% polyacrylamide gel at 65V for approximately 2 hours. Scatter plots show EMSA intensity (calculated as 0-255 grayscale average comparing shift band intensity against image background) with motif score and ChIP Peak Height for each binding site.

4. Supplementary Tables

Table S1: 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 S2: Summary of Proteomic Changes – Discovery LC-MS/MS Method (Experiment 9). Number of proteins meeting statistical significance and differential expression thresholds for [Day 7 vs. Day 0] and [Day 11 vs. Day 7] comparisons is shown. A file with all protein measurements is provided in supplementary material.

<table>
<tr><th></th><th></th><th colspan="3">p-value <0.05 | q-value <0.05 | DI>2</th></tr>
<tr><th>TIME POINTS</th><th>CONDITIONS</th><th>DOWN</th><th>UP</th><th>TOTAL</th></tr>
<tr><td>Day 7 vs. Day 0</td><td>[Hypoxia vs. Log Phase]</td><td>109(85%)</td><td>19 (15%)</td><td><b>128</b></td></tr>
<tr><td>Day 11 vs. Day 7</td><td>[Re-aeration vs. Hypoxia]</td><td>72 (70%)</td><td>31 (30%)</td><td><b>103</b></td></tr>
<tr><td></td><td></td><td></td><td></td><td><b>138</b></td></tr>
</table>

Table S3: Summary of Proteomic Changes – Targeted MRM Method (Experiment 6). Number of proteins meeting statistical significance and differential expression thresholds for [Day 7 vs. Day 0] and [Day 11 vs. Day 7] comparisons is shown. A file with all protein measurements is provided in supplementary material.

<table>
<tr><th></th><th></th><th colspan="3">p-value <0.05 | q-value <0.05 | DI>2</th></tr>
<tr><th>TIME POINTS</th><th>CONDITIONS</th><th>DOWN</th><th>UP</th><th>TOTAL</th></tr>
<tr><td>Day 7 vs. Day 0</td><td>[Hypoxia vs. Log Phase]</td><td>31 (84%)</td><td>6 (16%)</td><td><b>37</b></td></tr>
<tr><td>Day 14 vs. Day 7</td><td>[Re-aeration vs. Hypoxia]</td><td>5(38%)</td><td>8 (62%)</td><td><b>13</b></td></tr>
<tr><td></td><td></td><td></td><td></td><td><b>44</b></td></tr>
</table>

Table S4: Summary of gene prediction results.

		5-fold Cross-validation on TF Induction Data (same condition)			**Generalization to Hypoxia Time Course (independent condition)**		
	Total genes with Binding TFs (peak impulse>1%)	Optimal Models Trained (F-test p<0.1) (pFDR<0.01)	Predicted with Binding TFs (F-test p <0.25)	Binding TFs predict better than average Random TFs (FDR<0.15)	>2-fold change in hypoxia expression	Predicted with Binding TFs (F-test p<0.25)	Binding TFs predict better than average Random TFs (FDR<0.19)
Genes	3072	2755	953	873	808	651	533
%		89% of total	36% of trained models	32% of trained models		80% of changing genes	66% of changing genes

Table S5: Sampling replicates for transcriptomics from hypoxia time course.

	Total # of replicates	SG2 replicates	SG6 replicates	SG7 replicates
Day 0	11	3	4	4
Day 1	11	3	4	4
Day 2	3	3		
Day 3	3	3		
Day 5	3	3		
Day 7	10	3	4	4
Day 8	11	3	4	4
Day 9	3	3		
Day 11	4		4	
Day 12	4			4
Day 14	4		4	

5. References

1 MacQuarrie, K. L., Fong, A. P., Morse, R. H. & Tapscott, S. J. Genome-wide transcription factor binding: beyond direct target regulation. *Trends in genetics : TIG* **27**, 141-148, doi:10.1016/j.tig.2011.01.001 (2011).

2 Farnham, P. J. Insights from genomic profiling of transcription factors. *Nature reviews. Genetics* **10**, 605-616, doi:10.1038/nrg2636 (2009).

3 Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res* **16**, 962-972, doi:10.1101/gr.5113606 (2006).

4 Vilar, J. M. & Saiz, L. DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise. *Curr Opin Genet Dev* **15**, 136-144, doi:10.1016/j.gde.2005.02.005 (2005).

5 Blasco, B. *et al.* Virulence regulator EspR of Mycobacterium tuberculosis is a nucleoid-associated protein. *PLoS Pathog* **8**, e1002621, doi:10.1371/journal.ppat.1002621 (2012).

6 Blasco, B. *et al.* Atypical DNA recognition mechanism used by the EspR virulence regulator of Mycobacterium tuberculosis. *Mol Microbiol* **82**, 251-264, doi:10.1111/j.1365-2958.2011.07813.x (2011).

7 Hunt, D. M. *et al.* Long-range transcriptional control of an operon necessary for virulence-critical ESX-1 secretion in Mycobacterium tuberculosis. *J Bacteriol* **194**, 2307-2320, doi:10.1128/JB.00142-12 (2012).

8 Zeitlinger, J. *et al.* Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**, 395-404 (2003).

9 Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117, doi:10.1016/j.cell.2008.04.043 (2008).

10 Gautam, U. S., Chauhan, S. & Tyagi, J. S. Determinants outside the DevR C-terminal domain are essential for cooperativity and robust activation of dormancy genes in Mycobacterium tuberculosis. *PLoS ONE* **6**, e16500, doi:10.1371/journal.pone.0016500 (2011).

11 Gold, B., Rodriguez, G. M., Marras, S. A., Pentecost, M. & Smith, I. 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* **42**, 851-865 (2001).

12 Rodriguez, G. M., Voskuil, M. I., Gold, B., Schoolnik, G. K. & Smith, I. ideR, An essential gene in mycobacterium tuberculosis: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infect Immun* **70**, 3371-3381 (2002).

13 Schroder, J. & Tauch, A. 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 microbiology reviews* **34**, 685-737, doi:10.1111/j.1574-6976.2010.00228.x (2010).

14 Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31**, 64-68 (2002).

15 Guelzim, N., Bottani, S., Bourgine, P. & Kepes, F. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* **31**, 60-63 (2002).

16 Lee, T. I. *et al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804, doi:10.1126/science.1075090 (2002).

17 Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. (Chapman and Hall/CRC, 2006).

18 Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A* **100**, 11980-11985 (2003).

19 Mangan, S., Zaslaver, A. & Alon, U. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J Mol Biol* **334**, 197-204 (2003).

20 Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824-827 (2002).

21 Kashtan, N., Itzkovitz, S., Milo, R. & Alon, U. Topological generalizations of network motifs. *Phys Rev E Stat Nonlin Soft Matter Phys* **70**, 031909 (2004).

22 Balazsi, G., Heath, A. P., Shi, L. & Gennaro, M. L. The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest. *Mol Syst Biol* **4** (2008).

23 Park, H. D. *et al.* Rv3133c/dosR is a transcription factor that mediates the hypoxic response of Mycobacterium tuberculosis. *Mol Microbiol* **48**, 833-843 (2003).

24 Sherman, D. R. *et al.* Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -crystallin. *Proc Natl Acad Sci U S A* **98**, 7534-7539, doi:10.1073/pnas.121172498 (2001).

25 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* **107**, 5154-5159 (2010).

26 Colangeli, R. *et al.* The multifunctional histone-like protein Lsr2 protects mycobacteria against reactive oxygen intermediates. *Proc Natl Acad Sci U S A* **106**, 4414-4418 (2009).

27 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* **3**, e87 (2007).

28 Travers, A. & Muskhelishvili, G. Bacterial chromatin. *Curr Opin Genet Dev* **15**, 507-514, doi:10.1016/j.gde.2005.08.006 (2005).

29 Rustad, T. R., Harrell, M. I., Liao, R. & Sherman, D. R. The enduring hypoxic response of Mycobacterium tuberculosis. *PLoS ONE* **3**, e1502, doi:10.1371/journal.pone.0001502 (2008).

30 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* **65**, 684-699, doi:10.1111/j.1365-2958.2007.05827.x (2007).

31 Nesbitt, N. M. *et al.* A thiolase of Mycobacterium tuberculosis is required for virulence and production of androstenedione and androstadienedione from cholesterol. *Infect Immun* **78**, 275-282, doi:10.1128/IAI.00893-09 (2010).

32 Gao, C. H., Yang, M. & He, Z. G. Characterization of a Novel ArsR-Like Regulator Encoded by Rv2034 in Mycobacterium tuberculosis. *PLoS ONE* **7**, e36255, doi:10.1371/journal.pone.0036255 (2012).

33 Gonzalo-Asensio, J. *et al.* PhoP: a missing piece in the intricate puzzle of Mycobacterium tuberculosis virulence. *PLoS ONE* **3**, e3496, doi:10.1371/journal.pone.0003496 (2008).

34 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* **281**, 1313-1316, doi:10.1074/jbc.C500388200 (2006).

35 Ryndak, M., Wang, S. & Smith, I. PhoP, a key player in Mycobacterium tuberculosis virulence. *Trends Microbiol* **16**, 528-534, doi:10.1016/j.tim.2008.08.006 (2008).

36 Abramovitch, R. B., Rohde, K. H., Hsu, F. F. & Russell, D. G. aprABC: a Mycobacterium tuberculosis complex-specific locus that modulates pH-driven adaptation to the macrophage phagosome. *Mol Microbiol* **80**, 678-694, doi:10.1111/j.1365-2958.2011.07601.x (2011).

37 Singh, A. *et al.* Mycobacterium tuberculosis WhiB3 maintains redox homeostasis by regulating virulence lipid anabolism to modulate macrophage response. *PLoS Pathog* **5**, e1000545, doi:10.1371/journal.ppat.1000545 (2009).

38 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* **104**, 11562-11567, doi:10.1073/pnas.0700490104 (2007).

39 Daniel, J., Maamar, H., Deb, C., Sirakova, T. D. & Kolattukudy, P. E. Mycobacterium tuberculosis uses host triacylglycerol to accumulate lipid droplets and acquires a dormancy-like phenotype in lipid-loaded macrophages. *PLoS pathogens* **7**, e1002093, doi:10.1371/journal.ppat.1002093 (2011).

40 Deb, C. *et al.* A novel lipase belonging to the hormone-sensitive lipase family induced under starvation to utilize stored triacylglycerol in Mycobacterium tuberculosis. *J Biol Chem* **281**, 3866-3875, doi:10.1074/jbc.M505556200 (2006).

41 Low, K. L. *et al.* Triacylglycerol utilization is required for regrowth of in vitro hypoxic nonreplicating Mycobacterium bovis bacillus Calmette-Guerin. *J Bacteriol* **191**, 5037-5043, doi:10.1128/JB.00530-09 (2009).

42 Daniel, J. *et al.* Induction of a novel class of diacylglycerol acyltransferases and triacylglycerol accumulation in Mycobacterium tuberculosis as it goes into a dormancy-like state in culture. *J Bacteriol* **186**, 5017-5030, doi:10.1128/JB.186.15.5017-5030.2004 (2004).

43 Garton, N. J., Christensen, H., Minnikin, D. E., Adegbola, R. A. & Barer, M. R. Intracellular lipophilic inclusions of mycobacteria in vitro and in sputum. *Microbiology* **148**, 2951-2958 (2002).

44 Garton, N. J. *et al.* Cytological and transcript analyses reveal fat and lazy persister-like bacilli in tuberculous sputum. *PLoS medicine* **5**, e75, doi:10.1371/journal.pmed.0050075 (2008).

45 Baek, S. H., Li, A. H. & Sassetti, C. M. Metabolic regulation of mycobacterial growth and antibiotic sensitivity. *PLoS biology* **9**, e1001065, doi:10.1371/journal.pbio.1001065 (2011).

46 Sirakova, T. D. *et al.* Identification of a diacylglycerol acyltransferase gene involved in accumulation of triacylglycerol in Mycobacterium tuberculosis under stress. *Microbiology* **152**, 2717-2725, doi:10.1099/mic.0.28993-0 (2006).

47 Sartain, M. J., Dick, D. L., Rithner, C. D., Crick, D. C. & Belisle, J. T. Lipidomic analyses of Mycobacterium tuberculosis based on accurate mass measurements and the novel "Mtb LipidDB". *Journal of lipid research* **52**, 861-872, doi:10.1194/jlr.M010363 (2011).

48 Sherrid, A. M., Rustad, T. R., Cangelosi, G. A. & Sherman, D. R. Characterization of a Clp protease gene regulator and the reaeration response in Mycobacterium tuberculosis. *PLoS ONE* **5**, e11622, doi:10.1371/journal.pone.0011622 (2010).

49 Wayne, L. G. & Sohaskey, C. D. Nonreplicating persistence of mycobacterium tuberculosis. *Annu Rev Microbiol* **55**, 139-163, doi:10.1146/annurev.micro.55.1.139 (2001).

50 Arnvig, K. & Young, D. Non-coding RNA and its potential role in Mycobacterium tuberculosis pathogenesis. *RNA biology* **9**, 427-436, doi:10.4161/rna.20105 (2012).

51 Dunn, T. M., Hahn, S., Ogden, S. & Schleif, R. F. An operator at -280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. *Proc Natl Acad Sci U S A* **81**, 5017-5020 (1984).

52 Wedel, A., Weiss, D. S., Popham, D., Droge, P. & Kustu, S. A bacterial enhancer functions to tether a transcriptional activator near a promoter. *Science* **248**, 486-490 (1990).

53 Czaplewski, L. G., North, A. K., Smith, M. C., Baumberg, S. & Stockley, P. G. Purification and initial characterization of AhrC: the regulator of arginine metabolism genes in Bacillus subtilis. *Mol Microbiol* **6**, 267-275 (1992).

54 Dandanell, G., Valentin-Hansen, P., Larsen, J. E. & Hammer, K. Long-range cooperativity between gene regulatory sequences in a prokaryote. *Nature* **325**, 823-826, doi:10.1038/325823a0 (1987).

55 Belitsky, B. R. & Sonenshein, A. L. An enhancer element located downstream of the major glutamate dehydrogenase gene of Bacillus subtilis. *Proc Natl Acad Sci U S A* **96**, 10290-10295 (1999).

56 Oehler, S., Eismann, E. R., Kramer, H. & Muller-Hill, B. The three operators of the lac operon cooperate in repression. *The EMBO journal* **9**, 973-979 (1990).

57 Narang, A. Effect of DNA looping on the induction kinetics of the lac operon. *J Theor Biol* **247**, 695-712, doi:10.1016/j.jtbi.2007.03.030 (2007).

58 Flashner, Y. & Gralla, J. D. Dual mechanism of repression at a distance in the lac operon. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 8968-8972 (1988).

59 Ninfa, A. J., Reitzer, L. J. & Magasanik, B. Initiation of transcription at the bacterial glnAp2 promoter by purified E. coli components is facilitated by enhancers. *Cell* **50**, 1039-1046 (1987).

60 Reitzer, L. J. & Magasanik, B. Transcription of glnA in E. coli is stimulated by activator bound to sites far from the promoter. *Cell* **45**, 785-792 (1986).

61 Ueno-Nishio, S., Mango, S., Reitzer, L. J. & Magasanik, B. Identification and regulation of the glnL operator-promoter of the complex glnALG operon of Escherichia coli. *J Bacteriol* **160**, 379-384 (1984).

62 Ueno-Nishio, S., Backman, K. C. & Magasanik, B. Regulation at the glnL-operator-promoter of the complex glnALG operon of Escherichia coli. *J Bacteriol* **153**, 1247-1251 (1983).

63 Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Meth* **4**, 651-657 (2007).

64 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).

65 Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502, doi:10.1126/science.1141319 (2007).

66 Ehrt, S. *et al.* Controlling gene expression in mycobacteria with anhydrotetracycline and Tet repressor. *Nucleic Acids Res* **33**, e21, doi:10.1093/nar/gni013 (2005).

67 Ehrt, S. & Schnappinger, D. Controlling gene expression in mycobacteria. *Future microbiology* **1**, 177-184, doi:10.2217/17460913.1.2.177 (2006).

68 Klotzsche, M., Ehrt, S. & Schnappinger, D. Improved tetracycline repressors for gene silencing in mycobacteria. *Nucleic Acids Res* **37**, 1778-1788, doi:10.1093/nar/gkp015 (2009).

69 Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049-1061, doi:10.1016/j.cell.2008.02.039 (2008).

70 Mazzoni, E. O. *et al.* Embryonic stem cell-based mapping of developmental transcriptional programs. *Nat Methods* **8**, 1056-1058, doi:10.1038/nmeth.1775 (2011).

71 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858, doi:10.1101/gr.078212.108 (2008).

72 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

73 Lun, D. S., Sherrid, A., Weiner, B., Sherman, D. R. & Galagan, J. E. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biology* **10**, R142, doi:10.1186/gb-2009-10-12-r142 (2009).

74 Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research* **34**, W369-373, doi:10.1093/nar/gkl198 (2006).

75 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064.

76 Gao, C. H., Yang, M. & He, Z. G. An ArsR-like transcriptional factor recognizes a conserved sequence motif and positively regulates the expression of phoP in mycobacteria. *Biochemical and biophysical research communications* **411**, 726-731, doi:10.1016/j.bbrc.2011.07.014 (2011).

77 Arnvig, K. B. *et al.* Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of Mycobacterium tuberculosis. *PLoS Pathog* **7**, e1002342, doi:10.1371/journal.ppat.1002342 (2011).

78 Minch, K., Rustad, T. & Sherman, D. R. Mycobacterium tuberculosis Growth following Aerobic Expression of the DosR Regulon. *PLoS One* **7**, e35935, doi:10.1371/journal.pone.0035935 (2012).

79 Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

80 Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264, doi:10.1093/biostatistics/4.2.249 (2003).

81 Aagaard, C. *et al.* A multistage tuberculosis vaccine that confers efficient protection before and after exposure. *Nat Med* **17**, 189-194, doi:10.1038/nm.2285 (2011).

82 Dolganov, G. M. *et al.* A novel method of gene transcript profiling in airway biopsy homogenates reveals increased expression of a Na+-K+-Cl- cotransporter (NKCC1) in asthmatic subjects. *Genome Res* **11**, 1473-1483, doi:10.1101/gr.191301 (2001).

83 Weisberg, S. *Applied linear regression*. 2nd edn, (Wiley, 1985).

84 Miller, A. J. *Subset selection in regression*. (Chapman and Hall, 1990).

85 Storey, J. D. A direct approach to false discovery rates. *J Roy Stat Soc B* **64**, 479-498 (2002).

86 Storey, J. D. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Stat* **31**, 2013-2035 (2003).

87 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445, doi:10.1073/pnas.1530509100 (2003).

88 Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716 (1974).

89 Lilliefors, H. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* **62**, 339 (1967).

90 Gardner, T. S., di Bernardo, D., Lorenz, D. & Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102-105 (2003).

91 Yang, X., Shen, Q., Xu, H. & Shoptaw, S. Functional regression analysis using an F test for longitudinal data with large numbers of repeated measures. *Statistics in medicine* **26**, 1552-1566, doi:10.1002/sim.2609 (2007).

92 Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15 (2003).

93 Sackrowitz, H. & Samuel-Cahn, E. P values as random variables - Expected P values. *Am Stat* **53**, 326-331 (1999).

94 Pfluger, R. & Hothorn, T. Assessing equivalence tests with respect to their expected p-value. *Biometrical J* **44**, 1015-1027 (2002).

95 Marbach, D. *et al.* Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. *Genome research*, doi:10.1101/gr.127191.111 (2012).

96 mod, E. C. *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787-1797, doi:10.1126/science.1198374 (2010).

97 Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* **96**, 1151-1160 (2001).

98 Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *P Natl Acad Sci USA* **98**, 5116-5121 (2001).

99 Xu, X. L., Olson, J. M. & Zhao, L. P. A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. *Hum Mol Genet* **11**, 1977-1985 (2002).

100 Ernst, J., Vainas, O., Harbison, C. T., Simon, I. & Bar-Joseph, Z. Reconstructing dynamic regulatory maps. *Mol Syst Biol* **3**, 74, doi:10.1038/msb4100115 (2007).

101 Reitman, Z. J. *et al.* Profiling the effects of isocitrate dehydrogenase 1 and 2 mutations on the cellular metabolome. *Proc Natl Acad Sci U S A* **108**, 3270-3275, doi:10.1073/pnas.1019393108 (2011).

102 Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M. & Milgram, E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Analytical chemistry* **81**, 6656-6667, doi:10.1021/ac901536h (2009).

103 DeHaven, C. D., evans, A. M., Dai, H. & Lawton, K. A. Organization of GC/MS and LC/MS metabolomics data into chemical libraries. *Journal of Cheminformatics* **2** (2010).

104 Chauhan, S. & Tyagi, J. S. Cooperative binding of phosphorylated DevR to upstream sites is necessary and sufficient for activation of the Rv3134c-devRS operon in Mycobacterium tuberculosis: implication in the induction of DevR target genes. *J Bacteriol* **190**, 4301-4312, doi:10.1128/JB.01308-07 (2008).