Transcription Factor Binding Site Mapping Using ChIP-Seq

SUMA JAINI,1,4 ANNA LYUBETSKAYA,2 ANTONIO GOMES,2 MATTHEW PETERSON,1 SANG TAE PARK,3 SAHADEVAN RAMAN,4 GARY SCHOOLNIK,5 and JAMES GALAGAN1,2,4,6

1Department of Biomedical Engineering, Boston University, Boston, MA 02215; 2Bioinformatics Program, Boston University, Boston, MA 02215; 3Macrogen Clinical Laboratory, Macrogen Corp, Rockville, MD 20850; 4National Emerging Infectious Diseases Laboratory, Boston University, Boston, MA 02118; 5Department of Medicine and Department of Microbiology and Immunology, Stanford Medical School, Stanford, CA 94305; 6Broad Institute of MIT and Harvard, Cambridge, MA 02142.

ABSTRACT Transcription factors (TFs) play a central role in regulating gene expression in all bacteria. Yet until recently, studies of TF binding were limited to a small number of factors at a few genomic locations. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) provides the ability to map binding sites globally for TFs, and the scalability of the technology enables the ability to map binding sites for every DNA binding protein in a prokaryotic organism. We have developed a protocol for ChIP-Seq tailored for use with mycobacteria and an analysis pipeline for processing the resulting data. The protocol and pipeline have been used to map over 100 TFs from Mycobacterium tuberculosis, as well as numerous TFs from related mycobacteria and other bacteria. The resulting data provide evidence that the long-accepted spatial relationship between TF binding site, promoter motif, and the corresponding regulated gene may be too simple a paradigm, failing to adequately capture the variety of TF binding sites found in prokaryotes. In this article we describe the protocol and analysis pipeline, the validation of these methods, and the results of applying these methods to M. tuberculosis.

This chapter describes a chromatin immunoprecipitation followed by sequencing (ChIP-Seq) method tailored for the study of Mycobacterium tuberculosis transcription factors (TFs) but amenable for the study of other prokaryotes. Noteworthy features of this method include the following: (i) it is conducted using a standard and readily reproducible growth condition for each TF studied; (ii) the binding behavior of each TF is studied using a tagged variant of the TF whose production is driven by an inducible promoter; (iii) each TF is studied using the same concentration of an exogenously added chemical inducer where the strength of TF gene induction is varied to systematically measure the effect of TF concentration on binding site strength and location; (iv) the resulting binding site data is spatially well resolved and highly reproducible, and binding strength is correlated with the degree of binding site motif conservation; and (v), because this method does not require knowledge of the physiological conditions that normally cause TF gene expression or of antibodies specific to each TF, it is applicable for the high-throughput study of multiple TFs; thus far it has generated binding site data for over 119 annotated M. tuberculosis TFs. Despite the use of a standard growth condition for all TFs and an inducible promoter system, the binding site data was found to agree well with data from a subset of TFs expressed under their normal physiological conditions.

Received: 9 September 2013, Accepted: 18 October 2013, Published: 18 April 2014

Editors: Graham F. Hatfull, University of Pittsburgh, Pittsburgh, PA, and William R. Jacobs, Jr., Howard Hughes Medical Institute, Albert Einstein College of Medicine, Bronx, NY


Correspondence: J. Galagan, jgalag@broadinstitute.org

© 2014 American Society for Microbiology. All rights reserved.
ChIP-Seq is a well-established method for identifying binding sites for DNA binding proteins (2–4). Chromatin immunoprecipitation (ChIP) is the first step (Fig. 1). Proteins bound to the genomes of these cells are cross-linked to DNA with formaldehyde. The cells are then broken open, and the DNA is sheared through sonication or enzymatic digestion. DNA fragments bound by a protein of interest (a TF, for example) are immunoprecipitated using an antibody to the protein. Cross-linking is then reversed to remove the proteins, and the precipitated DNA fragments are isolated and sequenced to generate reads from the ends of the fragments. Sequenced reads are aligned to the corresponding genome sequence, and genomic locations from which the DNA fragments are derived are identified as regions that are overrepresented with aligned reads.

Ideally, only genomic regions that are bound by the protein of interest would display read coverage. In practice, nonspecific DNA fragments will also be isolated and sequenced, resulting in a background coverage on genome regions that are not bound. To assess this background coverage, one or more control experiments are typically used. One common example is the use of mock ChIP runs; they include every step of the ChIP process with the exception of the antibody purification. These experiments control for the many nonantibody steps of ChIP that may lead to nonspecific DNA isolation. Antibodies can also be used against epitope tagged proteins. In those cases the control experiment can use the same antibodies for ChIP in strains that lack the epitope tag. Such experiments assess the degree to which the antibody recognizes nonspecific targets. A genomic DNA preparation can also be generated as a control. Control experiments help to identify the differential efficiency of isolation and sequencing of different locations of the genome. The genome regions that contain binding sites for the protein of interest are characterized by greater coverage in comparison with the background coverage.

ChIP-Seq also produces a strand-specific signature of enrichment that can be used to identify true binding peaks. When DNA fragments from ChIP are sequenced, reads are typically generated from one end of the fragment or the other. Protein binding sites will occur between the ends of this fragment. Reads will thus align to one side or the other of the binding site and thus on different strands. If coverage is visualized for the two strands separately, this process gives rise to a bimodal enrichment profile: the coverage of the forward strand will be shifted upstream with respect to the actual binding site, while the coverage on the reverse strand will be shifted downstream (5). The distance between the forward and reverse coverage profile is determined by the size of the DNA fragment sequence, which can be estimated during the sequencing process. Nonspecific binding, by contrast, often results in enrichment that lacks this bimodal shift.

DNA binding proteins, especially TFs, typically bind to short DNA sequences (on the order of 15 bp or less). Enriched peaks, however, typically span a region of several hundred base pairs as a consequence of the larger fragment size generated during ChIP (typically around 250 bp). Moreover, when multiple closely spaced binding sites for a protein exist in a particular location, the read coverage for these sites can merge into a single broad enriched region.

We have developed a protocol for ChIP-Seq that has been tailored for use with mycobacteria, although we have also successfully utilized the same protocol for other bacteria. The protocol is provided in the appendix. A number of factors contribute to obtaining optimal results with this protocol. The most important factor is the quality of the antibody. Strong binding to the protein of interest and minimal nonspecific binding are critical. During IP, sufficient time should be allowed for the antibodies to interact with their target proteins. Also, incubation with antibodies should be performed at 4°C in order to minimize any protein degradation. During capturing of immuno-complexes, washes should be performed with care to minimize the loss of agarose beads. The nature of protein-G coated beads to capture
the antibody complex may also influence the quality of the preparation. In our experience, protein-G coated agarose beads are found to be superior to protein-G coated magnetic beads. Agarose beads are fragile, however, so care must be taken to maintain the integrity of the beads during centrifugation.

Another critical factor is to have proper lysing and shearing conditions. Cell lysis can be performed separately from DNA shearing or in a single step using Covaris. The time and strength of shearing should be optimized such that most of the sheared DNA is between 200 and 500 bp. Depending on the cell lysis method used, it is important to make sure the cells are lysed efficiently to release the cellular contents. Avoid contamination of the lysate with cellular debris during centrifugation and separation of the lysate.

For library preparation, the last amplification step is important. To prevent amplification of nonspecific DNA, do not exceed 18 cycles of amplification. Also, there should not be a significant amount of primer dimers or adapter dimers. In order to prevent the carryover of primer/adapter dimers, adjust the number of adapters and primers accordingly. The standard Illumina library prep protocol is provided for 10 ng of ChIP DNA. If there is insufficient DNA, ChIP should be performed using more cells.

**DATA ANALYSIS**

We have developed a computational pipeline to perform data processing and analysis of ChIP-Seq data for mycobacteria and other microbes (Fig. 2). The pipeline consists of two stages: (i) detection of enriched regions from raw ChIP-Seq data and (ii) identification of binding sites and motifs within enriched regions.
Detection of Enriched Regions from Raw ChIP-Seq Sequencing Data

Sequence reads are mapped to the target genome sequence using MAQ (6). The “pileup” command from the SAMtools (7) suite is used to calculate coverage along the forward and reverse strands along the genome. From this coverage, regions of enrichment along the genome are identified using a log normal distribution. The log normal distribution is defined by the probability density function (PDF):

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

Here, $\mu$ and $\sigma$ are the mean and standard deviation of the natural logarithm of the random variable, respectively. For each experiment, positions along the genome greater than five times the mean coverage or higher are excluded to avoid fitting outliers, and the parameters of the distribution are estimated using maximum likelihood. Although the log normal distribution has been the best fit for most microbial ChIP-Seq data in our experience, we have also tested other distributions including the Gaussian, negative binomial, and generalized extreme value distributions.

The resulting distribution is used to score each position of the genome to identify enriched positions along the genome. Positions with a $P$ value of 0.01 or lower are called enriched. Since TF binding is expected to result in contiguous regions of enrichment, only regions of enrichment of a threshold length or longer are included for further analysis. This threshold is set by the user and is typically 100 bp for our analyses.

A cross-correlation filter is 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 between 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 is defined as the value $n$ that maximizes the cross-correlation. Regions with a shift of less than a particular threshold are removed from further analysis.

Regions are then compared against a database of known artifacts that have been identified in the organisms we have worked on. In some cases, these artifacts are general and include annotated repeat regions. In other cases, artifacts can be organism specific. For example, in M. tuberculosis we observed that certain regions showed statistical enrichment in nearly all ChIP-Seq experiments regardless of the ChIP’d TF. It was found that most of these regions displayed strong
binding by the histone-like protein Lsr2 (8) and may reflect nonspecific interactions. Our pipeline flags these artifacts and removes them for separate consideration. The output at this stage of the analysis pipeline is a list of enriched regions and their locations. However, as noted above, enriched regions can typically span hundreds of nucleotides, while binding sites are more often less than 15 bases long. Moreover, a single enriched region can contain multiple individual binding sites.

**Identification of Binding Sites and Motifs within Enriched Regions**

To resolve individual binding sites within an enriched region, we treat binding site prediction as a signal detection problem. Conceptually, binding sites can be considered a source of an input signal that, through the process of ChIP-Seq, give rise to the broader output signal observed in the ChIP-Seq coverage. Multiple binding sites give rise, to a first approximation, to an output coverage that is the sum of the outputs of each of the individual binding sites. This process can be modeled as a linear convolution.

In this model, the output signal arising from a point input is called an impulse response (or point spread function). In the case of the ChIP-Seq, the impulse response is a consequence of “transmitting” the input signal through the process of randomly sequencing the ends of large DNA fragments that overlap the point source and “receiving” this transmission in terms of coverage after aligning these reads. The impulse function essentially “blurs” the output signal arising from a point input. Input signals are modeled as the sum of multiple point sources, or impulse functions, that are each scaled to a particular magnitude. In the case of ChIP-Seq, impulse functions correspond to binding sites where the scaling associated with each site corresponds roughly to relative enrichment (and thus relative occupancy). The output signal is then the sum of the correspondingly scaled impulse functions, or a convolution in mathematical terms (specifically, a discrete convolution since DNA coordinates are integer based).

The operation of recovering the binding site locations from regions of enriched coverage is then a deconvolution. In the case where the impulse function is known, this process is straightforward (9). For ChIP-Seq, however, this function is not known and depends in part on the details of each specific experiment. Thus, the impulse function must be estimated at the same time that the coverage signal is being deconvolved. This operation is termed a blind deconvolution (blind because the impulse function is not known *a priori*), and techniques have also been developed to solve this problem (10).

We previously developed a method, CSDeconv, that uses a reestimation method to solve the blind deconvolution problem (11) (Fig. 3). The basic approach begins by generating an initial estimate of the impulse function. This is typically generated by selecting a set of peaks with high coverage that are then used to fit a model of the impulse function based on an initial estimate of the binding site locations in these regions. This estimated impulse function can then be used to perform blind deconvolution on all enriched regions, which results in new binding site locations for all enriched regions. These new binding site locations are then used to fit an updated model of the impulse function, and the process iterates in this fashion until convergence criteria are reached.

We more recently enhanced the basic CSDeconv algorithm in a number of ways to develop a new algorithm called BRACIL (Gomes A, Abeel T, Peterson M, Azizi E, Lyubetskaya A, Carvalho L, Galagan J, submitted for publication). First, we incorporated *de novo* motif discovery into the peak detection process (Fig. 3). By performing motif discovery during the blind deconvolution process, we can improve the prediction of binding site locations. By using binding sites locations predicted by the blind deconvolution algorithm as input, we are able to improve motif discovery. Second, we developed a more principled model for the impulse function; this and other enhancements have improved the accuracy and runtime of the algorithm such that we are able to rapidly analyze binding sites for even large mammalian chromosomes.

We have also incorporated the ability to detect potential cooperative binding. Closely spaced binding sites have been shown to mediate cooperative binding that can substantially alter the apparent affinity of individual sites (13). The deconvolution approach described above implicitly assumes that TFs bind individual sites independently of all other sites. The approach can be generalized, however, to explicitly model dependencies between sites for the same TF. Using this approach, it is possible to predict known cooperative interactions between closely spaced sites for the DosR transcription factor in *M. tuberculosis* (12).

The final output of our blind deconvolution method is a list of binding sites with high spatial resolution and accuracy (11, 12). Based on a test using ChIP-Seq data for the GABP TF in humans and the DosR TF in *M. tuberculosis*, the approach is able to identify binding locations to within an average absolute difference of less
than 24 bp (11). Moreover, the method can accurately predict multiply spaced binding sites within the same ChIP-Seq enriched region. In the case of DosR, binding sites located less than 57 bp apart in the same intergenic region could be resolved, while several closely spaced binding sites for GABP were observed, two as close as 20 bp apart.

**COMPREHENSIVE MAPPING OF TF BINDING SITES IN M. TUBERCULOSIS**

As part of a consortium effort (1), we developed a high-throughput system for comprehensively mapping and functionally validating regulatory interactions in *M. tuberculosis*. For this effort, we performed ChIP-Seq (2, 4, 14) using FLAG-tagged TFs episomally expressed under the control of the well-known mycobacterial tetracycline-inducible promoter (15–18). An anhydro-tetracycline (Atc)-inducible episomal vector containing a Gateway Recombination™ (Invitrogen) cassette was modified to contain an in-frame N- or C-terminal FLAG epitope tag. TFs from a Gateway entry clone library (supplied by the NIAID-funded PFGRC) were recombined into this vector to create N- or C-terminally epitope-tagged expression vectors. The full complement of these epitope-tagged expression vectors is being made available to all researchers through the NIAID-funded BEI repository.

This system allowed us to map DNA binding of the vast majority of the ∼180 *M. tuberculosis* regulators in a consistent and comparable manner independent of regulatory function. The use of tagged TFs obviates the need to develop antibodies to native proteins, an effort that would have been prohibitive for all *M. tuberculosis*

![FIGURE 3 Blind deconvolution analysis of ChIP-Seq data. This schematic representation illustrates the steps of BRACIL (12) in the analysis of ChIP-Seq data. BRACIL uses a blind deconvolution algorithm that takes advantage of ChIP-Seq coverage and genome sequence to refine the resolution of ChIP-Seq binding regions into single-nucleotide resolution. The top panel illustrates the steps of the blind-deconvolution algorithm. The algorithm starts with a guess about the shape of the impulse response as well as the location of binding sites. Both the shape of the impulse response and the binding site locations are updated iteratively until convergence. The predicted binding site locations are used in motif discovery to predict a binding motif. This motif is used to constrain the search space for deconvolution and refine prediction of binding site locations. A set of high-resolution binding site locations is obtained as a final prediction. doi:10.1128/MicrobiolSpectrum.MGM2-0035-2013.f3](ASMscience.org/MicrobiolSpectrum)
TFs. Moreover, we are able to use a ChIP grade antibody to FLAG for IP that enables high sensitivity and specificity of pull down. Control IP experiments in strains lacking FLAG-tagged proteins revealed little nonspecific binding of the antibody used. What nonspecific binding occurred could be systematically removed for all tagged TFs in our analysis pipeline. The antibody used also displayed a high level of enrichment for tagged proteins during IP, leading to high signal to noise in subsequent sequencing and alignment.

The use of an inducible promoter system ensures expression of targeted TFs, which allowed us to study the full complement of M. tuberculosis regulators in a standard and highly reproducible baseline state without a priori knowledge of the conditions that normally induce their expression. All regulators studied this way were induced with ATc during mid-log-phase growth. We confirmed that protein expression per se did not lead to spurious nonspecific binding. ChIP-Seq of three non–DNA binding proteins using the inducible system resulted in no enriched binding (1). The tetracycline-inducible promoter is a relatively weaker promoter that induces the expression to a moderate level, preventing the overproduction of the protein beyond a certain level. The standard level of induction was 100 ng Atc per ml of culture, although the inducible system provided the ability to induce at multiple levels (see below). The use of 100 ng/ml Atc results in different levels of expression for different TFs relative to their physiological expression levels. For example, in the case of KstR and Rv0081, induction with 100 ng/ml resulted in roughly four times and three times overexpression, respectively, at the mRNA level relative to the expression observed for known physiological stimuli (1). In contrast, for DosR the same level of Atc resulted in expression induction essentially equivalent to that observed under physiological conditions.

The small size of the M. tuberculosis genome (~4.4 Mb) serves as an advantage for ChIP-Seq mapping. ChIP-Seq coverage for binding sites scales with the overall number of reads generated relative to the size of the reference genome. A typical ~36-bp sequencing lane on an Illumina GAIIx is sufficient to provide an average of 500-fold coverage for the M. tuberculosis genome. With this degree of coverage, binding sites for individual TFs can be identified with coverage that spans several logs in magnitude (Fig. 4A). We routinely observed enrichments of over 100,000-fold relative to a baseline of 150-fold, while the highest affinity peaks for certain TFs could result in binding regions with over 1 million-fold coverage. The differences in coverage between different binding sites reflect the probability of occupancy of each site in the population of cells on which ChIP-Seq was performed. Occupancy, in turn, reflects a number of factors including the concentration and modification state of the TF, the affinity of the binding site for the TF, the accessibility of the binding site, and the availability of molecular cofactors. The high coverage that can be generated for ChIP-Seq in M. tuberculosis provides insights into the variation of these factors on a genome-wide scale with unprecedented resolution (see below).

**VALIDATION OF THE ChIP-Seq SYSTEM**

The comprehensive mapping of binding sites for all TFs in the M. tuberculosis genome provided the opportunity to perform substantial validation of the results of ChIP-Seq as applied to a large number of different TFs in M. tuberculosis (1). We highlight two results of this validation here.

**Binding Is Sensitive and Reproducible**

The inducible and FLAG-tagged ChIP-Seq system displays high sensitivity. For example, we confirmed the sensitivity of the system using high-confidence direct regulation from two well-studied regulators, the activator DosR (Rv3133c) (11, 19–24) and the repressor KstR (Rv3574) (25–28). In both cases we can identify all of the previously reported direct binding sites. The system also correctly identifies the known binding motifs for these factors. Our system also replicates biochemically identified binding sites for Rv2034 reported by Gao and colleagues (29, 30). Using electrophoretic mobility shift assay (EMSA), these authors demonstrated that Rv2034 binds to its own promoter more strongly than to promoters of other genes they tested, binds both upstream of the DosR operon as well as near the DosR gene, binds upstream of the GroEL2 gene, and binds the promoter of PhoP. ChIP-Seq of Rv2034 using the inducible FLAG-tagged system replicates the location and strength of these binding regions (1). ChIP-Seq also recapitulates the core binding motif identified using EMSA. To date, all binding interactions previously published for TFs that we have successfully mapped with our system have been identified in our experiments.

Binding is also highly reproducible. Biological replicates using the inducible FLAG-tagged system show that coverage for enriched sites is highly correlated between experiments (R² > 0.83 for all TFs with replicates; two examples are shown in Fig. 4B). There is also high reproducibility in binding location, with distances between replicate binding sites less than the length
of predicted binding site motifs for the vast majority of sites.

We further confirmed that binding sites were reproducible between different cellular conditions. To comprehensively map all *M. tuberculosis* TFs, a standard normoxic culture condition was selected that was both practical and not specific to any subset of TFs. To test the degree to which the binding sites observed in normoxia could be detected under different physiological conditions, ChIP-Seq was also performed on 11 TFs under hypoxic conditions. Substantial concordance between the binding enrichment in normoxia and hypoxia was seen (Fig. 5). For the large majority of the regions, the evidence of binding in normoxia was also observed in hypoxia. Moreover, all binding sites observed in hypoxia were also detected in normoxia. In addition, the relative coverage of corresponding binding regions was broadly conserved between hypoxia and normoxia conditions.

Finally, we confirmed that the inducible FLAG-tagged system could replicate binding sites found by an independent lab using a different ChIP-Seq system. In a report by Blasco and colleagues, the TF EspR was mapped by ChIP-Seq using antibodies to the wild-type EspR protein and thus with EspR induced from its native promoter (31). To further validate the inducible tagged TF system, we performed ChIP-Seq on EspR using induction with 100 ng/ml Atc and compared the results

![Graph](https://via.placeholder.com/150)

**FIGURE 4** Example ChIP-Seq results from *M. tuberculosis*. (A) The top panel displays the fold read coverage for a single binding region with two known binding sites for the TF KstR. ChIP-Seq coverage visually resolves both binding sites and confirms the experimental observation that the site closest to Rv3571 is a weaker affinity. Total coverage is shown in blue, and the forward and reverse coverage is shown in red and green, respectively. The binding event also displays the expected shift in position between the forward and reverse reads. The bottom panel displays the genome-wide fold coverage for the same experiment. 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 gray lines are multiples of the standard deviation of background coverage. (B) Binding site identification is highly reproducible. Bar plots show the distance between corresponding sites in two replicates for two TFs. The blue line indicates the length of known motifs. Insets show relationship of peak height between corresponding peaks in two replicates (R2 > 0.83 for all TFs). Figures from reference 1. doi:10.1128/microbiolspec.MGM2-0035-2013.f4
Figure 5 continues on next page
FIGURE 5 (continued)
to those from the Blasco report. As shown in Fig. 6, we observed a high level of agreement between the two methods applied to map the same TF in independent groups. In particular, the inducible FLAG-tagged system recapitulates the distribution of binding locations and heights, the motif, and the specific pattern of binding at individual loci that was found by Blasco.

**Binding is Correlated with Bound Sequence**

The physiological nature of the binding identified by ChIP-Seq is also supported by an analysis of the sequences underlying each of the bound sites. For nearly all TFs, a consensus motif can be determined based on the ChIP-Seq data, and an instance of this motif can be identified underlying the vast majority of binding sites. Moreover, as a general trend, variations in motif instances can be associated with differences in apparent binding affinity. An example for KstR is shown in Fig. 7. In the bottom portion of the heat map, each row represents the sequence region for a single binding site, with the nucleotides color coded. The rows are ordered from the binding peaks with the most enrichment at the top down to binding peaks with the lowest enrichment at the bottom. A clear consensus motif can be detected as shown, which corresponds to the published binding motif for KstR (25–28), and the core of this motif is visible in the heat map across all the binding sites. Individual binding sites, however, display differences in their underlying motifs. The strongest binding sites contain sequences that contain not only the core palindrome, but also conserved accessory bases known to play an important role in shaping affinity (32). Less-enriched peaks, by contrast, are often found to lack these accessory bases. Furthermore, in some cases only one-half of the palindrome is present. Binding to such half sites was described in the earliest ChIP-Seq experiments in eukaryotes (4). This pattern suggests that, to a first approximation, differences in enrichment at different binding sites can often be explained by differences in the affinities of the underlying sequence motif.

The difference in affinities of different binding sites is also reflected in experiments in which ChIP-Seq was performed on TFs induced to different levels. ChIP enrichment is a function of the number of cells in which a site is bound (33, 34), which is governed by the affinity of the site and the concentration of the factor. By overexpressing a factor, we expect to increase the occupancy of strong sites up to a saturation limit while also occupying, and thus detecting, weaker affinity sites. This was experimentally confirmed. In the example shown in Fig. 6, when KstR is expressed at a low level, only a small number of low peaks are identified, and a number of gold standard sites are missed. As expression increases, overall peak heights increase while maintaining roughly the same relative heights, and additional lower peaks are revealed. At the highest levels of expression, the increase in enrichment of strong sites begins to level off.

It is important to note that the binding sites detected do not simply reflect the association of induced TFs with all available motifs. For the majority of factors mapped, only a fraction of computational identified instances of motifs in the genome are typically bound based on ChIP-Seq. In the example for KstR in Fig. 7, numerous instances of the strongest KstR motif can be detected throughout the genome that are not bound in any of the ChIP-Seq experiments we have performed. These instances include examples of nearly the identical sequence in two different genomic locations, one of which is bound, while the other is not. Binding of the majority of TFs thus displays genome context specificity: only a fraction of the possible binding sites appear occupied, and this differential occupancy is highly reproducible. This phenomenon is well known in eukaryotic systems (34) but is less well appreciated in bacteria. Examples are known of TFs in prokaryotes that appear to be able to bind to any available motif instance regardless of genomic location (35). However, the data from M. tuberculosis, as well as corresponding data from ChIP-Seq in *Escherichia coli* (unpublished results), indicates that for many bacterial TFs, factors other than the underlying sequence determine whether individual motif instances are bound in vivo. The determinants of this specificity are not fully understood (34).

**FIGURE 5** Binding sites replicate between normoxia and hypoxia. Each panel compares the results of ChIP-Seq under both normoxic (x axis, top traces) and hypoxic (y axis, bottom traces). Strong concordance was seen in both peak heights (scatter plots) and coverage profiles (sequencing traces) in experiments performed under both conditions. While no binding sites were identified in normoxia that were not identified in hypoxia, a small number of sites exhibited greatly increased binding under hypoxic conditions. The three sites showing the largest increases are shown in red on the scatter plots. doi:10.1128/microbiolspec.MGM2-0035-2013.f5
FIGURE 6  Independent replication of EspR binding sites. The ChIP-Seq experiment performed with the native EspR antibody (31) compares well to the ChIP-Seq with the inducible promoter. (A) Binding sites are categorized by their locations relative to target genes. Motifs and binding sites detected by independent protocols are very similar. (B) Coverage tracks between two experiments are in concordance. doi:10.1128/microbiolspec.MGM2-0035-2013.f6
THE DIVERSITY OF TF BINDING IN PROKARYOTES

ChIP-Seq has been applied extensively to map TF binding sites in a range of eukaryotic organisms. The data from these studies have provided a wealth of information about the global binding patterns of TFs that has overturned many previously held assumptions about the nature, diversity, and possible functions of TF binding sites (34). The data from M. tuberculosis represent the first large-scale application of ChIP-Seq to bacteria (1). The resulting data have confirmed several surprises that have also emerged from mapping eukaryotic cells. These surprises call into question some of the simplifying assumptions held about bacterial transcriptional regulation (36). In particular, the data reveal that binding of TFs in M. tuberculosis (i) occurs in many more diverse genomic locations than expected and (ii) involves much more weak binding than was previously appreciated.

NONPROMOTER TF BINDING

The canonical model of bacterial transcriptional initiation focuses on the role of binding in the proximal promoter region. Binding in this region enables TFs to facilitate or directly block the recruitment of the polymerase to the transcription start site. Data from ChIP-Seq mapping of over 119 TFs in M. tuberculosis confirms that binding in upstream intergenic regions is enriched over what would be expected by chance. But surprisingly, binding in this region is the exception. As shown in Fig. 8, binding to intergenic regions represents less than 40% of the binding events for any TF. The majority of binding events occur outside of upstream intergenic regions. This has been confirmed by ChIP-Seq of TFs in both E. coli and Salmonella (data not shown).

A number of explanations for this observation are conceivable. The most straightforward explanation is the presence of errors in the annotation of coding regions. The majority of genes in all prokaryotic...
genomes are computational predictions that are known to be error-prone in predicting start codons. In some cases, binding sites that occur at the 5′ ends of annotated coding regions may thus reflect intergenic binding relative to an actual downstream start codon. Another explanation is the well-known fact that promoter regions are not strictly limited to intergenic regions. Although canonical promoter signals are enriched in intergenic regions and may be selected against in genic regions \( (37, 38) \), many examples of promoter regions occurring in coding regions have been described in prokaryotes \( (39–43) \).

Binding outside of promoter regions may also reflect long-range interactions between binding sites and the proximal promoter. There is substantial evidence in the literature that functional binding in prokaryotes can and does occur at much larger distances from promoters \( (40, 44–54) \). Several examples are known for \textit{M. tuberculosis} in particular. The regulation of GroEL2 by Rv2034 described above occurs from a binding site 746 bp upstream of the start codon for the gene \( (30) \). In another example, maximal expression of the \textit{espA} gene requires binding by EspR to the \textit{espA} activating region located between 1,004 and 884 bp upstream \( (55) \). A deletion
transcription. Recently, the distinction between proteins that modulate DNA structure and proteins that regulate transcription has blurred (57).

This is well illustrated by the example of EspR (Fig. 6). ChIP-Seq of EspR revealed binding to over 165 loci in the *M. tuberculosis* genome, distributed nearly equally between intergenic and genomic regions. Moreover, EspR binding data display substantial overlap with the binding sites for Lsr2, a known nucleoid associate protein in *M. tuberculosis* (8, 61, 62). EspR contains a helix-turn-helix DNA binding domain typical of many TFs and a C-terminal domain that mediates dimerization (63). Structural studies have led to a model in which EspR acts as a dimer of dimers in which each helix-turn-helix domain in a dimer can bind to distantly separated binding sites (64). This model was corroborated by atomic force microscopy that revealed DNA bending and DNA loop formation in conjunction with EspR binding. Together these data led to the hypothesis that EspR does not behave like a traditional TF but instead regulates transcription globally like a NAP through long-range interactions and DNA structure modifications (31). The results of the large-scale mapping of *M. tuberculosis* TFs suggest, however, that EspR may not be altogether unusual. Rather, considered in the context of the diverse binding of the majority of TFs in both prokaryotes and eukaryotes, these results suggest the possibility that other regulators may also act to modulate DNA structure on either a local or global scale.

**EXTENSIVE TF BINDING**

A nearly universal finding of TF ChIP-Seq studies is the unexpectedly large number of binding sites that are found for many, though not all, TFs. Even well-studied TFs can reveal a large number of novel sites. For example, in the case of DosR and KstR, 47 and 27 sites, respectively, had been associated with these factors. ChIP-Seq mapping identified 361 novel sites for DosR and 310 novel sites for KstR. Across all TFs mapped, the number of binding sites associated with each factor can be roughly fit to a power law distribution (p[k] ~ k^−α). This pattern reveals that there are some TFs that have numerous binding sites and thus likely interact with many genes. TFs of this sort have been termed “hubs” in the regulatory network literature. In contrast, most other TFs have fewer binding sites, and there is a long tail of TFs that have only a few binding sites. In some examples, we have observed a TF to bind only to its own promoter. Such cases may be explained by the TF modulating the expression dynamics of the operon in which it resides (65) or may reflect regulation of neighboring genes through divergent promoters.

One of the surprises to come out of the first set of 50 TFs mapped using ChIP-Seq in *M. tuberculosis* was the presence of Rv0081 as a hub TF. Rv0081 is the only other TF that is part of the initial hypoxic response (19, 23), but it has not been well studied. Rv0081 binds at 560 regions with 880 predicted binding locations, and its overexpression differentially regulates over 400 genes equally split between activation and repression. Prior to mapping, there were few clues that Rv0081 might have such a large regulatory role. Its identification as a hub, and the more recent identification of other previously unknown hubs, is one benefit of an unbiased and comprehensive approach to mapping *M. tuberculosis* TFs.

For TFs with many binding sites, the distribution of binding affinities often follows a common broad pattern. There is a set of binding peaks with higher apparent affinity and a gradual long tail of binding sites with weaker affinity. It should be cautioned that there are numerous variations within this theme. There are examples of TFs for which a clear set of strong peaks can be differentiated from the set of weak peaks. In other cases, there is simply a gradual decrease in affinities across binding sites. Many other patterns can also be seen, but a common element to most of these patterns is the presence of numerous weaker binding sites.

Extensive weak binding has been observed for TFs in every eukaryotic organism in which ChIP-Seq mapping of TFs has been performed (34, 56, 66–72). The data from *M. tuberculosis* (1) extends this observation to prokaryotes and indicates that extensive weak binding may be a general property of TFs. The physiological significance of this weak binding is extensively debated. As described above, substantial data support the contention that weak binding represents sequence-specific binding sites with true affinity for the corresponding TF.
The question remains, however, as to whether such binding sites have physiological significance.

There are specific examples of low-affinity binding sites with confirmed regulatory roles. For both DosR and KstR, although the majority of known regulatory interactions correspond to the strongest binding peaks, many known interactions correspond to weak peaks. For KstR in particular, the majority of novel weaker binding sites have greater affinity than the weakest-affinity binding site for a previously identified regulatory interaction. The work by Gao and colleagues on Rv2034 provides additional examples (29, 30). The confirmed binding of Rv2034 to the DosR promoter is one of the weaker binding sites for Rv2034 in the ChIP-Seq data (1.4% of the maximal binding peak). Rv2034 was also shown to regulate PhoP and bind to its promoter in EMSA experiments, and ChIP-Seq mapping identified a corresponding weakly enriched binding region at the far threshold for calling a binding site.

More broadly, the analysis of potential regulatory interactions for the first set of 50 TFs mapped by the consortium effort revealed a correspondence between binding affinity and the ability to assign a regulatory effect. By integrating expression data derived from the induction of each TF with the ChIP-Seq binding data for the corresponding TF, we sought to identify binding sites that mediated strong regulatory effects (1). This analysis was able to assign putative regulatory roles for 25% of binding sites, a number in line with analyses in previous systems (34). Of particular interest, stronger binding sites were more often associated with regulation than weaker sites. This result suggests a correlation between binding affinity and regulatory impact.

This is consistent with an analysis of binding sites detected in yeast using ChIP-ChIP (68). Consistent with ChIP-Seq data in other organisms, ChIP-ChIP in yeast revealed that weak binding sites likely represented the majority of binding events for most TFs. The substantial noise associated with ChIP-ChIP compared to ChIP-Seq necessitates caution in the analyses of these data. Yet despite this, a clear relationship was observed between the predicted binding energy of promoters for different TFs and the regulatory effect of perturbing that TF. Most notably, this trend was observed even for promoters whose effect fell below standard significance thresholds.

Together with the results shown in Fig. 7, these data suggest a more analog role of TF binding and regulation than typically considered. Rather than a TF modulating the expression of a regulon in a binary manner—in which all regulon genes are bound and induced/repressed in an on/off manner—the analog view suggests a continuum of binding and regulatory impact (68, 69). At one end, very-high-affinity sites may be bound at even low TF concentration and have strong regulatory effects. At the other end, weaker affinity sites may be more selectively bound or only bound in a fraction of cells, because TF concentrations are varied. These sites may have weaker regulatory effects or may serve to fine-tune other regulatory interactions. For example, substantial evidence exists that weak sites may play a significant role in modulating the effects of other binding sites through cooperative interactions (73). The distribution of binding site affinity may also play a role in sculpting the relative timing of expression of genes within a regulon (74, 75).

These data also suggest why strong sites are more likely to have been previously identified, since most methods for detecting regulation are designed to find substantial changes in gene expression. A common criterion is to select only genes that display a greater than 2-fold change in expression, yet the data above suggest that physiological binding sites may have relevant regulatory effects that are below this threshold. More generally, it is unlikely that any instrumentation-based threshold would happen to match the biological thresholds for physiological relevance for all TFs, even if such biological thresholds existed.

Of course, the null hypothesis is that many of these weak sites may not be functional in any meaningful or detectable sense. They may reflect the limitations of tuning the binding affinity of TFs to exclude spurious sites that are likely to arise in a large genome (although the existence of TFs that only bind a few sites argues this is not impossible). They may also lead to noise in transcription that is simply filtered out by other mechanisms (i.e., through the degradation of unstable transcripts arising from nonstandard transcription start sites). Many hypotheses are possible. The growing data from ChIP-Seq in both eukaryotes and now prokaryotes makes this an important and outstanding question in gene regulation research.

TBDB AND DATA AVAILABILITY

ChIP-Seq data for M. tuberculosis TFs generated by the NIAID TB Systems Biology Project have been released through the Tuberculosis Database (TBDB.org) (Fig. 9). TBDB is an online database providing integrated access to genome sequence, expression data, literature curation, and systems biology data for
M. tuberculosis and related genomes (76). TBDB currently houses genome assemblies for numerous strains of M. tuberculosis as well as assemblies for over 20 strains related to M. tuberculosis and useful for comparative analysis. It also houses resequencing data for over 31 M. tuberculosis strains selected as part of the M. tuberculosis Phylogeographic Diversity Sequencing Project. These data provide a global view of the genomic diversity of M. tuberculosis at the level of SNPs and indels. TBDB stores pre- and postpublication gene-expression data from M. tuberculosis and its close relatives, including over 3,000 M. tuberculosis microarrays, 95 RT-PCR datasets, 2,700 microarrays for human and mouse tuberculosis-related experiments, and 260 arrays for Streptomyces coelicolor. In addition, metabolic reconstructions have been performed on all organisms in the site, and these models are hosted as Biocyc Pathway/Genome databases (http://biocyc.org/) in TBDB. To enable wide use of these data, TBDB provides a suite of tools for searching, browsing, analyzing, and downloading the data.

TBDB also provides a growing set of tools for utilizing the M. tuberculosis ChIP-Seq data (Fig. 9). Through TBDB, users can search for regulatory binding sites by regulator, by target, or by genomic coordinate. Users can also browse a regulatory network constructed from these data. From the results of any of these searches, users may view the regulatory network for the gene of interest, select and view raw ChIP-Seq data in the dynamic real-time genome browser GenomeView (77), view the summary page for each experiment, or view static images of the ChIP-Seq peak data. Users may browse experiments directly and view the entire genome for each experiment. Users can also download all raw data (Fig. 9).

APPENDIX
Day 1
Protein-DNA cross-linking

- Grow 50 ml of M. tuberculosis cells to mid log or OD₆₀₀ = 0.5 to 0.6.
- Add formaldehyde to 1% (final concentration): 1.35 ml of 37% formaldehyde to 50 ml of culture.
Rock at room temperature (RT) for 30 min.
Stop cross-linking by adding final 250 mM of glycine (4.17 ml of 3 M glycine to 50 ml); rock at RT for 15 min.
Spin down the cells at 4°C for 10 min and 3,000 rpm.
Wash two times with ice-cold 1× PBS (50 ml) and spin down the cells at 4°C for 10 min and 3,000 rpm.
Resuspend cells in 0.5 to 0.6 ml of freshly made buffer 1 + PI (protease inhibitor, Complete Mini).
Lyse the cells using MagNA Lyser four times at 4,000 rpm and 45 sec. Keep the samples on ice between cycles.
Spin down the cells at 4°C for 10 min and 13,000 rpm.
Transfer the lysate to Covaris tubes.

Cell lysis and DNA shearing
- Do Covaris for 18 min for lysate, or 25 min if you skip the above three steps (amplitude = 20%, intensity = 5, cycles/burst = 200).
- Spin down the lysate for 10 min at 13,000 rpm and 4°C.
- Transfer supernatant to a new tube.
- Adjust salt concentration to:
  - 10 mM Tris HCl pH 8 (Add 10 μl of 1M Tris/ml of sample.)
  - 150 mM NaCl (Add 30 μl of 5 M NaCl/ml of sample.)
  - 0.1% NP40 (Add 10 μl of 10% NP40/ml of sample.)
  - Invert tubes to mix.

Immunoprecipitation
- Add 5 μl of anti-FLAG antibody to the lysate.
- Incubate overnight at 4°C on the rotating platform.

Day 2
- Rinse 50 μl of protein-G agarose beads (end of the tip [200 μl] should be cut by scissors for taking beads) with 1 ml of IPP150 buffer.
- Spin for 2 min at 2,000 × g to pellet the beads and discard the buffer.
- Transfer the lysate-Ab to the tube containing beads.
- Incubate for 2 hours at RT (0.5 hour at 4°C and 1.5 hours at RT) on a rotating platform.

Washes and elution of protein-DNA complexes
- Wash at least five times with 1 ml of IPP150 buffer. (Invert the tubes for 2 min between the washes and spin for 2 min at 2,000 × g to pellet the beads.)
- Wash at least two times with 1 ml of 1× TE. (Invert the tubes for 2 min between the washes and spin for 2 min at 2,000 × g to pellet the beads.)
- Add 150 μl of “elution from the beads” buffer.
- Incubate at 65°C for 15 min.
- Spin for 5 min at 2,000 × g to pellet the beads; keep the supernatant (first elution).
- Add 100 μl of 1× TE + 1% SDS (has to be fresh) to the pellet and incubate at 65°C for 5 min for second elution.
- Spin for 5 min at 2,000 × g to pellet the beads, keep the supernatant, and pool the samples (first and second elution).

Reverse cross-linking
- Add proteinase K (1 mg/ml final concentration) and incubate at 37°C for 1 to 2 hours.
- Incubate at 65°C overnight.

Day 3
DNA purification
- Purify the DNA with the PCR purification kit from Qiagen.
- Elute with 30 μl of EB.

Library preparation
The library is prepared using the standard Illumina protocol (http://www.broadinstitute.org/annotation/tbsysbio/Protocols/ChIPSeq_Protocol_2.pdf).
Transcription Factor Binding Site Mapping Using ChIP-Seq

Specific Buffers

Buffer 1 and buffer 1 + PI

<table>
<thead>
<tr>
<th>Composition</th>
<th>Stock solutions</th>
<th>Vol to add for 40 ml</th>
<th>Vol to add for 10 ml</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 mM KHEPES pH 7.9</td>
<td>1 M KHEPES</td>
<td>800 µl</td>
<td>200 µl</td>
</tr>
<tr>
<td>50 mM KCl</td>
<td>1 M KCl</td>
<td>2,000 µl</td>
<td>500 µl</td>
</tr>
<tr>
<td>0.5 mM DTT</td>
<td>0.5 M DTT</td>
<td>40 µl</td>
<td>10 µl</td>
</tr>
<tr>
<td>10% glycerol</td>
<td>glycerol</td>
<td>4 ml</td>
<td>1 ml</td>
</tr>
<tr>
<td>+ Water</td>
<td>37.2 ml</td>
<td>8.290 ml</td>
<td></td>
</tr>
</tbody>
</table>

To make buffer 1 + PI, take 2 ml of buffer 1 and add a half tablet of Mini Protease Inhibitor.

IPP150 buffer

<table>
<thead>
<tr>
<th>Composition</th>
<th>Stock solutions</th>
<th>Vol to add for 250 ml</th>
<th>Vol to add for 50 ml</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 mM Tris-HCl pH 8</td>
<td>1 M</td>
<td>2.5 ml</td>
<td>500 µl</td>
</tr>
<tr>
<td>150 mM NaCl</td>
<td>5 M</td>
<td>7.5 ml</td>
<td>1.5 ml</td>
</tr>
<tr>
<td>0.1% NP40</td>
<td>10%</td>
<td>2.5 ml</td>
<td>500 µl</td>
</tr>
<tr>
<td>+ Water</td>
<td>237.5 ml</td>
<td>47.5 ml</td>
<td></td>
</tr>
</tbody>
</table>

“Elution from beads” buffer

<table>
<thead>
<tr>
<th>Composition</th>
<th>Stock solutions</th>
<th>Vol to add for 10 ml</th>
</tr>
</thead>
<tbody>
<tr>
<td>mM Tris-HCl pH 8</td>
<td>1 M</td>
<td>500 µl</td>
</tr>
<tr>
<td>10 mM EDTA</td>
<td>0.5 M</td>
<td>200 µl</td>
</tr>
<tr>
<td>1% SDS</td>
<td>20%</td>
<td>500 µl</td>
</tr>
<tr>
<td>+ Water</td>
<td>8.8 ml</td>
<td></td>
</tr>
</tbody>
</table>

REFERENCES


12. (Reference deleted.)


52. Reitzer LJ, Magasanik B. 1986. Transcription of glnA in *E. coli* is stimulated by activator bound to sites far from the promoter. *Cell* 45:785–792.


