# Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites

# Supplementary Material

Stephen A. Ramsey, Theo A. Knijnenburg, Kathleen A. Kennedy, Daniel E. Zak,
Mark Gilchrist, Elizabeth S. Gold, Carrie D. Johnson, Aaron E. Lampano,
Vladimir Litvak, Garnet Navarro, Tetyana Stolyar,
Alan Aderem, and Ilya Shmulevich

Institute for Systems Biology
1441 N. 34th St., Seattle, WA 98103, USA

September 30, 2011

This document contains supplementary information for the article *Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites*, including supplementary methods, tables, and figures. The dataset used for training and testing the model, and the MATLAB source code for the software developed for binding site predictions and model training, are freely available for download at magnet.systemsbiology.net/hac.

## S1 Supplementary Methods

This section gives additional details on the computational and experimental methods used in the study, to supplement the Methods section of the main article.

### S1.1 Computation

Unless otherwise indicated, data processing was performed using software written in MATLAB (release R2009b; MathWorks, Natick, MA) or in Perl (version 5.8.8). TF binding prediction and model training were performed using software written in MATLAB. Computational work was performed on a cluster of 64-bit, eight-core Intel Xeon systems, each with 32 GB RAM and running the CentOS 5.4 operating system (GNU/Linux kernel version 2.6.18).

### S1.2 Genomic regions

In keeping with the goal of evaluating HAc ChIP-Seq-based features for their utility in predicting TF binding sites in the same cell type (macrophages) from which the ChIP data were derived, the study was carried out in transcript-proximal genomic regions for genes that are expressed in murine macrophages (as was the promoter motif scanning analysis in the transcriptional profiling study of Ramsey *et al.* [1]). Genes that were not detected as expressed in either unstimulated or activated macrophages were excluded. This restriction was to ensure that the conclusions of the study would be based on TF binding prediction performance for genes where TF binding might conceivably have a functional consequence in macrophage activation. The genomic regions were selected using transcriptional profiling data, as described below.

**Transcriptional profiling:** A list of likely transcription start sites in both resting and activated murine macrophages were identified using transcriptional profiling data [2]. Murine bone marrow-derived macrophages (BMMs) were cultured from female C57BL/6 mice (age 8-12 weeks) as previously described [1], and on day 6, cells were plated into six-well tissue culture plates. On day 7, cells are incubated for 4 h in either (i) complete RPMI 1640 medium (Invitrogen) with recombinant human macrophage colony simulating factor (rhM-CSF) or (ii) complete RPMI and rhM-CSF and one of the following purified Toll-like receptor (TLR) agonists (sourced as described in [1]): lipopolysaccharide (LPS, 10 ng/mL), polyriboinosinic polyribocytidylic acid (poly I:C, 6 μg/mL), or synthetic triacylated lipopeptide ($\mathrm{Pam_3CSK_4}$, henceforth abbreviated as “PAM3”, 300 ng/mL). Macrophage preparations were performed in three biological replicates for each treatment type, and RNA was isolated using Trizol (Invitrogen) following the instructions accompanying the reagent. For each replicate, 1 μg of RNA was amplified and labeled using the Affymetrix single-step protocol and hybridized to Affymetrix Mouse Exon 1.0 ST Array GeneChips (Affymetrix, Santa Clara, CA). The GeneChips were scanned using the Affymetrix GeneChip Scanner 3000 and processed into probe-level intensity (“`.cel`”) files using the Affymetrix GeneChip Operating Software.

The probe intensity data were analyzed as follows. For each Core Affymetrix transcript meta-probeset (Affymetrix Mouse Exon Array 1.0 ST annotations version 1, release 2), a $P$ value representing the likelihood of seeing probe intensities as high as what was observed for the probes within that meta-probeset by chance was obtained using a GC-based background model and using the Affymetrix Detect Above Background (DABG) algorithm as implemented in the software package `xps` [3] in the R system (version 2.9.0) [4]. For each transcript listed in the Ensembl mouse genome database (v56) [5], $P$ values within all replicates for each condition were combined using the geometric mean, and the overall minimum $P$ value for the transcript was taken as the representative $P$ value for the transcript's probe-level intensities being attributable to background, in all conditions. A $P$ value cutoff was determined by requiring a false discovery rate of less than 0.01%. For each transcript for which the $P$ value was less than the cutoff in at least one condition, the transcript start site (TSS) was obtained from the Ensembl database using the BioMart tool [6]. After excluding transcripts that could not be uniquely mapped to an Ensembl transcript, a total of 13,906 transcript start sites were obtained, corresponding to 12,863 independent genes. This corresponds to 53% of the genes in the mouse genome, consistent with a previous transcriptomic analysis of genes expressed in macrophages [1].

**Mapping genomic regions:** The base-pair-wise union of all regions ± 5 kbp from the transcription start sites were mapped and the sequences for these regions were obtained using the UCSC Genome Browser [7]. As the reference mouse genome assembly, NCBI37/mm9 (July 2007) was used. The resulting sequences were used to generate the TF binding site motif scanning feature, the GC content feature, and the nucleosome occupancy prediction score feature for the TF binding site prediction algorithm. The genomic regions were then divided into consecutive non-overlapping 100 bp intervals.

## S1.3 Motif scanning

For each TF $t$ (where the index $t$ runs from 1 to 5, spanning the list of the five transcription factors analyzed in this study and shown in Table S2), all motif position-weight matrices (PWMs) that are annotated as being recognized by that TF were obtained from the TRANSFAC Professional database [8]. The PWMs that were used are shown in Table S2. For notational clarity, let each matrix be denoted by $M_{tm}$, where $m$ runs over the set of motifs for TF $t$. Sequence files were scanned for each of these matrices using the likelihood-based PWM scanning implementation of Lähdesmäki *et al.* [9], using a $0^{\text{th}}$-order background Markov model (trained on the sequence file for the same regions in which the TF prediction was done). The scanning produced a likelihood ratio $r_{ijtm}$ for each PWM $M_{tm}$, at each base position $j$, in each interval $i$ (where $i$ labels one of the 100 bp intervals from the genomic regions selected as described in Sec. S1.2). The final TF motif scanning feature value for TF $t$ at interval $i$, denoted by $v_{1it}$, was computed as follows,

$$v_{1it} = \text{trank}\left(\max_m \left(\max_j (q(r_{ijtm}))\right)\right)$$

where "trank" is the tied rank (computed using the MATLAB Statistics Toolbox), normalized to a maximum value of 1, and $q$ is a function defined as follows:

$$q(x) = \begin{cases} 0, & x \leq 1, \\ \log_{10}(x), & x > 1. \end{cases}$$

## S1.4 ChIP-Seq

BMMs were cultured from female C57BL/6 mice as previously described [1] and stimulated on day 7 for the combinations of stimulus and stimulus durations listed in Table S2. Immunoprecipitation (IP) was carried out as described in [10] (for acetyl-H4 and NFκB/p65 targets, using Protein A Dynabeads from Invitrogen), using antibodies as described previously for ATF3 [10], IRF1 [1], NFκB/p50 [1], and C/EBPδ [11]. The antibody used for the NFκB/p65 IP was from Abcam (catalog number ab7970-1). The antibody used for the acetyl-H4 IP (Millipore catalog number 06-866) is a rabbit polyclonal IgG derived using an immunogen consisting of an 18 amino acid peptide from the N-terminal tail of histone H4, acetylated on lysines 5, 8, 12, and 16 of the peptide (see the Millipore online catalog page for this reagent, for additional information). A sequencing library for the Illumina Genome Analyzer was derived from the IP using the Illumina reagent kit, as previously described [2]. Single-ended, 36-cycle sequencing was performed on an Illumina Genome Analyzer, and the raw image data were processed using the Illumina Genome Analysis Pipeline Software on a dedicated sequence data processing system [12]. Reads were aligned to the mouse genome using eland_extended with an `ELAND_SEED_LENGTH` value of 25 and an `ELAND_MAX_MATCHES` value of 15, and with the $3'$-most base excluded. Reads aligned to the same position and strand were counted only once to eliminate duplicates from PCR amplification (consistent with the approach of [13]). For all ChIP-Seq samples, aligned reads were processed into extended fragments (consistent with the approach of [14]) of length 158 bp, the estimated typical insert size in the sequencing library. This estimated size was determined by assaying representative ChIP-Seq samples (after the PCR amplification step) using the Agilent Bioanalyzer to determine the typical fragment size in the sequencing library, and subtracting the combined size of the two Illumina adaptor molecules. Replicate ChIP-Seq experiments (typically, two replicates per combination of condition and target) were averaged by combining their groups of extended fragments. The number of overlapping fragments was then counted at survey points every 10 bp, starting with the first bp of the chromosome. The resulting counts were multiplied by $10^6$ and divided by the total number of aligned reads to obtain the replicate-combined ChIP-Seq "signal" along the chromosome. A control ChIP-Seq signal was also obtained from three IPs of BMMs with immunoglobulin G derived from rabbits that were not immunized with the specific target antigen (Santa Cruz Biotechnology). Within regions of the genome where the control ChIP-Seq signal was above one fragment per million, the ChIP-Seq signal for target-specific IPs was set to zero.

## S1.5 HAc ChIP-Seq

The background-masked acetyl-H4 ChIP-Seq signal tracks (which were at 10 bp resolution, and obtained as described in Sec. S1.4) were sampled within each 100 bp interval using the maximum function. These 100 bp values were then used as a feature for TF binding site prediction. For each TF, and within the binding site prediction model, HAc ChIP-Seq data were used from both unstimulated cells and from cells stimulated with LPS for the same duration as in the TF ChIP-Seq experiment (see Table S2).

## S1.6 HAc valley scores

In this section, the motivation and implementation of the HAc "valley scores" are described.

### S1.6.1 Motivation

Preliminary analysis of ChIP-Seq data for TFs and histone acetylation in murine macrophages revealed that transcription factor binding sites appear to be concentrated at local minima of histone acetylation (with characteristic sizes of approximately 200-400 bp) within regions of high histone acetylation (see Fig. S1). In order to quantitatively investigate this observation, a "valley score" signal was derived from the HAc ChIP-Seq data, to measure the depth of a local minima within a histone acetylated region. A high "peak" of the valley score signal corresponds to a strong "dip" in a region of otherwise high HAc ChIP-Seq signal (see Fig. S1 below, and Fig. 1A in the main article). A preliminary analysis of valley scores at TF binding sites revealed that the probability distribution for the valley score at a randomly selected genomic location is significantly affected by whether or not there is a TF binding site at that location (see Fig. 1B, main article). As a potential mechanism that could explain the statistical association between TF binding sites and HAc local minima, we hypothesize that HAc local minima may represent small nucleosome-excluded pockets within loci of high histone acetylation. Such nucleosome-excluded pockets might be more readily accessible to TFs than adjacent nucleosome-bound chromatin. DNase I digestion assays have confirmed that the chromatin accessibility of non-nucleosomal pockets is strongly increased by the acetylation of flanking histones [15]. Based on these considerations, the HAc valley score signal was evaluated for its potential utility for predicting TF binding. As described in the main article and in detail below, the highest HAc valley score values within 100 bp intervals were used as a feature for motif-based prediction of TF binding sites in macrophages. The predictive utility of the valley score feature was compared to that of a feature derived directly from the HAc ChIP-Seq signal (see Sec. S1.5).

### S1.6.2 Implementation

Valley scores were computed based on the HAc ChIP-Seq signal (see Sec. S1.5) sampled at a resolution of 10 bp. First, the HAc ChIP-Seq signal was smoothed by convolving it with a Gaussian kernel with a standard deviation of 60 bp. Next, local minima of the HAc ChIP-Seq signal were identified: for each sample point, the maximum signal value in the window 100 to 500 bp to the right of that point as well as in the window 100 to 500 bp to the left of that point were computed using a sliding window approach. If the signal value at the sample point was less than 90% of the minimum of these two local maxima, this sample point was called a "valley". The "valley score" assigned to this point is the minimum of these two local maxima. For all sample points that were not identified as a valley, the valley score signal was set to zero, thus reducing the data track to only the local minima of the HAc ChIP-Seq signal. Finally, the valley score signal was downsampled to a resolution of 100 bp, by taking the maximum value within each 100 bp interval.

## S1.7 TF ChIP-Seq peak detection

The ground-truth set of TF binding sites for the five TFs ATF3, C/EBPδ, IRF1, NFκB/p50, and NFκB/p65 was obtained from the background-masked ChIP-Seq signal track for each TF using conservative signal level thresholds, as described below. A survey point was identified as a binding location if and only if three conditions were simultaneously satisfied: (i) the ChIP-Seq signal in the TF-specific IP was at least five times the control ChIP-Seq signal (see Sec. S1.4); (ii) the ChIP-Seq signal in the TF-specific IP was at least two fragments per million; and (iii) the ChIP-Seq signal in the TF-specific IP corresponded to at least six overlapping fragments (extended reads) at that survey point. Any 100 bp interval for which one of the 10 bp survey points within the interval satisfied these thresholds was identified as containing a TF binding site. TF binding sites identified in adjacent 100 bp intervals (due to the average insert length of 158 bp resulting from the ChIP-Seq library preparation) were accounted for in the computation of the false negative error rate, to prevent double-counting (see Sec. S1.11). The numbers of binding sites obtained for the five TFs, within the ∼7% of the genome analyzed for this study, are given in Table S2.

## S1.8 GC content

From the sequence files for genomic regions analyzed in this study (see Sec. S1.2), the fractions of G or C bases within adjacent 10 bp intervals (a typical motif size, [16]) was computed. Within each 100 bp interval, the maximum of 10-bp-average GC content value was computed, and used as the value for the GC content feature.

## S1.9 Conservation

Chromosome-specific files containing the PHAST 30-way vertebrate species conservation scores [17] at every base pair position were obtained from the UCSC Genome Browser [7] in Wiggle (WIG) format. Because the average size of evolutionarily conserved elements across vertebrate genomes (100-120 bp, according to [18]) is significantly larger than the typical TF-binding *cis*-regulatory element size, the sequence conservation data were downsampled within 100 bp intervals to produce a TF-generic conservation feature that could be used for identifying potential *cis*-regulatory regions, similar to the approach of [19]. The downsampling was performed in two steps. First, at survey points every 10 bp along the chromosome, the average PHAST alignment score was computed for the 5 bp flanking positions on each side of the survey point. Second, within each 100 bp interval of the genomic regions analyzed for the study (see Sec. S1.2), the maximum of this 10 bp-averaged conservation score value was computed among all 10 bp survey points contained within the interval, and used as the value for the conservation feature.

## S1.10 Nucleosome Occupancy Prediction Score

The nucleosome occupancy prediction model of Kaplan *et al.* [20] is a sequence-based probabilistic model for predicting the likelihood that, at a given position in the genome, a nucleosome will be present. The model was based on *in vitro* nucleosome binding data for 40,000 different 150 bp sequences. Based on the observation in [20] that the nucleosome occupancy score tends to have a local minimum at transcription start sites, and based on the theory that transcription factors may preferentially bind nucleosome-excluded regions [21], the nucleosome occupancy score was selected as a possible feature that may anticorrelate with transcription factor binding. Nucleosome occupancy probabilities were obtained at every base pair position within the genomic regions analyzed this study (see Sec. S1.2), by passing the genomic sequences through the nucleosome positioning prediction program version 3 of Kaplan *et al.* [20] (64-bit version), with the tabbed output option. Within each 100 bp interval, the geometric mean of the nucleosome occupancy probabilities ("P occupied") for all positions within that interval was computed. The resulting geometric-mean probabilities were used as the nucleosome occupancy feature.

## S1.11 Model Performance Metric

For each transcription factor $t$, and for each prediction model, the prediction score cutoff $\sigma$ was varied and the resulting sensitivity $S(\sigma, t)$ and false positive error rate $E(\sigma, t)$ values were obtained as described in the main article (Methods section, Performance Metric). The range of $\sigma$ values were determined by computing quantiles of the distribution of nonzero $\sigma_t$ prediction scores, for quantiles defined by $1 - 10^R$, where $R$ is a vector that uniformly samples values between $\log_{10}(1/X)$ ($X$ is the total number of 100 bp intervals) and $\log_{10}(E_{\max})$, where $E_{\max}$ is the maximum false discovery rate for computing the performance as the area under the $S$ vs. $E$ curve (i.e., receiver operating characteristic or "ROC" curve). Due to the extremely sparse nature of TF binding in the genome, the value $E_{\max} = 0.01$ was used, so that model training would not be biased by the sensitivity levels recorded above a FPR of 0.01 (which would have little practical utility for TF binding site prediction, due to the signficant genomic distance scales over which mammalian *cis*-regulatory elements are distributed relative to the transcription start site of the gene they control [22]). With the 100 bp binning used in this study, a model with a FPR of 0.01 would have an erroneous prediction, on average, once per 10 kbp of promoter sequence (consistent with the threshold used in [23]). The practice of using, as a measure of prediction performance, the ROC curve for the specificity range relevant to a particular application (the so-called "partial area under the ROC curve" measure [24]) has been used in several studies of TF binding or *cis*-regulatory module prediction [25, 26, 27]. For the case of model training, the $S$ vs. $E$ curve was sampled at 20 points; for the case of model validation, 100 samples were used. This yielded $(E, S)$ values for the range $0 < E \leq E_{\max}$, which was then numerically integrated

$$A(t) = \int_0^{E_{\max}} S(\sigma, t) dE(\sigma, t) \tag{S1}$$

using the trapezoidal rule. A nonnegative, TF-specific cost term was then defined by

$$C(t) = 1 - \frac{A(t)}{E_{\max}}. \tag{S2}$$

Because $S(\sigma, t) \leq 1$, the maximum value for $A(t)$ is $E_{\max}$, and thus $0 \leq C(t) \leq 1$. Normalizing the cost in this manner simplifies the implementation of penalty terms to enforce constraints (see Sec. S1.12) and the handling of cases where $C(t)$ can not be computed using Eqs. S1–S2. Specifically, because the prediction model was globally optimized, some sets of model parameters were evaluated for which (due to degeneracy of the prediction scores $\sigma_{it}$) it was not possible to obtain at least 20 unique $(E, S)$ samples in the range $0 < E \leq E_{\max}$. In such cases, instead of using Eqs. S1–S2, the cost $C(t)$ was computed as twice the difference between the desired number of samples (20) and the actual number of unique $(S, E)$ samples (within the range $0 < E \leq E_{\max}$) that was obtained for that set of model parameters.

## S1.12 Model Training

For each feature combination $F$, the model parameters $\{\vec{\lambda}, \vec{\mu}, \vec{\omega}\}$ were optimized to minimize the average of the TF-specific cost function across the set $T$ of TFs selected for training,

$$C = \langle C(t) \rangle_{t \in T}$$

subject to the constraints that the weights have unit L1 norm,

$$\sum_{f \in F} |\omega_f| = 1;$$

and that the min/max thresholds be properly ordered:

$$\lambda_f \leq \mu_f, \;\; \forall f \in F.$$

For the general case of features that may anti-correlate with *cis*-regulatory function, negative weight components are permitted. Initial parameter values for the optimization were selected by a procedure that can be described in three cases. (i) For the case of the model consisting only of the rank-transformed motif scanning feature (i.e., the motifs-only model), initial parameter values for the optimization were $\lambda_1 = 0.98$ and $\mu_1 = 1$. (ii) For the case of a model consisting of motifs plus one additional feature, the initial parameter values for the optimization were chosen as follows. The initial values for $\lambda_1$ and $\mu_1$ were taken from the best parameters for the motifs-only model, and the initial value for $\omega_1$ was 0.99. The initial values for the thresholds for the additional feature were computed as

$$\lambda_f = \min_{i,t}(v_{fit}), \tag{S3}$$

and

$$\mu_f = \max_{i,t}(v_{fit}), \tag{S4}$$

respectively. (iii) For the case of a model consisting of motifs plus "HAc ChIP VS (S)" plus one additional feature (see Table S2 and Fig. S5), the initial $\lambda_1$ and $\mu_1$ values for the motif feature and the "HAc ChIP VS (S)" feature were obtained from the best parameter set for the two-feature model consisting those two features. The remaining feature's initial threshold values were computed as shown in Eqs. S3–S4, respectively. The initial weight value for the additional feature was taken to be 0.01/1.01, and the initial weight values for the motif feature and HAc feature were taken to be the weight values from the optimized two-feature model containing those two features, divided by 1.01 (dividing by 1.01 normalized the initial weight vector for the three-feature model).

Model parameters were optimized in a two-stage approach. The first stage was a global optimization using a branch/fit algorithm, SNOBFIT (version 2.1) [28], with the best parameter set from this stage being used as the initial point for the second stage. The second stage optimizer was the `fminsearch` function in MATLAB (a simplex search algorithm), with additive penalty terms included in the cost function to ensure that the constraints were not violated. The function tolerance for `fminsearch` was $10^{-4}$. Because models with more features have a higher-dimensional parameter space, the total number of iterations for each optimization stage was limited to 300 times the square of the number of features in the model. The best parameter set from the second stage optimization were used as the model parameters for testing the model's prediction performance on the remaining TF. Robustness of the optimized parameter set was verified by re-optimizing various models using the two-stage approach as described above, and comparing best parameter values before and after re-optimization.

## S1.13 Model Testing

The performance of each model was tested across all five leave-one-out cross-validation scenarios (i.e., on each of the five TFs, and in each case using model parameters that were trained using the other four TFs). The performance metric was the area under the sensitivity vs. FPR curve, up to a maximum FPR of 0.01, as described in Sec. S1.11. As model consisting of only the motif scanning feature was used as a reference model. The prediction performance of the multiple-feature models were then compared to the reference model (see main article, section Methods, subsection Model Testing).

As a negative control for the motif-based TF binding site prediction approach, the prediction performance of a set of random predictions was measured. To make random predictions, a prediction score was assigned randomly (drawn from the unit interval with uniform distribution) to each interval. The resulting random prediction scores were analyzed for prediction performance, as described in Sec. S1.11.

# S2 Supplementary Tables

| # | Feature description | Abbrev. in Fig. 1 | Methods sec. | Source |
|---|---|---|---|---|
| 1 | Motif scanning information | Motifs | Sec. S1.3 | genome, [8] |
| 2 | HAc ChIP-Seq signal, stimulated cells | HAc ChIP (S) | Sec. S1.5 | ChIP-Seq |
| 3 | HAc ChIP-Seq signal, unstimulated cells | HAc ChIP (U) | Sec. S1.5 | ChIP-Seq |
| 4 | HAc valley scores, stimulated cells | HAc ChIP VS (S) | Sec. S1.6 | ChIP-Seq |
| 5 | HAc valley scores, unstimulated cells | HAc ChIP VS (U) | Sec. S1.6 | ChIP-Seq |
| 6 | Maximum 10-bp-subinterval GC content | GC Content | Sec. S1.8 | genome, n/a |
| 7 | 30-way vertebrate conservation score | Conservation | Sec. S1.9 | [17] |
| 8 | Nucleosome occupancy prediction score | Nucleos. Occ. | Sec. S1.10 | genome, [20] |

Table S1: Features used for predicting transcription factor binding sites. The column "#" gives the $f$ index value of the feature track (see Eqs. 1–2, main article). The column "Abbrev. in Fig. 1" indicates the abbreviation used to represent the feature, in the legend of Fig. 2B (main article). Each feature is described in a subsection of Sec. S1 as indicated in the "Methods sec." column of the table. The "Source" column describes how the feature track was obtained ("genome" means that the track was obtained from genomic sequence, and, where applicable, it is followed by the source of the pattern used for extracting the feature track from genomic sequence (see Sec. S1.2); "ChIP-Seq" means that the track was derived from ChIP-Seq data, as described in Secs. S1.4–S1.6.

| TF | Ref. | Condition | Sites (TPGRs) | Sites (genome) | TRANSFAC Motif PWMs |
|---|---|---|---|---|---|
| ATF3 | [10] | LPS, 4 h | 3,642 | 10,349 | ATF3_Q6, ATF_01, ATF_B, CREBATF_Q6, CREB_Q3 |
| C/EBPδ | [11] | PAM3, 6 h | 8,464 | 23,353 | CEBPDELTA_Q6, CEBP_01, CEBP_C, CEBP_Q2, CEBP_Q2_01, CEBP_Q3 |
| IRF1 | [29] | LPS, 4 h | 2,261 | 5,705 | IRF1_01, IRF1_Q6, IRF_Q6_01, IRF_Q6 |
| NFκB/p50 | [30] | LPS, 1 h | 2,186 | 4,387 | NFKAPPAB50_01, NFKAPPAB_01, NFKB_C, NFKB_Q6_01, NFKB_Q6_B0 |
| NFκB/p65 | [30] | LPS, 1 h | 4,188 | 8,475 | NFKAPPAB65_01, NFKAPPAB_01, NFKB_C, NFKB_Q6_01, NFKB_Q6_B0 |

Table S2: Transcription factors (TFs) whose binding sites were used for training and testing the TF binding site prediction model. The column “Ref.” gives a reference to an article describing the role of the indicated TF in macrophage activation under Toll-like receptor stimulation (e.g., under stimulation with lipopolysaccharide). The column “Condition” indicates the activation conditions for the macrophages in which binding of the indicated TF were measured (the reagents and their concentrations are specified in Sec. S1.2) The column “Sites (TPGRs)” gives the number of ChIP-Seq-derived ground-truth TF binding sites detected within the set of transcript-proximal genomic regions (TPGRs) used for this study (see Sec. S1.2). The column “Sites (genome)” gives the number of ChIP-Seq-derived TF binding sites detected, genome-wide, for each of the TFs in activated macrophages. The column “TRANSFAC Motif PWMs” gives the list of motif position-weight matrices (PWMs) used for predicting binding sites for that TF or TF family (the “V$” TRANSFAC PWM prefix, indicating a vertebrate-derived motif, is not shown). The relative numbers of binding sites in the TPGRs (comprising ∼7% of the genome) vs. the number of binding sites in the entire genome shows a significantly higher density of binding sites in transcript-proximal genomic regions than in the genome overall.

| Model name | Features | Figure | Mean | Std Dev |
|---|---|---|---|---|
| Random model | none | Fig. 2 | 0.000064 | 0.00001 |
| Motifs only | 1 | Fig. 2 | 0.00175 | 0.00160 |
| Motifs + HAc ChIP (S) | 1, 2 | Fig. 2 | 0.00199 | 0.00052 |
| Motifs + HAc ChIP (U) | 1, 3 | Fig. 2 | 0.00196 | 0.00047 |
| Motifs + HAc ChIP VS (S) | 1, 4 | Fig. 2 | 0.00266 | 0.00023 |
| Motifs + HAc ChIP VS (U) | 1, 5 | Fig. 2 | 0.00203 | 0.00035 |
| Motifs + GC content | 1, 6 | Fig. 2 | 0.00183 | 0.00005 |
| Motifs + Conservation | 1, 7 | Fig. 2 | 0.00179 | 0.00003 |
| Motifs + Nucleos. Occ. | 1, 8 | Fig. 2 | 0.00176 | 0.00002 |
| Motifs + HAc ChIP VS (S) + HAc ChIP (S) | 1, 4, 2 | Fig. S5 | 0.00251 | 0.00030 |
| Motifs + HAc ChIP VS (S) + HAc ChIP (U) | 1, 4, 3 | Fig. S5 | 0.00257 | 0.00028 |
| Motifs + HAc ChIP VS (S) + HAc ChIP VS (U) | 1, 4, 5 | Fig. S5 | 0.00267 | 0.00023 |
| Motifs + HAc ChIP VS (S) + GC content | 1, 4, 6 | Fig. S5 | 0.00269 | 0.00023 |
| Motifs + HAc ChIP VS (S) + Conservation | 1, 4, 7 | Fig. S5 | 0.00266 | 0.00024 |
| Motifs + HAc ChIP VS (S) + Nucleos. Occ. | 1, 4, 8 | Fig. S5 | 0.00263 | 0.00024 |

Table S3: Transcription factor binding site prediction models that were studied, and the features that each model used. Column "Model name" gives the name of the model, as it appears in the legend of the figure specified in the "Figure" column. The column "Features" gives the feature index numbers of the features that were used in each model; the feature index numbers refer to the order in which the features appear in Table S2. The column "Mean" gives the area under the sensitivity vs. FPR curve (i.e., receiver operating characteristic, or "ROC" curve), up to an FPR of 0.01, averaged over the five-fold cross-validation. The column "Std Dev" gives the standard deviation of the difference between the ROC curve area in the indicated model and the reference model, over the five-fold cross-validation (except in the case of the reference model and the random model, where the standard deviation of the area over the five-fold cross-validation is given. The complete ROC curves for the one- and two-feature models are shown in Fig. S4.

| TF name | Binding site type | Induced dip | Constitutive dip | Peak to dip |
|---|---|---|---|---|
| C/EBPδ | Inducible | 135 | 4691 | 2031 |
| | Constitutive | 0 | 263 | 38 |
| | Displaced | 0 | 43 | 6 |
| | | | $P$ value: | $8.1 \times 10^{-12}$ |
| IRF1 | Inducible | 58 | 1247 | 654 |
| | Constitutive | 0 | 31 | 7 |
| | Displaced | 0 | 7 | 3 |
| | | | $P$ value: | $2.1 \times 10^{-1}$ |
| NFκB/p50 | Inducible | 45 | 1341 | 674 |
| | Constitutive | 2 | 177 | 57 |
| | Displaced | 0 | 44 | 6 |
| | | | $P$ value: | $3.8 \times 10^{-4}$ |
| NFκB/p65 | Inducible | 53 | 2137 | 810 |
| | Constitutive | 0 | 130 | 22 |
| | Displaced | 0 | 100 | 27 |
| | | | $P$ value: | $5.7 \times 10^{-4}$ |

Table S4: At macrophage TF binding sites, classifications of TF binding (inducible, constitutive, or displaced) are associated with HAc feature classifications. For each of four TFs for which ChIP-Seq data were available in unstimulated murine bone marrow macrophages, TF binding sites within the transcript-proximal genomic regions analyzed in this study (see Sec. S1.2) were classified according to whether the binding sites were present in both unstimulated and stimulated macrophages (constitutive); present in stimulated macrophages only (inducible); or present in unstimulated macrophages only (displaced). At the subset of these locations where a nonzero HAc valley score was evident in either cell condition, the HAc peaks were classified according to whether a HAc local minimum was present only in stimulated cells, with no peak in unstimulated cells ("induced dip"), a HAc local minimum was present in both unstimulated and stimulated cells ("constitutive dip"), or a peak was evident in unstimulated cells and no peak was evident in stimulated cells ("peak to dip"). The three-by-three contingency tables for these classifications were then analyzed using the chi-square test. The null hypothesis that the TF binding and HAc classifications are independent, can be excluded for three out of the four TFs tested ($P < 0.001$). This suggests that condition-dependent changes in the HAc signal are associated with condition-dependent TF binding. More specifically, the "transient dips" (induced dip and peak to dip) occur much more frequently for the inducible TF binding events when compared to the constitutive and displaced binding events.

| TF | Conserv. | HAc ChIP (S) | HAc ChIP VS (S) |
|---|---|---|---|
| ATF3 | 5.9% | 25.0% | 190% |
| C/EBPδ | 6.7% | 43% | 152% |
| IRF1 | 0.7% | 28.4% | 22.8% |
| NFκB/p50 | 2.7% | -13.6% | 27.3% |
| NFκB/p65 | 8.1% | 55.3% | 149% |

Table S5: Relative change in prediction performance for each of the five TFs, in three different prediction models. Shown here are the percent changes in the partial area under the ROC curve for the prediction performance of each of the five TFs, in the test step (see Sec. S1.13), for three different two-evidence models (each relative to the motifs-only model). The partial areas under the ROC curves are computed for FPR $< 0.01$ (see Sec. S1.11). In each case, the prediction performance for binding sites for the indicated TF is based on model parameters obtained by training using binding site data from the other four TFs (see Sec. S1.13 for a details on the TF-based cross-validation). The three models shown here are as follows. *Conserv.* = motifs plus conservation; *HAc ChIP (S)* = motifs plus HAc ChIP-Seq signal from activated macrophages; *HAc ChIP VS (S)* = motifs plus HAc ChIP-Seq valley scores from activated macrophages. Incorporating HAc ChIP-Seq valley scores into the prediction model improves the prediction performance vs. the motifs-only model for all five TFs, and improves performance vs. the "motifs + HAc ChIP" model for four out of five TFs. Incorporating conservation data into the prediction model gives only a modest improvement to prediction performance, consistent with the findings of other studies [9, 31, 19].

# S3 Supplementary Figures

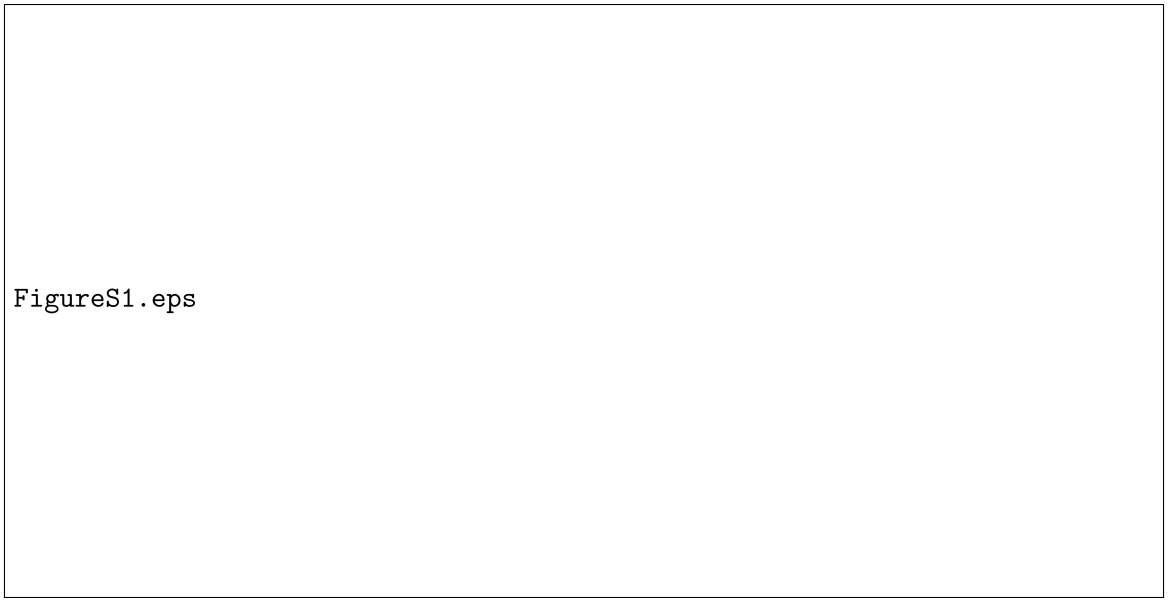

Figure S1: Transcription factor binding, as measured by ChIP-Seq, correlates with local minima of histone acetylation (HAc) ChIP-Seq signal in regions of overall high HAc signal. Shown here is a 40 kbp region of mouse chromosome 11 including the cytokine genes *Ccl3* and *Ccl4* (genes annotated at bottom of figure) that is hyper-acetylated in LPS-treated macrophages (as well as at the *Ccl5* promoter as shown in Fig. 1A of the main text, and at many other LPS-inducible promoters as described in [32]). Also shown are tracks representing signal from three ChIP-derived features sampled at 10 bp resolution. The orange track shows the average of ChIP-Seq data for all five transcription factors assayed in this study (see Table S2) in activated murine macrophages (see Sec. S1.7), giving an overall measure of TF-binding potential. The dark blue track shows valley scores indicating local minima of the HAc ChIP signal (see Sec. S1.6), which are seen to co-occur with many of the TF binding peaks. The magenta track shows the HAc ChIP-Seq signal in macrophages that have been stimulated for 1 h with LPS.

FigureS2.eps

Figure S2: Motif scanning confers specificity, as shown in the histograms of normalized motif match scores at sites where there is TF binding (white) and where there is no measured TF binding (gray). The histograms are constructed using the set of motif scanning scores across all five TFs (see Table S2 and Sec. S1.3).

FigureS3.eps

Figure S3: At LPS-inducible TF binding locations, HAc valley scores (which represent local minima within regions of high HAc; see Sec. S1.6) are more pronounced in LPS-stimulated cells than in unstimulated cells. (A) Histogram of HAc ChIP valley scores (VS) measured in unstimulated (white bars) and LPS-stimulated (gray bars) macrophages, at locations where there is LPS-dependent (inducible) IRF1 binding. The HAc valley scores from LPS-stimulated cells at these binding locations are clearly seen to be skewed towards higher scores, when compared to HAc valley scores measured in unstimulated cells. (B) Histogram of HAc ChIP signal intensity values from unstimulated (white bars) and LPS-stimulated (gray bars) macrophages, at locations where there is LPS-dependent (inducible) IRF1 binding. Although there is a slightly higher frequency at high HAc ChIP signal in stimulated cells at these locations, the effect is very slight. Thus, it appears that at LPS-dependent IRF1 binding sites, HAc valley scores are more LPS-responsive than the HAc ChIP-Seq signal intensity.

FigureS4.eps

Figure S4: Sensitivity vs. false positive rate (FPR) curves for the motif scanning-only reference model and all models with motif scanning plus one additional feature. (A) Complete sensitivity vs. FPR curve (receiver operating characteristic, or "ROC" curve), on linear scale. The FPR curve for the random model is slightly convex, due to the fact that a ground-truth binding site can span adjacent 100 bp intervals, and a predicted binding site in either interval would be counted as a correct positive prediction (see Sec. S1.11; and in the main article, see the Methods subsection "Performance Metric"). A zoomed-in view of the portions of these curves at low FPR (i.e., FPR $< 0.01$) is shown in Fig. 2 (main article). For reference purposes, the full area under the ROC curves for the random, motifs-only, and HAc ChIP VS (S) models are as follows: 0.579, 0.803, and 0.825, respectively. Please note that the full area under the ROC curve is not the measure of prediction performance used for comparing models in this study; instead, the partial area under the ROC curve (for FPR $< 0.01$) was used (see Sec. S1.11). (B) Partial ROC curve (for FPR $> 10^{-4}$) for the same models, with the horizontal (FPR) axis on log scale. Here, the significantly higher sensitivity of the model with HAc ChIP-Seq valley scores ["Motifs + HAc ChIP VS (S)"] vs. the motifs-only model, for FPR values in the range of $10^{-4}$ to $10^{-2}$, is apparent.

FigureS5.eps

Figure S5: Performance of models with the motif scanning feature, the HAc ChIP-Seq valley score feature (from stimulated cells), and one additional feature. The reference model (motifs only) and the best-performing two-feature model (motifs + HAc ChIP VS (S)) are shown for comparison purposes. Prediction performance is measured as the area under the sensitivity vs. false positive rate (FPR) curve, or "ROC" curve, for FPR $< 0.01$ (see Sec. S1.11).

Figure S6: Histogram of nucleosome occupancy scores at TF binding sites vs. sites where no TF binding was detected (for the five TFs assayed). While TF binding sites are slightly biased toward higher nucleosome occupancy scores, the histograms are essentially overlapping.

# References

[1] Ramsey SA, Klemm SL, Zak DE, Kennedy KA, Thorsson V, et al. (2008) Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. PLoS Comput Biol 4:e1000021. doi:10.1371/journal.pcbi.1000021.

[2] Aderem A, et al. (2009). Systems Approach to Immunity and Inflammation Project website. URL `http://www.systemsimmunology.org`.

[3] Stratowa C (2009). xps: Methods for processing and analysis of Affymetrix oligonucleotide arrays including exon arrays, whole genome arrays, and plate arrays. URL `http://www.bioconductor/org/packages/2.6/bioc/html/xps.html`.

[4] Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. J Comp Graph Stat 5:299–314.

[5] Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2008) Ensembl 2008. Nucleic Acids Res 36:D707–14. doi:10.1093/nar/gkm988.

[6] Smedley D, Haider S, Ballester B, Holland R, London D, et al. (2009) BioMart–biological queries made easy. BMC Genomics 10:22. doi:10.1186/1471-2164-10-22.

[7] Karolchik D, Hinrichs AS, Kent WJ (2009) The UCSC genome browser. Curr Protoc Bioinformatics Chapter 1:Unit1.4. doi:10.1002/0471250953.bi0104s28.

[8] Wingender E, Dietze P, Karas H, Knuppel R (1996) TRANSFAC: a database on transcription factors and their dna binding sites. Nucleic Acids Res 24:238–241.

[9] Lähdesmäki H, Rust AG, Shmulevich I (2008) Probabilistic inference of transcription factor binding from multiple data sources. PLoS ONE 3:e1820. doi:10.1371/journal.pone.0001820.

[10] Gilchrist M, Thorsson V, Li B, Rust AG, Korb M, et al. (2006) Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. Nature 441:173–178. doi:10.1038/nature04768.

[11] Litvak V, Ramsey SA, Rust AG, Zak DE, Kennedy KA, et al. (2009) Function of C/EBPdelta in a regulatory circuit that discriminates between transient and persistent TLR4-induced signals. Nat Immunol 10:437–443. doi:10.1038/ni.1721.

[12] Illumina (2008) Genome Analyzer Pipeline Software User Guide. Illumina, San Diego, CA, USA, v0.3 edition.

[13] Robertson AG, Bilenky M, Tam A, Zhao Y, Zeng T, et al. (2008) Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. Genome Res 18:1906–17. doi:10.1101/gr.078519.108.

[14] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657. doi:10.1038/nmeth1068.

[15] Nelson D, Perry ME, Chalkley R (1979) A correlation between nucleosome spacer region susceptibility to DNase I and histone acetylation. Nucleic Acids Res 6:561–74.

[16] Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23:137–44. doi:10.1038/nbt1053.

[17] Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, et al. (2007) 28-way vertebrate alignment and conservation track in the UCSC genome browser. Genome Res 17:1797–1808. doi:10.1101/gr.6761107.

[18] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–50. doi:10.1101/gr.3715005.

[19] Won KJ, Ren B, Wang W (2010) Genome-wide prediction of transcription factor binding sites using an integrated model. Genome Biol 11:R7. doi:10.1186/gb-2010-11-1-r7.

[20] Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, et al. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458:362–6. doi:10.1038/nature07667.

[21] Goh WS, Orlov Y, Li J, Clarke ND (2010) Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. PLoS Comput Biol 6:e1000649. doi:10.1371/journal.pcbi.1000649.

[22] Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 424:147–51. doi:10.1038/nature01763.

[23] Ovcharenko I, Loots GG, Giardine BM, Hou M, Ma J, et al. (2005) Mulan: multiple-sequence local alignment and visualization for studying function and evolution. Genome Res 15:184–94. doi:10.1101/gr.3007205.

[24] McClish DK (1989) Analyzing a portion of the ROC curve. Med Decis Making 9:190–5.

[25] Osada R, Zaslavsky E, Singh M (2004) Comparative analysis of methods for representing and searching for transcription factor binding sites. Bioinformatics 20:3516–25. doi:10.1093/bioinformatics/bth438.

[26] Li L, Liang Y, Bass RL (2007) GAPWM: a genetic algorithm method for optimizing a position weight matrix. Bioinformatics 23:1188–94. doi:10.1093/bioinformatics/btm080.

[27] Won KJ, Agarwal S, Shen L, Shoemaker R, Ren B, et al. (2009) An integrated approach to identifying *cis*-regulatory modules in the human genome. PLoS One 4:e5501. doi:10.1371/journal.pone.0005501.

[28] Huyer W, Neumaier A (2008) SNOBFIT – stable noisy optimization by branch and fit. ACM Trans Math Softw 35:1–25. doi:10.1145/1377612.1377613.

[29] Honda K, Taniguchi T (2006) IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. Nat Rev Immunol 6:644–658. doi:10.1038/nri1900.

[30] Hoffmann A, Baltimore D (2006) Circuitry of nuclear factor kappaB signaling. Immunol Rev 210:171–186. doi:10.1111/j.0105-2896.2006.00375.x.

[31] Whitington T, Perkins AC, Bailey TL (2009) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. Nucleic Acids Res 37:14–25. doi:10.1093/nar/gkn866.

[32] Saccani S, Pantano S, Natoli G (2001) Two waves of nuclear factor kappaB recruitment to target promoters. J Exp Med 193:1351–9.