Reconstructing the Regulatory Network of TB: Transcription Factor Binding Distribution and Properties

Authors: Anna Lyubetskaya, Matthew Peterson, James Galagan.

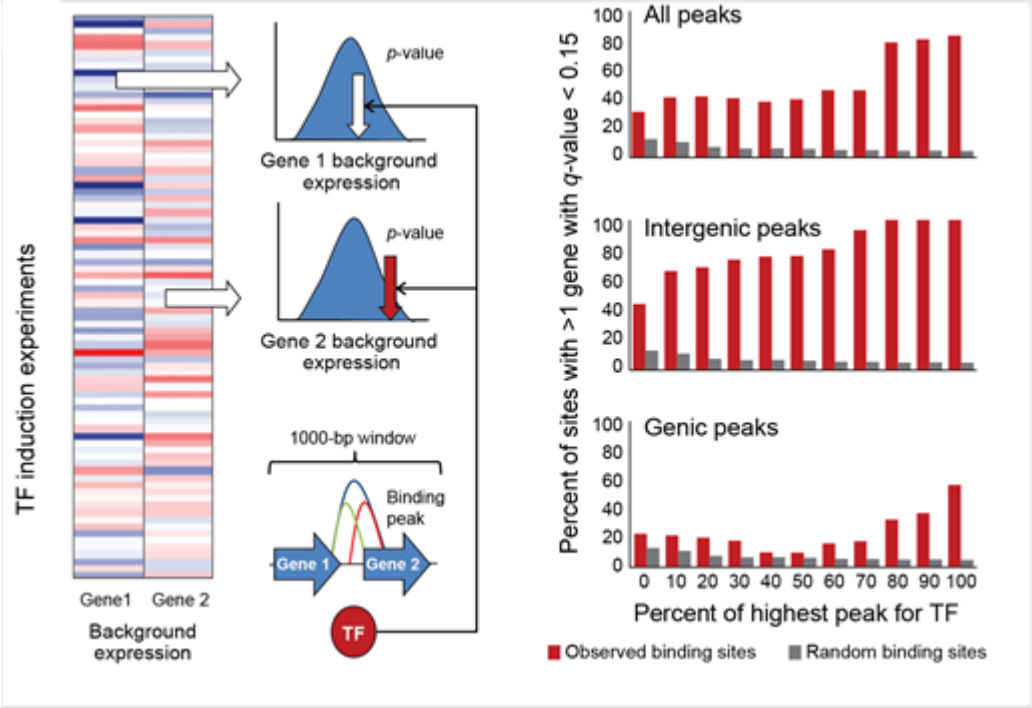
Bioinformatics Program, Boston University, Boston, USA.

E-mail: avl@bu.edu

As a part of the project to identify, validate, and perturb key genes and network interactions predicted to underlie the metabolic adaptations of *M. tuberculosis* and reprogramming of host cells during TB infection, we built a transcriptional network of MTB. So far, we successfully carried out ChIP-Seq experiments for 90 TFs and developed a pipeline to computationally analyze the data. We verified the quality of our data by comparing multiple biological replicates for at least 15 TFs. For 10 key TFs, we also confirmed that our experiments, although carried out in conditions containing oxygen, apply to hypoxic coditions characteristic of MTB infection. While we detected all previously known binding sites for a few well-studied regulators in MTB (KstR and DosR), we also found many more binding instances and significantly extended transcription factor regulons.

In order to validate transcriptional function of predicted binding sites, we carried out complimentary overexpression experiments for all 200 TFs of MTB (replicate experiments were performed for a number of TFs). This data was used to assign a probability of observing the expression level for each gene identified to be bound by a given transcription factor in the overexpression microarrays (Figure 1). For each site, we examined all genes around the site to determine if the overexpression of the corresponding TF significantly altered expression of these genes. Binding sites were validated if any gene in the window displayed an expression level greater than a threshold value after correction for multiple testing.

Applying this method to all sites from analyzed TFs, we could assign a potential regulatory role to 25% of binding sites. Stronger binding sites were more often associated with regulation than weaker sites, suggesting a possible correlation between binding strength and regulatory impact. However, it appeared that clusters of weak binding sites had a stronger regulatory role than weak singletons suggesting cooperative mechanism of interaction. Also, strong binding sites were often located in the proximity of weak sites which suggested the role of weak sites in modulating affinity.



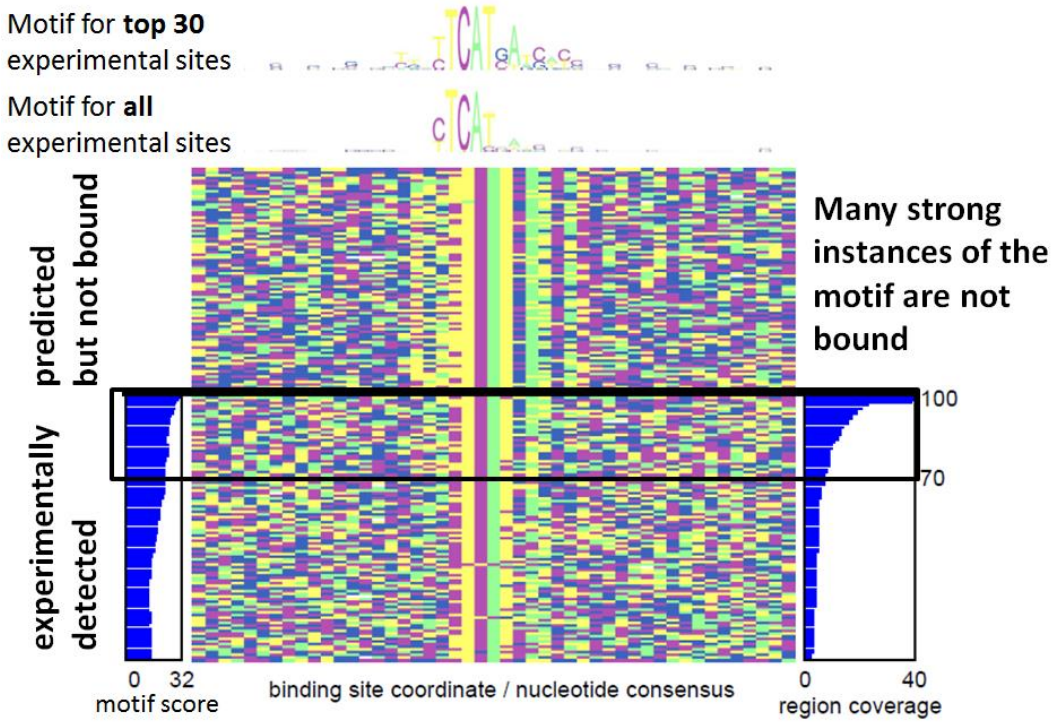
The canonical model of the transcriptional regulation in prokaryotes restricted binding site location to proximal promoter region and suggested that the binding sequence is the main determinant of the binding. The distance between binding sites and associated target genes displayed a pattern partially consistent with expectation: about half of binding sites were located within 1000bp of the start codon of the gene they were predicted to regulate. Most binding sites located in upstream intergenic regions were validated by expression data. However, 76% of binding sites fell into annotated coding regions and a significant proportion could be assigned regulation.

Integration of independent binding and expression datasets allowed us to test which binding site characteristics – binding motif strength, ChIP-Seq coverage, relative location of site and potential target genes, and presence of other binding sites – are essential for assigning regulatory role.

Although a conservative binding motif was found for most transcription factors, only a fraction of motif instances appeared bound in the experiment. The experimentally determined motif for weak binding sites was often a degraded version of the motif detected for the strong binding sites. Some low-affinity binding sites appeared occupied by the transcription factor while many high-affinity binding sites were not.

For example, we detected 100 binding sites for Rv0602c (Figure 2). Site coverage ranges from 40 times above the median to 1 as indicated in the right side of the figure. The strong experimentally detected binding sites are characterized by a TCATGA motif. With coverage, this motif degrades to the TCAT core with some conservation at accesory positions as reflected by the motif score in the left side of the figure. However, if we use

the strong motif to scan the genome, we find many additional instances unbound in the experiment. Interestingly, we find exactly exactly same 13 nulceotides bound in one area of the genome and not bound in another.



By comparing experimental and computational binding site distributions, we defined 'hot' areas of the genome (that were depleted of binding in the experiment despite the existence of motif instances) and 'cold' areas (that were bound by more TFs than expected from the regression model). A nucleoid-associated protein LSR2 with a known role in organizing DNA was associated with these regions, as well as other TFs with no known structural function (for example, Rv0081). Our data suggested that some transcription factors had both distinct regulatory role and significant impact on DNA organization.