Supplementary Material for: Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data.

Roger Pique-Regi[1,†,*], Jacob F. Degner [1,2,†,*], Athma A. Pai[1],
Daniel J. Gaffney[1,3], Yoav Gilad[1,*], Jonathan K. Pritchard[1,3,*]

[1]Department of Human Genetics, University of Chicago
[2]Committee on Genetics, Genomics and Systems Biology, University of Chicago
[3]Howard Hughes Medical Institute, University of Chicago
[†]These authors contributed equally.

[*]To whom correspondence should be addressed: rpique@uchicago.edu,
jdegner@uchicago.edu, gilad@uchicago.edu, pritch@uchicago.edu

Contents

1 Overview

CENTIPEDE applies a hierarchical Bayesian mixture model to infer regions of the genome that are bound by particular transcription factors. It starts by identifying a set of candidate binding sites (e.g., sites that match a certain position weight matrix (PWM)), and then aims to classify the sites according to whether each site is bound or not bound by a TF. CENTIPEDE is an unsupervised learning algorithm that discriminates between two different types of motif instances using as much relevant information as possible. In brief, the procedure is as follows:

1. Scan the genome for all approximate matches to a target PWM of interest. Each site that matches the PWM is considered a candidate binding site (Section 2.1).

2. Extract the relevant data for every candidate binding site. For this paper we extracted the following. Data for the prior (Section 2.2): PWM match score, conservation score, distance to the nearest transcription start site. Data for the likelihood (Section 2.3): DNase-seq and histone ChIP-seq data from windows of 200 bp and 400 bp, respectively, centered on each PWM match.

3. Fit a mixture model in which there are two kinds of sites (labeled "bound" and "unbound"). The two classes of sites are characterized by having (potentially) different distributions for each of the measured data types. The parameters to be estimated by the model are: (i) the distributions of each data type in the bound and unbound classes, separately, and (ii) the posterior probability that each site is in the bound, and unbound classes, respectively. We use an EM (expectation maximization) algorithm to find the values of (i) and (ii) that maximize the overall likelihood of the data across all sites (Section 3).

4. Report the set of candidate binding sites with high posterior probability (e.g., > 0.99) of being in the bound class.

5. Validation of the high posterior TF-bound sites using ChIP-seq (Section 4.1) or sequence conservation if ChIP-seq is not available (in this case, conservation is not included in the model; Section 4.3).

CENTIPEDE separates the available information into two components. We think of the genomic information (e.g., PWM match score) as telling us something about the propensity for a particular site to be actively bound, without directly depending on the tissue or cell type or experimental conditions. Based on this information we compute a prior probability that each site is bound (using a logistic function, the parameters of which are estimated by the mixture model).

In contrast to the genomic prior information, we treat experimental data (DNaseI and histone modifications) as being dependent on the particular cell type or experimental condition. Each data type is modeled with a mixture of two distributions with an independent set of parameters for the bound and unbound state. Although we focus on a single cell-type here, this framework extends naturally to the setting in which experimental data are available from multiple tissues, cell types or experimental conditions.

For unsupervised mixture models, the labeling of the clusters is sometimes arbitrary (i.e., in this context it might be unclear which cluster corresponds to "bound" sites). This problem was largely solved by forcing one class (the "unbound" state) to have a uniform distribution of DNase cuts across the region around the motif. In a handful of CENTIPEDE runs, label switching could be detected if, for example, "unbound" sites were more conserved, had more DNase-seq reads, or tended to be closer to the nearest TSS than "bound" sites. In practice, we minimized label switching by choosing an appropriate starting point (Section 3.2).

2 Data sources and processing

2.1 PWM scanning.

Candidate binding sites were identified using either pre-estimated position weight matrices (PWMs) or words that we determined to be enriched in hypersensitive sites (see Section 5.1). PWMs were defined based on TRANSFAC (TRANSFAC®Professional 2009.1 (2009-03-27)) and JASPAR (`http://jaspar.genereg.net/`, core set downloaded on 2009-09-01). All matrices from those two sources were considered for inclusion (see below for further discussion). We generated random PWMs by permuting the JASPAR matrices using an online tool available at `http://jaspar.genereg.net/`. We further eliminated any randomly generated PWMs that showed significant similarity to a known PWM using TomTom (Leaving 47 of 200 PWMs, [1]).

We scanned the human genome sequence (hg18) for matches to each PWM using our implementation of the following commonly used formula [2]:

$$\begin{aligned}\text{PWM score}\,(l) &= \sum_{w=0}^{W-1} \log_2\left(\frac{P\,(\text{seeing } S_{l+w} \text{ at position } w|\text{PWM})}{P\,(\text{seeing } S_{l+w}|\text{ background model})}\right) \qquad (1)\\ &= \sum_{w=0}^{W-1} \log_2\left(p\left[S_{l+w}, w\right]\right) - \sum_{w=0}^{W-1} \log_2\left(0.25\right)\\ &= \sum_{w=0}^{W-1} \log_2\left(p\left[S_{l+w}, w\right]\right) - \log_2\left(0.25\right) W\end{aligned}$$

where PWM score(l) denotes the PWM score for a specified PWM of length W in the sequence between positions l and $l + W - 1$ in a DNA sequence, where S_{l+w} denotes the nucleotide observed at position $l + w$, where the PWM model is given by the probability $p\left[S_{l+w}, w\right]$ of observing the nucleotide S_{l+w} (A,C,G,T) at position w (W is the motif length), and where it is assumed that the probability of observing each nucleotide on the background distribution is 0.25. The probabilities of the PWM model were taken directly from the database, or estimated from the base counts at each position. In the latter case, we estimated the probabilities by adding 0.05 pseudo-counts to avoid having a very strong score against a sequence for which binding affinity may have not been tested (this is especially helpful for PWMs obtained using electrophoretic mobility shift assays (EMSA), which often have small sample sizes).

We used a relatively permissive threshold for defining candidate binding sites, considering all sites for which PWM Score(l) $> log_2(10000) = 13.288$). The median number of matches for a PWM was $\sim$ 96,000. We filtered out the matrices that were unusually abundant ($> 500,000$ matches in the genome), or that were not found (i.e., 0 matches). Using these criteria and combining both databases (TRANSFAC and JASPAR), we obtained 756 position weight matrices (PWMs), out of the original ~ 1000 PWM matrices from both datasets.

In many cases there is similarity between multiple PWMs; this occurs for two major reasons. First, a single TF may be represented by two or more similar PWMs (e.g., NRSF also known as REST is represented by M00256 in TRANSFAC, and MA0138 in JASPAR). Second, closely related TFs may have similar motifs (e.g. E-box, GC-box, GATA, and CAAT). Since it is difficult to know *a priori* which motifs are best, we fit the model first using all motifs. Motifs that do not correspond closely to any active TF are filtered out by our conservation criterion (Section 4.3). Additionally, our map reports ambiguity in cases where a single location is estimated to be bound by PWMs corresponding to multiple factors (Section 5.3).

2.2 Data for the prior.

For each PWM or word and for each motif match in the genome (as defined in the previous section), we obtained the following genomic data for use in the prior (see also Table S1):

- PWM score(l), calculated based on eq. 1.
- The average PhastCons conservation score on the motif match region across the plancental mammals on the 44-way multiple alignment (`ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/phastCons44way/placentalMammals/`)
- Distance to the closest annotated Ensembl transcription start site (TSS) (`ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/ensGene.sql`). The following distance formula was used to calculate the TSS Proximity$_l$ score, where TSS Distance$_l$ is the distance in base pairs from the motif center to the nearest TSS:

$$\text{TSS Proximity}_l = \left(1 + \frac{\text{TSS Distance}_l}{1000}\right)^{-1} \tag{2}$$

Table S1: Summary of data sources used as prior information.

Data type	Source	Publications
PWM matrices	TRANSFAC	[3]
PWM matrices	JASPAR	[4]
PhastCons conservation score	UCSC browser	[5]
Ensembl gene annotion	UCSC browser	[6]

2.3 Experimental data.

Table S2 summarizes the experimental information used in this paper. The ENCODE Project data were downloaded (on 2010-07-01) from the main ENCODE data distribution center at the University of California Santa Cruz (UCSC), publicly available at `ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/`. We thank the ENCODE Project for making these data available pre-publication (see release dates on Table S2).

We generated additional DNase-seq data for the GM18507 and GM18507 HapMap LCLs (8 lanes each on the Illumina GAII sequencer) using the same protocol as for the ENCODE GM12878 sample (`http://genome.ucsc.edu/ENCODE/protocols/general/Duke_DNase_protocol.pdf`). We thank Greg Crawford for assistance in implementing the DNase-seq protocol.

Table S2: Summary of experimental data used by CENTIPEDE.

Data type	Cell-line	Source	Lab	Publications	Embargo end date	# mapped reads
DNase-seq	GM12878	EncodeDCC	Crawford	[7]	2009-12-20	62.3M
DNase-seq	GM18507	GEO	Pritchard	This study	NA	77.6M
DNase-seq	GM19239	GEO	Pritchard	This study	NA	76.3M
H3K4me1	GM12878	EncodeDCC	Bernstein	NA	2009-10-05	20.5M
H3K4me2	GM12878	EncodeDCC	Bernstein	NA	2009-10-05	14.0M
H3K4me3	GM12878	EncodeDCC	Bernstein	NA	2009-10-05	13.5M
H3K9ac	GM12878	EncodeDCC	Bernstein	NA	2009-10-05	16.1M
H3K27ac	GM12878	EncodeDCC	Bernstein	NA	2009-10-05	13.9M
H3K27me3	GM12878	EncodeDCC	Bernstein	NA	2009-10-05	18.1M
H4k20me1	GM12878	EncodeDCC	Bernstein	NA	2009-10-05	16.9M

Read mapping. We used the raw sequencing reads ("*.fastq.gz" files) and mapped them on the human reference genome (hg18) using BWA [8] (using the default settings). Then we filtered out the reads with quality score < 10 or that mapped to more than one location in the genome. For reads obtained using the DNase-seq protocol, since their length is limited to 20bp, we further removed all reads that contained an insertion/deletion or that had more than one mismatch with respect to hg18. The aligned reads of multiple experimental replicates on the same cell-line type were combined into a single file. The total numbers of reads that were successfully mapped for each data type are reported in Tables S2 and S3.

2.4 Models for experimental data.

Based on preliminary analysis, we settled on the following models for the experimental data. For the DNase data, we observed that around bound locations there is typically both an overall increase in the number of reads, as well as a distinctive profile of cut-site locations (that often differs by strand) [9, 10, 11] (see Figure S8 for examples). We model the total number of DNase reads in a 200bp window around the motif site as coming from a negative binomial distribution (this distribution is like a Poisson with variability in the underlying rate). Then, conditional on the total number of reads in the region, the number of reads at each site is modeled as multinomial. In the model, we estimate different parameters for the negative binomial for the bound and unbound classes. Based on preliminary experiments we fix the parameters for the multinomial distribution to be uniform in the unbound case, and we estimate a distinct parameter for every position in the bound case. (The full model is quite parameter-rich, but since the data sets are quite large, we observed no performance cost.)

For the histone modification data, we found that the data were most informative if we counted the total numbers of reads corresponding to each modification in a 400 bp window around each candidate motif site. There is potentially some additional spatial information (i.e., the distributions of reads is not quite uniform around the binding sites) but including the spatial information has little effect on model performance so we dropped this from the final model. Initially we included separate parameters for each histone modification, but this proved to be confusing to the mixture model since there are strong correlations among some of the modifications. So for the analyses presented, we merged the modifications into two groups: marks that are associated with active (H3K4me1, H3K4me2, H3K4me3, H3K9ac and H3K27ac), and inactive (H3K27me1 and H4K20me1) chromatin, respectively [12]. The numbers of reads in each class were modeled as being negative binomially distributed, with distinct parameters for the bound and unbound cases.

3 The CENTIPEDE model

3.1 The model.

CENTIPEDE uses a probabilistic framework known as a hierarchical mixture model

$$P\left(D_l\right) = P\left(Z_l = 1 \middle| G_l\right) P\left(D_l | Z_l = 1\right) + P(Z_l = 0|G_l) P\left(D_l | Z_l = 0\right) \tag{3}$$

where, for each motif match 'l', D_l and G_l represent the observed experimental data and the prior information around the motif match. The data D_l can come from two underlying distributions that form the mixture model. One distribution corresponds to the bound state of transcription factors ($Z_l = 1$) while the other distribution corresponds to the unbound state ($Z_l = 0$).

The model is hierarchical because we model the prior probabilities of observing a bound motif given prior information G_l, such as conservation, distance to the TSS, score of the PWM. For each potential binding location 'l' we calculate a prior probability $\pi_l = P\left(Z_l = 1 \middle| G_l\right)$ that the site is used. This prior probability depends on various genomic information that can describe site 'l' and is modeled using a logistic model

$$log\left(\frac{\pi_l}{1-\pi_l}\right) = \beta_0 + \beta_1 \times \text{PWM Score}_l + \beta_2 \times \text{Cons. Score}_l + \beta_3 \times \text{TSS Proximity}_l + \dots \tag{4}$$

As experimental data "D_l", CENTIPEDE can combine many different types of experiments (e.g. read counts from DNase-seq, histone modification ChIP-seq). For a single experimental data-type (e.g., DNase-seq), the collection of reads in a given region (200bp) around the motif matches 'l' can be represented by an $L \times S$ matrix $\mathbf{X} = \{X_{ls}\}$. Each row $X_{l,\cdot} = (X_{l,1}, \dots, X_{l,S})$ corresponds to a single location in the genome with a motif match 'l' and the column index 's' indexes each position relative to the center of this motif match. Thus, the cells of this matrix will represent discrete counts of DNase-seq reads that map to a given distance from the motif center.

For example, for DNase-seq (window size 200bp) and a motif of length 19, $S = (200 + 19)2 = 438$, the first quarter of $X_{l,\cdot}$ contains the reads collected to the left of the motif mid-point on the forward strand, the second quarter the reads collected to the right of the motif mid-point on the forward strand, and the second half correspond to the reads collected on the reverse strand. This matrix can be visualized as an image (see Figure 1A - DNaseI cuts/site - where both strands have been combined $X_{l,1:S/2} + X_{l,(S/2+1):S}$).

For a single data type, we can write the problem for a particular motif instance (l) as

$$\begin{aligned} P\left(X_{l,\cdot}\right) &= \pi_l P\left(X_{l,\cdot}|Z_l=1\right)+ \\ &\quad (1-\pi_l)\, P\left(X_{l,\cdot}|Z_l=0\right) \end{aligned} \tag{5}$$

where here, we have simply re-written the prior probabilities of coming from the bound and unbound distributions as $\pi_l = P\left(Z_l = 1|\, G_l\right)$ and $P\left(Z_l = 0|\, G_l\right) = 1 - \pi_l$, respectively. If we have multiple experimental data-types this is modeled as

$$\begin{aligned} P\left(X_{l,\cdot}^{(1)}, X_{l,\cdot}^{(2)}, X_{l,\cdot}^{(3)}\right) &= \pi_l P\left(X_{l,\cdot}^{(1)}, X_{l,\cdot}^{(2)}, X_{l,\cdot}^{(3)}|Z_l=1\right)+ \\ &\quad (1-\pi_l)\, P\left(X_{l,\cdot}^{(1)}, X_{l,\cdot}^{(2)}, X_{l,\cdot}^{(3)}|Z_l=0\right) \end{aligned} \tag{6}$$

where $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, and $\mathbf{X}^{(3)}$ represent three different discrete-count data-types ($\mathbf{X}^{(k)}$ in general, for the k-th data type). Our model makes the simplifying assumption that, conditional on the state of the element (i.e. bound/unbound), the data types are independent: i.e.,

$$P\left(X_{l,\cdot}^{(1)}, X_{l,\cdot}^{(2)}, X_{l,\cdot}^{(3)}|Z_l\right) = P\left(X_{l,\cdot}^{(1)}|Z_l\right) P\left(X_{l,\cdot}^{(2)}|Z_l\right) P\left(X_{l,\cdot}^{(3)}|Z_l\right) \tag{7}$$

Violations of this assumption may reduce the performance of the method, which is why we merged the seven histone marks into two groups for the analysis.

Now we will describe how these conditional distributions are modeled. Each data-type will have the same distributional form but an independent set of parameters, so we will focus on a single data-type $P\left(X_{l,\cdot}|Z_l\right)$ (for additional data-types we would add the appropriate super-indices in all the notation that follows). We need to define the two key distributions $P\left(X_{l,\cdot}|Z_l=1\right)$ and $P\left(X_{l,\cdot}|Z_l=0\right)$ that distinguish bound instances from those not bound. First we define the following useful variable

$$R_l = \sum_{s=1}^{S} X_{l,\cdot} \tag{8}$$

to denote the total number of reads in the region surrounding the l-th site. We can then write

$$\begin{aligned} P\left(X_{l,\cdot}|Z_l\right) &= P\left(X_{l,\cdot}, R_l|Z_l\right) = \\ &= P\left(X_{l,\cdot}|R_l, Z_l\right) P\left(R_l|Z_l\right) \end{aligned} \tag{9}$$

Now we model the first and second term of eq. 9 for the bound case $Z_l = 1$ and unbound case $Z_l = 0$. The total number of reads is modeled with a negative binomial distribution,

$$\begin{aligned} P\left(R_l|Z_l=1\right) &= NegativeBinomial\left(R_l|\alpha_1, \tau_1\right) = \\ &= \frac{\Gamma\left(\alpha_1 + R_l\right)}{R_l!\Gamma\left(\alpha_1\right)} {\tau_1}^{\alpha_1} \left(1-\tau_1\right)^{R_l} \end{aligned} \tag{10}$$

$$\begin{aligned} P\left(R_l|Z_l=0\right) &= NegativeBinomial\left(R_l|\alpha_0, \tau_0\right) = \\ &= \frac{\Gamma\left(\alpha_0 + R_l\right)}{R_l!\Gamma\left(\alpha_0\right)} {\tau_0}^{\alpha_0} \left(1-\tau_0\right)^{R_l} \end{aligned} \tag{11}$$

which depend on α_1, τ_1 for the bound class, and α_0, τ_0 for the unbound class. With these two distributions we can capture open versus closed chromatin in DNaseI hypersensitivity assays, or enrichment of certain histone modifications associated with enhancers or repressors measured by ChIP-seq assays. If the positional distribution $P(X_{l,\cdot}|R_l, Z_l)$ is not important (or not very informative) we are able leave it un-specified (i.e, any configuration is equally likely)

$$\begin{aligned} P(X_{l,\cdot}|R_l, Z_l) &= C && (12)\\ P(X_{l,\cdot}|Z_l) &\propto P(R_l|Z_l) && (13)\end{aligned}$$

This is the option we chose for the histone modification ChIP-seq assays. In contrast, we found that for DNaseI the positional information can be very informative as DNaseI leaves a distinctive cleavage pattern when $Z_l = 1$. The spatial distribution of reads surrounding the binding site is modeled with a multinomial distribution

$$\begin{aligned} P(X_{l,\cdot}|Z_l = 1, R_l) &= Multinomial\,(X_{l,\cdot}|R_l, \{\lambda_1, \ldots, \lambda_S\}) \\ &= R_l! \prod_{s=1}^{S} \left(\frac{{\lambda_s}^{X_{l,s}}}{X_{l,s}!} \right) && (14)\end{aligned}$$

where the λ_s gives the probability that a read is obtained from position index 's' and $R_l \lambda_s$ is the expected value of $X_{l,s}$ given R_l. For $Z_l = 0$, the TF is not bound, so no specific footprint is expected. In this case we found it worked well to simply set the cut-site distribution as uniform ($\lambda_s = 1/S$):

$$\begin{aligned} P(X_{l,\cdot}|Z_l = 0, R_l) &= Multinomial\,(X_{l,\cdot}|R_l, \{1/S, \ldots, 1/S\}) \\ &= R_l! \prod_{s=1}^{S} \left(\frac{S^{-X_{l,s}}}{X_{l,s}!} \right) && (15)\end{aligned}$$

The parameters of the CENTIPEDE model ($\beta_1, \beta_2, \ldots; \alpha_0, \tau_0,\ \alpha_1, \tau_1, \lambda_1, \ldots, \lambda_S; \alpha'_0, \tau'_0,\ \alpha'_1, \tau'_1, \ldots$) are estimated by maximizing the likelihood function using an expectation maximization (EM) algorithm. To implement this, instead of maximizing the likelihood function

$$L(\text{Parameters}) = \prod_l P(D_l|\,\text{Parameters}) \qquad (16)$$

directly, we formulate a "complete" likelihood function:

$$\begin{aligned} L_C(\text{Param.}) &= \prod_{l=1}^{L} P(D_l, Z_l|\,\text{Parameters}) = \\ &= \prod_{l}^{L} \left(\frac{\Gamma(\alpha_1 + R_l)}{\Gamma(\alpha_1)} \tau_1^{\alpha_1} (1 - \tau_1)^{R_l} \prod_{s=1}^{S} \left(\frac{\lambda_s^{x_{s,l}}}{x_{s,l}!} \right) \right)^{Z_l} \\ &\quad \times \left(\frac{\Gamma(\alpha_0 + R_l)}{\Gamma(\alpha_0)} \tau_0^{\alpha_0} (1 - \tau_0)^{R_l} \prod_{s=1}^{S} \left(\frac{(1/S)^{x_{s,l}}}{x_{s,l}!} \right) \right)^{(1-Z_l)} \\ &\quad \prod_{l}^{L} \left(\pi_l^{Z_l} (1 - \pi_l)^{1-Z_l} \right) && (17)\end{aligned}$$

and we iteratively maximize $Q\left(\text{Param.}\right) = E_Z\left[L_C\left(\text{Param.}\right)|D\right]$. This maximization involves cycling through two phases. In the first phase (E-step), we estimate the expectations of each individual Z_l keeping all other parameters fixed. In the second phase (M-step), we treat the expectations of Z_l as fixed, and we find the maximum likelihood estimates of all other parameters.

For a given cycle of the EM algorithm, on the E-step, we calculate $E\left(Z_l\right)$ (also equal to the posterior probability p_l of binding for that site given the data and the current estimates of the model parameters):

$$\begin{aligned} p_l &= P\left(Z_l = 1\,|R_l\ , X_{l,\cdot}, \text{Param.}\right) = \qquad (18)\\ &= \frac{P\left(R_l, X_{l,\cdot}|Z_l = 1\right)\pi_l}{P\left(R_l, X_{l,\cdot}|Z_l = 1\right)\pi_l + P\left(R_l, X_{l,\cdot}|Z_l = 0\right)\left(1-\pi_l\right)} = \\ &= \frac{1}{1 + \frac{P(R_l, X_{l,\cdot}|Z_l=0)(1-\pi_l)}{P(R_l, X_{l,\cdot}|Z_l=1)\pi_l}} = \\ &= \frac{1}{1 + \left(\frac{1-\pi_l}{\pi_l}\right)\left(\frac{(1-\tau_0)^{R_l}\tau_0^{\alpha_0}\Gamma(R_l+\alpha_0)/\Gamma(\alpha_0)}{(1-\tau_1)^{R_l}\tau_1^{\alpha_1}\Gamma(R_l+\alpha_1)/\Gamma(\alpha_1)}\right)\left(\prod_{s=1}^{S}\left(\frac{1/S}{\lambda_s}\right)^{x_{s,l}}\right)} \qquad (19) \end{aligned}$$

The form of this posterior probablity p_l can be more easily interpreted in terms of the posterior odds:

$$\frac{p_l}{1-p_l} = \left(\frac{\pi_l}{1-\pi_l}\right)\left(\frac{\left(1-\tau_1\right)^{R_l}\tau_1^{\alpha_1}\Gamma\left(R_l+\alpha_1\right)/\Gamma\left(\alpha_1\right)}{\left(1-\tau_0\right)^{R_l}\tau_0^{\alpha_0}\Gamma\left(R_l+\alpha_0\right)/\Gamma\left(\alpha_0\right)}\right)\left(\prod_{s=1}^{S}\left(S\lambda_s\right)^{x_{s,l}}\right) \qquad (20)$$

illustrating that the posterior odds are equal to the product of the prior odds (given by the logistic model) and the likelihood ratios (LRs) for the models corresponding to each type of data (multinomial negative binomial for the DNase-seq). This easily extends to multiple independent types of experimental data:

$$\begin{aligned} &\text{Posterior Odds:} & \frac{p_l}{1-p_l} &= \\ &\text{Prior Odds:} & &= \left(\frac{\pi_l}{1-\pi_l}\right) \\ &\text{LR DNase-seq:} & &\times \left(\frac{\left(1-\tau_1^{(1)}\right)^{R_l^{(1)}}(\tau_1^{(1)})^{\alpha_1^{(1)}}\Gamma\left(R_l^{(1)}+\alpha_1^{(1)}\right)/\Gamma\left(\alpha_1^{(1)}\right)}{\left(1-\tau_0^{(1)}\right)^{R_l^{(1)}}(\tau_0^{(1)})^{\alpha_0^{(1)}}\Gamma\left(R_l^{(1)}+\alpha_0^{(1)}\right)/\Gamma\left(\alpha_0^{(1)}\right)}\right)\left(\prod_{s=1}^{S}\left(S\lambda_s\right)^{x_{s,l}^{(1)}}\right) \\ &\text{LR Act. Histones:} & &\times \left(\frac{\left(1-\tau_1^{(2)}\right)^{R_l^{(2)}}(\tau_1^{(2)})^{\alpha_1^{(2)}}\Gamma\left(R_l^{(2)}+\alpha_1^{(2)}\right)/\Gamma\left(\alpha_1^{(2)}\right)}{\left(1-\tau_0^{(2)}\right)^{R_l^{(2)}}(\tau_0^{(2)})^{\alpha_0^{(2)}}\Gamma\left(R_l^{(2)}+\alpha_0^{(2)}\right)/\Gamma\left(\alpha_0^{(2)}\right)}\right) \\ &\text{LR Rep. Histones:} & &\times \left(\frac{\left(1-\tau_1^{(3)}\right)^{R_l^{(3)}}(\tau_1^{(3)})^{\alpha_1^{(3)}}\Gamma\left(R_l^{(3)}+\alpha_1^{(3)}\right)/\Gamma\left(\alpha_1^{(3)}\right)}{\left(1-\tau_0^{(3)}\right)^{R_l^{(3)}}(\tau_0^{(3)})^{\alpha_0^{(3)}}\Gamma\left(R_l^{(3)}+\alpha_0^{(3)}\right)/\Gamma\left(\alpha_0^{(3)}\right)}\right) \end{aligned}$$

In the maximization step we update each of the parameters according to the estimators that maximize the EM Q function; in this case equivalent to the complete likelihood function with the Z_l replaced by their expected values

p_l. For each of the multinomial parameters, $\lambda_{s=.}$, the estimators are:

$$\lambda_s = \frac{\sum_l^L p_l x_{s,l}}{\sum_{s=1}^S \sum_l^L p_l x_{s,l}} \tag{21}$$

The estimates of the negative binomial parameters τ_0 and τ_1 are:

$$\tau_0 = \frac{\alpha_0 \left(L - \sum_l^L p_l\right)}{\sum_l^L R_l + \alpha_0 \left(L - \sum_l^L p_l\right) - \sum_l^L R_l p_l} \tag{22}$$

$$\tau_1 = \frac{\alpha_1 \sum_l^L p_l}{\alpha_1 \sum_l^L p_l + \sum_l^L R_l p_l}, \tag{23}$$

respectively. Finally, in the case of the β's in the logistic model and in the case of the negative binomial parameters α_0, and α_1, there are no analytic solutions that maximize Q, but the derivatives are known so we can maximize them with a Newton-Raphson (BFGS) step using the R function "optim".

3.2 Implementation details.

This section describes several issues that have been considered in the implementation of the CENTIPEDE algorithm: i) mappability, ii) model robustness, iii) estimator shrinkage, and iv) EM initialization and convergence.

Mappability. Filtering the reads that map to multiple locations means that some repeat-rich regions of the genome contain gaps in the experimental data. In order to avoid problems with mappability that could bias our results, we filtered out motif instances that had more than 20% un-mappable bases in the 200bp window centered on the motif (mappability determined with simulation scripts modified from [13]). Briefly, we simulated reads for every 20 base pair segment of the genome, and determined whether our mapping scheme would reliably map a read originating from that location. These motif instances were discarded for all the analyses that are presented throughout this report. We also determined if the remaining instances with a few un-mappable bases could pose a problem to the CENTIPEDE model. In general, the model behaved well even if those missing locations were treated as having zero reads. We also explored an extended version of the model that can handle missing data by imputation, but it only showed a mild improvement at a much higher computational cost (results not shown). In fact, we saw a greater improvement in performance by "shrinking" the estimates for the multinomial profile (see below).

Model robustness. Several filters were used to discard sequencing reads that had low quality or that mapped to multiple locations along the genome (see Section 2.3). In general we find that the mixture of two negative binomials is very flexible in discriminating between two classes of sites, and provides a very good fit to the empirical distribution for the number of reads observed across all 200bp windows in the genome (Figure S7). However, we find that a few locations have extremely large numbers of reads at a single base location. For this reason, we decided to clip the reads that map to a single base in the genome to 30 reads at maximum. This happens very rarely, less than 10 times per 100,000 motif matches. Additionally, all the multiplications, factorials, Gamma functions, and parameters involving large numbers were computed as log-transformed which avoided potential numerical underflows/overflows.

Estimator shrinkage. In the statistics field, shrinkage estimators are often used to improve the acuracy of parameter estimators and avoid overfitting. We tried an implicit shrinkage by replacing the multinomial distribution by a multivariate Pólya distribution (also known as Dirichlet compound multinomial distribution) but explicit shrinkage of the λ parameters

$$\tilde{\lambda}_s = 0.5\lambda_s + 0.5\frac{1}{S} \tag{24}$$

achieved similar or better results at a much smaller fraction of the computational cost. When we compared the model predictions to ChIP-seq, shrinkage of the λ parameters helped to improve the overall performance. However, the improvement was mostly in terms of sensitivity while specificity usually decreased. For this reason, estimator shrinkage was not used in producing the main regulatory map where a low false positive rate (specificity) was preferered over a low false negative rate (sensitivity).

EM initialization and convergence. In general we found that the EM algorithm used in CENTIPEDE was very robust with respect to the starting point for the parameter estimates. Nevertheless, faster convergence was achieved if we used an informed starting point. In the implementation used here, we started the model by initializing the parameters as follows:

1. Fit the β's in the logistic model by replacing the p_l's (still not known) by a binary variable indicating the motif instances that are above the 90th%-tile on the DNase-seq distribution (for the histone-only model the 90th%-tile on the activating histone distribution was used instead).

2. Calculate the expected p_l using only the prior model using the β's calculated in the previous step.

3. Maximize all the parameters as in a regular EM algorithm iteration. All the parameters are updated using this initial estimate for p_l.

4. Regular EM algorithm until convergence.

Using this strategy the EM algorithm typically converges in less that 30 cycles. The criteria for convergence are: i) the likelihood of the model changes by less than 0.02, and ii) all the parameters change by less than 0.001 with respect to the previous EM iteration.

4 Validation of the CENTIPEDE predicted TF-bound sites

4.1 Validation with ChIP-seq.

We downloaded publicly available ChIP-seq data from the ENCODE project corresponding to six transcription factors (Table S3), including the raw reads and also the ChIP-seq peaks that ENCODE identified. In summary, the prediction performance of the methods developed here was assessed using: i) a correlation based approach, and ii)

Receiver Operation Curves (ROCs). The correlation approach considers that a method has a higher accuracy if the correlation between the predicted values and the ChIP-seq signal is large and the correlation with the background noise "control" is small. The ROC approach assumes that a "gold standard" exists with which we can confidently tell if the binary outcome of our prediction (TF bound versus not-bound) is true for a subset of the candidate motif sites.

Table S3: Summary of ChIP-seq validation data used in the main paper.

Antibody	Cell-line	Source	Lab	Embargo end date	# mapped reads
NRSF (REST)	GM12878	EncodeDCC	Myers	2009-08-25	22.7 M
GABPA	GM12878	EncodeDCC	Myers	2009-08-20	22.0 M
SRF	GM12878	EncodeDCC	Myers	2009-08-20	24.5 M
MAX	GM12878	EncodeDCC	Snyder	2009-10-09	11.2 M
JUND	GM12878	EncodeDCC	Snyder	2009-11-27	7.6 M
CTCF	GM12878	EncodeDCC	Bernstein	2009-10-05	15.4 M

Definition of ChIP-seq peaks, and motif instances as ChIP-seq positives. For NRSF (also known as REST), SRF, and GABPA we used the ChIP-seq peaks as reported by ENCODE at the UCSC genome browser portal. If multiple replicates were available we combined them into a single file. For three other factors (CTCF, JUND and MAX) ChIP-seq peaks were re-extracted using MACS [14]. This was necessary because peaks reported by the Myers lab ENCODE group were shorter peaks (100bp) compared to those extracted by other ENCODE groups and we wanted to have the same peak calling criteria across all TF. We combined all ChIP-seq and control replicates into two separate files respectively, and we used MACS version 1.3.7.1 with the default settings to detect peaks. We selected all the peaks with the lowest p-value resulting in a FDR of 0.02% for CTCF, 20% for JUND, and 0.1% for MAX. These differences in FDR are necessary due to the different amounts of coverage that are obtained in each ChIP-seq experiment. We also shortened the peaks we called with MACS to a 100bp window centered on the peak summit in order to obtain the same peak length distribution as for NRSF, SRF, and GABPA reported by ENCODE. Then, for each transcription factor all motif instances that overlapped these peaks were considered as ChIP-seq positives.

Definition of motif instances as ChIP-seq negatives. MACS does not report the peaks below a p-value cut-off so we need a procedure to extract motif instances with strong evidence that they are not bound by the factor. We defined the set of ChIP-seq negatives as all motif instances that did not overlap a ChIP-seq peak (i.e., did not have a p-value< 0.001) and had a fraction of total mapped reads on the control $\geq$ than that for the ChIP-seq treatment (see below for how these reads were extracted). This enrichment for control reads in the ChIP-seq negative set ensures that the evaluated prediction methods are not particularly biased towards the ChIP-seq control.

ROC analysis. After determining motif instances that are ChIP-seq positives / negatives, we calculated the performance of CENTIPEDE in predicting them. For a given cut-off of the posterior probability we define:

- True positives (TP) as the ChIP-seq positive instances that have been predicted correctly as bound.
- False positives (FP) as the ChIP-seq negative instances that have been predicted incorrectly as bound.
- False negatives (FN) as the ChIP-seq positive instances that have been predicted incorrectly as not bound.
- True negatives (TN) as the ChIP-seq negative instances that have been predicted correctly as not bound.

The sensitivity is defined as the fraction of ChIP-seq positive motifs that have been predicted correctly as bound (Sens.= TP / (TP+FN)). The false positive rate (FPR) is defined as the fraction of ChIP-seq negative motifs that have been predicted incorrectly as bound (FPR = FP / (FP+TN)). The FPR should not to be confused with the false discovery rate (FDR) defined as the fraction of motif instances that have been predicted incorrectly (FDR = FP / (TP+FP)). The receiver operator characteristic (ROC) is the curve that describes the trade-off between sensitivity and FPR that can be achieved by changing the threshold cut-off on the posterior probability. The larger the area under the curve (AUC), the better is the prediction accuracy. In Table S4 we also report the precision defined as fraction of all predicted sites that are also in ChIP-seq peaks (Prec.= TP/(TP+FP)).

Correlation analysis. In contrast to ROC analysis that requires a binary partition into bound/not-bound using ChIP-seq as a "gold standard", a correlation analysis does not require us to discard motif-cases that cannot be reliable determined (with a low FDR) by ChIP-seq. Furthermore, it is also becoming increasingly evident that TF binding is not perfectly dichotomous (bound/not-bound), but in many cases there would be partial binding corresponding to the fraction of cells that have TF binding at a particular site at any given time. We used Pearson correlation to measure the trend between the square root of the total number of ChIPseq-reads around a motif site and the posterior log-odds reported by CENTIPEDE.

The ChIPseq reads and "Control" reads around motif sites were extracted using the following procedure. We first mapped all the reads corresponding to all replicate samples for each TF and its respective "Control" experiment. BWA[8] mapping software was used and we filtered all the reads that did not map uniquely or had low quality scores. Then, for each motif instance along the genome:

1. Extract reads that fall within a 400bp window centered on the motif instance.
2. Calculate a ChIP-seq peak profile based on the motif instances with the top 20% number of reads (Figure S1).
3. Obtain the window that covers 75% of the peak mass using reverse-waterfilling (shaded area on Figure S1).
4. Calculate for each instance the total number of ChIP-seq reads and "Control" reads that fall inside of the window around the motif we calculated on the previous step (shaded area on Figure S1).

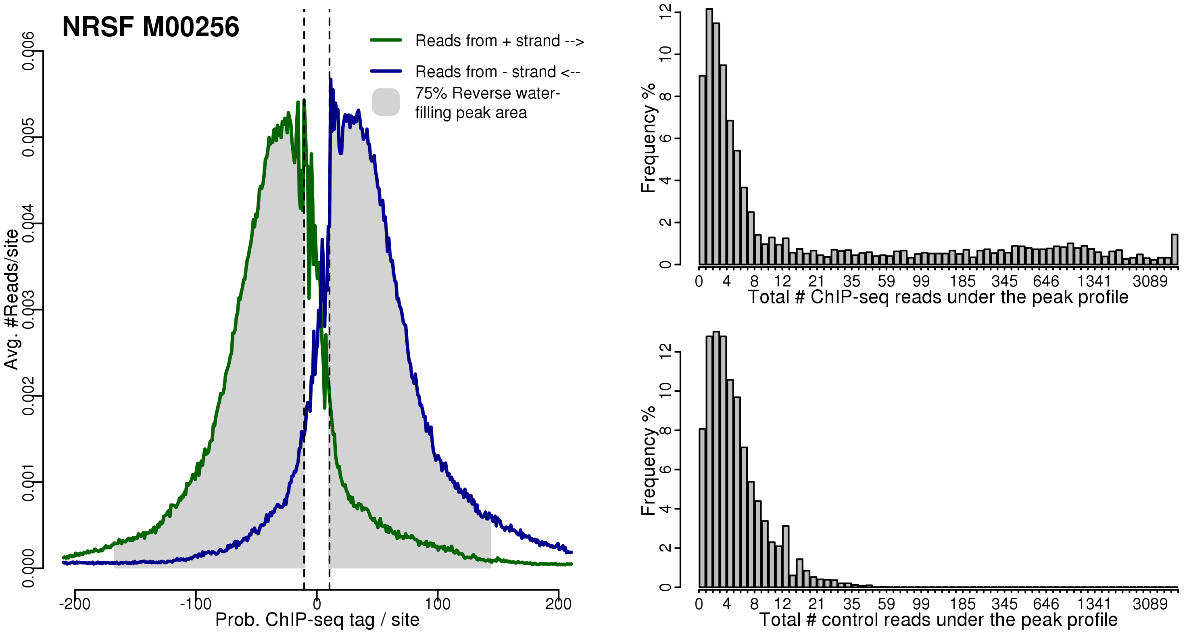


Figure S1: **Extraction of the ChIP-seq and Control signals**. This figure illustrates the procedure used for extracting ChIP-seq reads and "Control" reads around motif instances using NRSF (also known as REST) as an example. **Left panel**, corresponds to the ChIP-seq peak profile based on the motif instances with the top 20% number of reads, and the shaded gray area corresponds to the window that covers the 75% of the peak mass. **Right panel**, corresponds to the histogram of the number of reads falling inside the ChIP-seq profile for both ChIP-treatment (upper plot) and "Control" (bottom plot).

Table S4: **Summary of ChIP-seq calls and validation**.

			ChIP-seq		**CENTIPEDE with DNaseI footprint**						**CENTIPEDE w/ footprint & Histones**					
		# Motif	#Calls		Correlation w/			At 1% FPR			Correlation w/			At 1% FPR		
TF	Motif	instances	Negative,	Positive	ChIP-seq,	Control	AUC %	# Pred.,	%Sens.,	%Prec.	ChIP-seq,	Control	AUC %	# Pred.,	%Sens.,	%Prec.
In LCLs:																
NRSF (REST)	M00256	4966	1642	1193	0.78	0.32	95.17	569	46.27	97.01	0.5	0.49	88.47	267	16.51	73.78
SRF	MA0083	20841	11146	133	0.72	0.26	98.97	235	92.48	52.34	0.57	0.29	98.67	245	84.96	46.12
MAX	M00118	28296	13611	72	0.56	0.37	99.53	202	88.89	31.68	0.63	0.42	99.46	202	88.89	31.68
GABPA	M00108	36726	19150	924	0.72	0.56	99.52	1070	95.02	82.06	0.69	0.61	98.12	1579	98.16	57.44
CTCF	M01200	63748	23201	17093	0.79	0.1	97.53	15200	87.56	98.47	0.73	0.1	96.37	9541	52.62	94.27
JUND	M00199	107939	72172	14	0.34	0.34	97.91	733	78.57	1.5	0.34	0.38	98.06	2560	92.86	0.51
In K562:																
CTCF	M01200	63748	14664	10426	0.75	0.19	98.81	7093	66.62	97.93	0.57	0.19	97.62	3882	35.16	94.44
GABPA	M00108	36726	16559	1170	0.76	0.68	98.75	1246	92.31	86.68	0.69	0.66	98.33	1304	89.83	80.6
FOS (C-fos)	M00038	50416	19729	1417	0.6	0.2	97.04	1300	77.77	84.77	0.56	0.2	98.92	1383	83.63	85.68
JUN (C-jun)	M00036	7134	2938	265	0.68	0.34	93.91	219	71.32	86.3	0.7	0.34	98.01	234	76.98	87.18
NFE2	M00037	82927	47180	563	0.52	0.16	99.14	952	85.26	50.42	0.44	0.14	98.83	860	68.92	45.12
E2F4	M00739	12594	5150	418	0.77	0.28	99.94	464	98.56	88.79	0.68	0.31	99.8	467	99.28	88.87
SRF	MA0083	20841	11113	166	0.71	0.36	96.01	243	78.92	53.91	0.43	0.33	97.25	224	67.47	50
NRSF (REST)	M00256	4966	1811	1193	0.82	0.51	94.96	656	53.39	97.1	0.48	0.56	88.03	212	13.24	74.53
GATA2	M00128	12984	5551	75	0.57	0.12	99.11	124	90.67	54.84	0.48	0.13	99.55	125	92	55.2
GATA1	M00348	67847	31192	402	0.57	0.09	98.84	683	92.29	54.32	0.53	0.12	99.77	704	97.51	55.68
MAX	M00118	28296	3475	246	0.58	0.17	99.4	269	95.12	86.99	0.59	0.17	99.84	275	97.56	87.27
YY1	M00069	65357	22562	840	0.66	0.18	98.6	998	91.9	77.35	0.63	0.17	99.04	999	92.02	77.38

This table summarizes the ChIP-seq validation results for the 6 LCL transcription factors used in the main paper (Table S3) and 12 additional TFs with ChIP-seq data on K562 cell-lines available from the ENCODE Project. The columns report the total number of instances (after mappability filter) considered, the number of ChIP-seq positive/negative instances, and the prediction accuracies for the CENTIPEDE model using the DNaseI footprint, with and without including histone modifications. The prediction accuracy is measured in terms of: i) correlation with ChIP-seq (higher is better) and the Control (lower is better), ii) area under the ROC curve (% AUC) iii) and at 1% FPR, the total number of instances predicted to be bound, the % sensitivity and the % precision.

Table S5: **Summary of average performance**.

	Average in LCLs					Average in K562				
	At 1% FPR			Correlation w/		At 1% FPR			Correlation w/	
Method	%Sens.,	%Prec.	% AUC	Chip.,	Control	%Sens.,	%Prec.	% AUC	Chip.,	Control
CENTIPEDE w/ DNaseI footprint	81.47	60.51	98.11	0.65	0.32	82.84	76.6	97.88	0.67	0.27
CENTIPEDE w/ Footpr. w/o Shrinking	83.26	60.43	95.3	0.6	0.26	80.75	75.5	93.9	0.6	0.22
CENTIPEDE w/ Footprint w/o Prior	80.72	60.61	97.35	0.64	0.32	80.95	76.3	96.08	0.64	0.25
CENTIPEDE w/ DNaseI w/o Footprint	64.67	58.11	97.02	0.63	0.35	75.57	75.7	97.48	0.66	0.28
Number of DNaseI cuts	66.61	58.06	96.23	0.62	0.37	76.6	74.6	97.31	0.65	0.29
CENTIPEDE w/ Footprint + Histones	72.33	50.63	96.52	0.58	0.38	76.13	73.5	97.92	0.56	0.28
CENTIPEDE w/ Histones only	44.27	35.91	85	0.31	0.31	63.68	65.9	92.79	0.41	0.22
PWM	21.28	35.14	73.01	0.25	0.05	26.26	31.6	66.7	0.18	0.05
Conservation	9.73	30.71	77.4	0.26	0.13	7.75	26.7	73.58	0.22	0.1

This table summarizes the average performance at predicting ChIP-seq data for several methods across 6 LCL transcription factors and 12 additional TFs with ChIP-seq data on K562 cell-lines available from the ENCODE Project. The columns report: i) the average % sensitivity and % precision at 1% FPR, ii) the area under the ROC curve (% AUC), and iii) the correlation with ChIP-seq (higher is better) and the Control (lower is better).

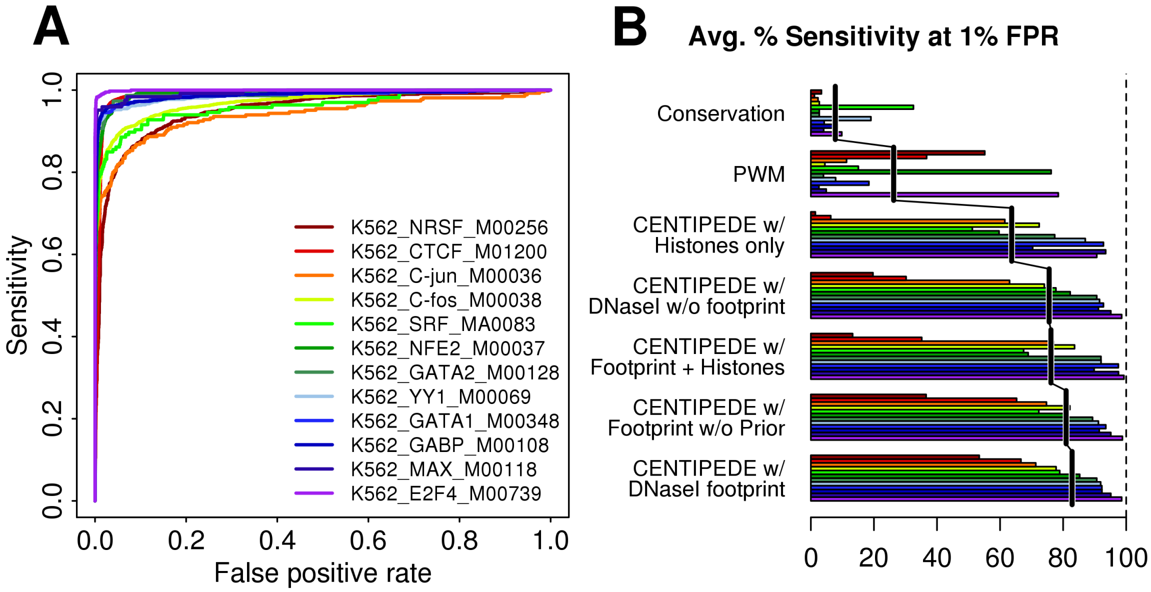


Figure S2: **ChIP-seq validation of CENTIPEDE predictions in the K562 cell-line**. This figure is the equivalent of Figure 3for the K562 cell-lines, where more ChIP-seq experiments are available. For K562 we see again that Centipede with DNaseI is the model with best average performance across all TFs, and that for some TFs, histone modification data can offer some improvement.

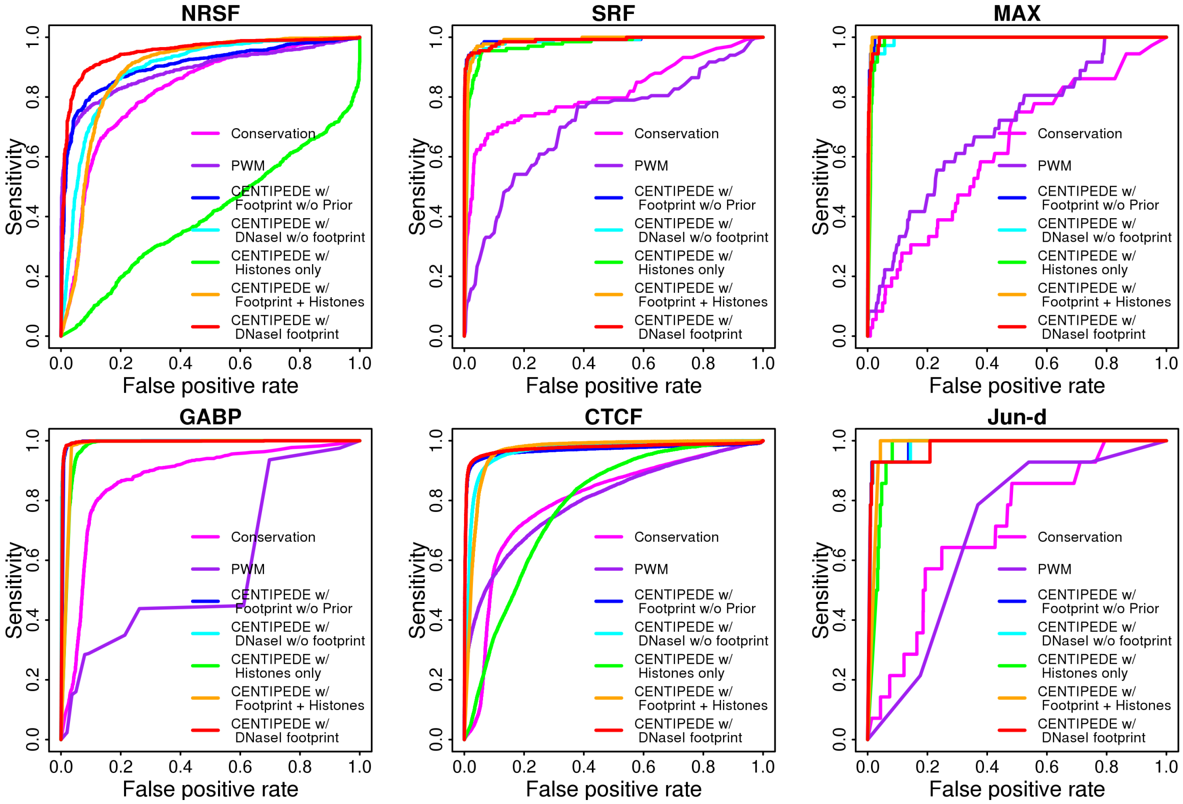


Figure S3: **Detailed ChIP-seq validation of CENTIPEDE predictions in LCLs**

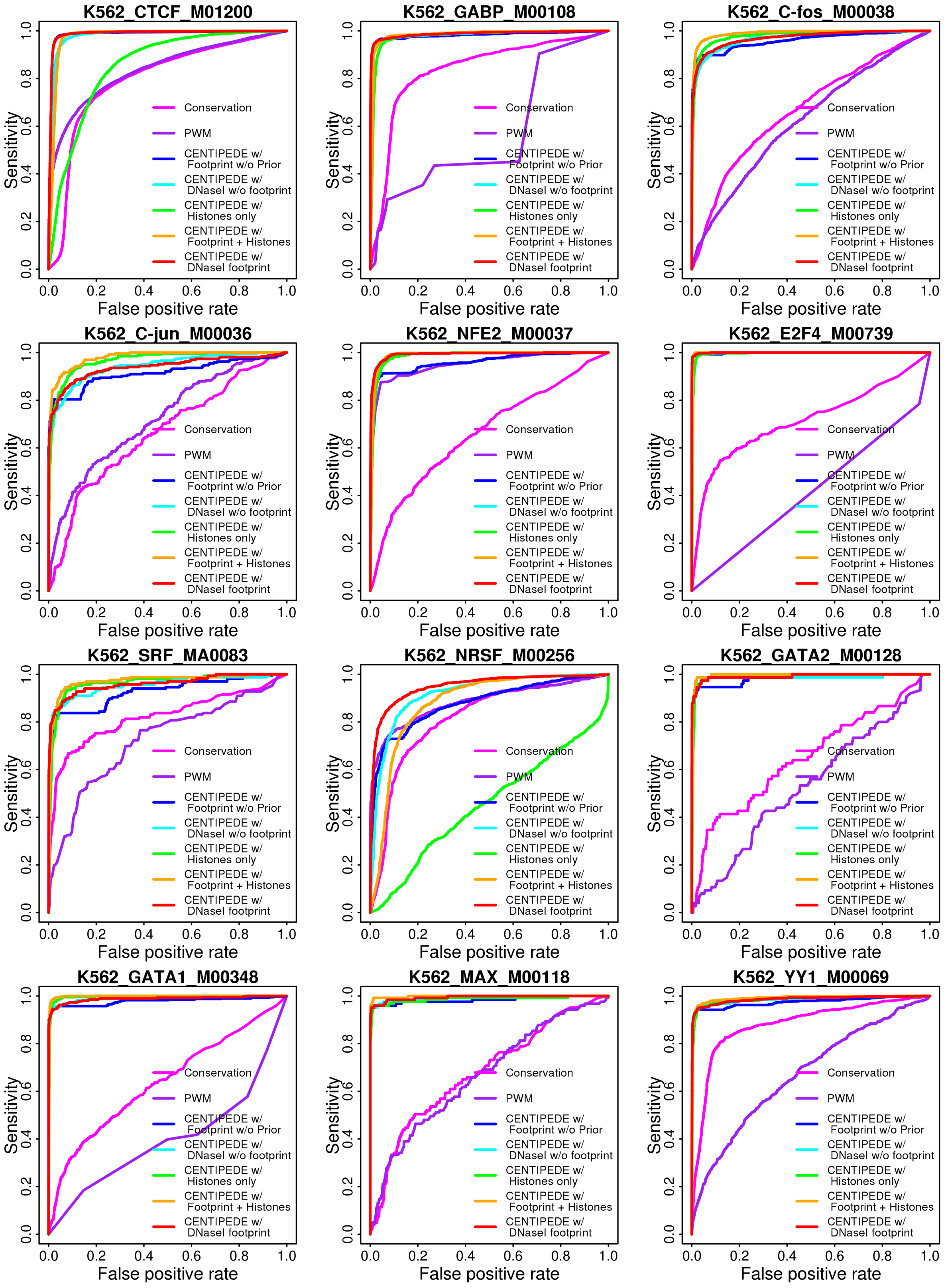


Figure S4: **Detailed ChIP-seq validation of CENTIPEDE predictions in K562**

4.2 Relationship to other methods

Several recent studies have also attempted to make *in silico* predictions of actively bound targets of individual TFs. A thorough comparison of all these methods is not possible as each method is generally tested on a different set of data, uses a different gold-standard for evaluation, is intended to solve a different problem, or is not readily extensible to apply directly to the data we use in this study. Nevertheless, despite these limitations, we discuss the main differences among these approaches here.

Ernst et al. 2010 [15] developed a computational method based on a logistic regression for combining multiple genome annotations to predict active TF binding. These annotations include both sequence-derived features (such as GC content, gene annotation and conservation) and experiment-derived data (such as histone modifications, and DNaseI hypersensitivity). The experimental data was tissue specific but they aimed at extrapolating potential TF binding to any tissue or cell condition, so data is combined together across tissues. In contrast, CENTIPEDE aims to use the experimental data to predict TF binding under specific experiment conditions. Another difference is that their approach is non-motif specific, and can be used to score individual bases for TF binding potential without using a motif. Unlike CENTIPEDE, however, their method requires a training step (i.e., it needs example regions that are confirmed to be TF binding sites). Xie et al. (2009) [16] used a comparative genomics approach, MotifMap, based on phylogenetic footprinting to predict the locations of functional regulatory sites, but did not integrate cell-specific experimental data.

Wang et al. (2009) [17] used an approach designed to integrate genomic data such as sequence conservation, CpG island richness, and DNase sensitivity in a model to predict active binding of USF1. Similar to our method, they started with a PWM scoring method to identify all candidate binding sites. Unlike the CENTIPEDE approach however, this approach required ChIP data to train their model parameters.

Won et al. (2010) [18] developed a computational method, Chromia, for integrating histone modification profiles to predict active TF binding in mouse embryonic stem cells. Like CENTIPEDE, Chromia also aims to identify binding sites which are active on the underlying experiment conditions and is motif-based (in contrast to Ernst et al. 2010 [15] approach discussed earlier). However, CENTIPEDE is more flexible in specifying the model for the experimental and sequence data in that: i) there is a hierarchical separation between the experimental-data model and the prior specified with a logistic regression for the sequence score and conservation; ii) the spatial distribution of the different data-types surrounding a binding site can be specified up to a single base-pair resolution. This may explain why we do not obtain a degradation in performace when we include conservation in the model as is the case for Chromia. Chromia uses an HMM to exploit the spatial distribution of data-types but at a much more limited resolution level (100bp windows and 3 states for the transitions form left, central to right flank of a motif). This may be sufficient for histone modification data, but is probably not enough to capture the DNaseI cleavage pattern. The Gaussian mixture model used by Chromia may be more flexible than CENTIPEDE to model dependency between histone modification experiments or multi-modal patterns, but requires supervised learning and the availability of suitable training data. Both Won et al. (2010) [18] and Ernst et al. 2010 [15] find histone modifications useful in predicting binding sites, which are also found useful with CENTIPEDE. However, Won et al. did not attempt to use DNaseI data at all, and Ernst et al. only used DNaseI hypersensitive sites as a binary feature (a very early low coverage DNaseI experimental track [9]). To the best of our knoweledge CENTIPEDE is the only computational approach available that is capable of learning TF specific DNaseI cleavage patterns[11] for predicting binding sites.

4.3 Conservation as validation of the CENTIPEDE predictions.

When ChIP-seq data are not available to validate the accuracy of the CENTIPEDE predictions we use conservation to assess whether the model is working properly. In this case we exclude conservation from the model, and we measure whether motif instances with high posterior have significantly higher conservation of the underlying DNA sequences across different species. Step by step:

1. Extract the PhastCons conservation scores for the placental mammals in the 44-way alignment (UCSC).
2. Calculate for each motif instance, the average PhastCons conservation score C_l.
3. Compute the CENTIPEDE posterior probabilities p_l excluding conservation from the model.
4. Fit a logistic model:
$$log(p_l/(1-p_l)) = \beta_0 + \beta_1 C_l \tag{25}$$
5. Calculate the z-score for conservation corresponding to β_1 using the Hessian matrix calculated at the maximum likelihood estimate. This gives exactly the same p-value as a generalized linear model (GLM) in R.

Finally, we applied a correction to these Z-scores that normalized across motifs with respect to the average distance to the nearest TSS by regressing out this effect.

5 Motif analysis

5.1 *De novo* motif discovery.

The approach that we use to build the CENTIPEDE model requires a set of candidate binding sites. Section 2.1 shows how we can recover the candidate binding sites using databases that contain known TF motifs. However, there is potentially a significant number of active regulatory elements for which a PWM has not yet been described. Since we know from this work (and previous work [9]) that DNaseI hypersensitive sites are enriched for matches to the PWMs of active regulatory elements, we reasoned that a comprehensive analysis of enriched k-mers on hypersensitive sites would be useful for discovering novel motifs (as well as recovering known PWMs). With the aim of verifying this hypothesis and discovering novel motifs we developed the approach presented in this section.

First, we counted the number of times every possible 10-mer occurred within the DNA sequences in the most DNaseI sensitive regions of the genome. These regions were defined using a 200bp sliding window centered on every single base pair and selecting the positions with more than 200 DNase-seq reads. In total, 6.4 Mb (0.21% of the human genome) met this criterion, and on average each 10-mer occurred 12.2 times within this region (where a k-mer and its reverse compliment are combined). We defined an “enriched” set of 10-mers as being those words that occurred more than 50 times in these DNaseI sensitive regions (corresponding to the top 3 percentile of the distribution). In addition, we constructed a “control” set of 20,000 10-mers that occurred 6 or fewer times (corresponding to the bottom 50 percentile) in these regions.

For each word in the enriched set, we ran the CENTIPEDE model on all the matches of the word along the genome. For the control words, we used all the matches up to 1 base away from the original word and used a rejection sampling strategy to match the distribution of DNaseI HS to that for the enriched words. This sampling procedure was used to control for the correlation of DNaseI regions with functional elements.

After applying the CENTIPEDE model to both the enriched words and the control words, we used the conservation Z-score procedure described in Section 4.3 to validate the predictions. We estimated that 735 of the enriched 10-mers showed significantly higher conservation, on average, for the sites predicted to be bound than for those sites predicted to be unbound (with a 10% FDR estimated using the control words). Using the CENTIPEDE models from each of these 735 significantly enriched 10-mers, we expanded the word-space covered by calculating the posterior probabilities for all locations in the genome that were at most one base-pair mismatch away from the original 10-mer seed word. Using the posteriors calculated, we estimated a PWM for each of these words by estimating the frequencies of each base in the high confidence predicted binding locations (Posterior Probability > 0.99). For the estimation of the PWMs, we first considered all bases within 50 bp of the 10-mer as potentially being informative, and then trimmed the resulting PWM of uninformative bases up to the first and last base that had an information content of $>$ 0.25 bits. (The estimated PWMs and their predicted binding sites for all words can be found at `http://centipede.uchicago.edu`.) In the following section, the overlap between the predicted binding sites for every pair of motifs will show that the procedure described in this section can indeed recover a large proportion of the known PWMs motifs, as well as discover 49 novel words (corresponding to approximately 24 non-overlapping sets).

5.2 Identification of groups of overlapping motifs.

There is redundancy in the set of motifs we analyzed. Some PWMs obtained from databases (Section 2.1) are derived from the same TF (or different TFs that recognize the same DNA sequence motif). Additionally, the motifs derived from words (Section 5.1) may overlap if they are a single nucleotide distance from one another. In this section we systematically explore the overlap between the predicted binding sites for all pairs of motifs (PWM or word derived). For each pair of motifs a and b:

1. Calculate the sets A and B of binding sites with a high posterior from the CENTIPEDE model > 0.99.

2. Calculate the intersection set $A \cap B$ with the instances form A and B that overlap on at least one single bp.

3. Measure the size of the sets ($|A \cap B|$, $|A|$, and $|B|$).

4. Finally, the amount of overlap between motifs a and b is:

$$\text{Overlap}\,(a, b) = \frac{|A \cap B|}{\min\left(|A|\,, |B|\right)} \tag{26}$$

From the analysis of all known PWMs (Section 2.1) and the word derived motifs (Section 5.1) we can retrieve 826,896 unique locations of being a TF-bound site with high probability. Of these locations, 431,724 were detected using PWMs, and 574,567 using words (179,395 recovered from both analyses). This indicates that a large proportion of the words can retrieve the original known PWM motif and expand a very restrictive PWM description. However, since our aim was to discover novel motifs with the 10-mer analysis, we removed all 10-mer derived words with at least 10% overlap to a known PWM. After this filtering step, we were left with 49 novel 10-mer motifs.

For the rest of the analyses we continued with the set of 239 known PWMs and 49 novel motifs. We estimated the groups of independent motifs from this set of 288 motifs by agglomerative clustering of the predicted binding sites. We recursively joined the pair of motifs with the largest overlap measure (eq. 26). Once a pair of motifs was joined, we used the average measure of overlap between all the leaves to merge two nodes (i.e., average linkage). Using a maximum pairwise distance cutoff of 10% between nodes (distance defined as 1-overlap), we estimate that there are 95 clusters of elements with little overlap. As expected, we find that different PWMs representing the same TF are grouped together and groups of related TFs (e.g. TFs with very similar PWMs) are almost invariably joined as members of the same cluster. Throughout the main text of this paper, we highlighted only a single member of each of these clusters (e.g. in the motifs listed in Figure 6). A complete list of motifs and their cluster membership can be found at `http://centipede.uchicago.edu`.

5.3 Novel words that have a match in existing protein binding microarray data

Protein binding microarrays (PBMs) are very useful for determining the set of sequences that a transcription factor binds [19, 20, 21, 22, 23]. Microarray probes that include, exhaustively, all possible short sequences (e.g. 8-mers) are tested for binding affinity to a protein of interest. Using the publicly available data deposited in the UniProbe database (`http://the_brain.bwh.harvard.edu/uniprobe/`), we checked if any of our novel words have a know binding partner. For 23 out of 49 words, we identified a protein that has a strong affinity for that sequence. In Table S6, we report the proteins with the highest enrichement score (ES) for these 23 words (although some of these words were significantly enriched for multiple proteins). However, when we conducted the same analysis for control words we found that roughly 50% also provided a ES above 0.45 (the default threshold). Thus, this analysis seems only suggestive of a potential binding partner and a more thorough analysis would be necessary to reach any definitive conclusion.

5.4 GO analysis of putative gene-targets of each TF.

Next, we asked whether the putative target genes for each TF were enriched for functional categories that one might expect to be coordinately regulated. For each motif, we took as putative targets the set of genes that had a high posterior binding site within 5 kb of an annotated TSS for that gene. We used the R package ‘GOstats’ to perform a hypergeometric test for overrepresentation conditional on the parent-child structure of the GO categories represented in the target gene set ([24, 25]). We took all Entrez genes with an assigned GO category as the gene universe against which we would test for significant overrepresentation in sets of target genes for a specific motif.

Table S6: Matching discovered novel words to TFs using PBM.

Novel 10-mer	Protein	PBM k-mer	E.S.	Reference
AGAACTACAA	Gat3	AGAACTAC	0.49477	“Zhu et al., GR 2009” [22]
AGGAAGGGGC	Plagl1	GAAGGGGC	0.469166	“Badis et al., Science 2009” [21]
AGGAAGGGGG	Zfp281	GAAGGGGG	0.48417	“Badis et al., Science 2009” [21]
AGGCAGGAAG	Ehf	GCAGGAAG	0.465621	“Badis et al., Science 2009” [21]
AGGGAGGAAG	Sfpi1	GGAGGAAG	0.467427	“Badis et al., Science 2009” [21]
CAAAATGGCG	E2F2	AAATGGCG	0.452616	“Badis et al., Science 2009” [21]
CACTGCGCAG	Zbtb3	CACTGCGC	0.452079	“Badis et al., Science 2009” [21]
CAGCCCCGGG	Tcfap2c	GCCCCGGG	0.48458	“Badis et al., Science 2009” [21]
CCCACCCTCC	Zfp740	CCCACCCT	0.458455	“Badis et al., Science 2009” [21]
CCCCCCTTCC	Zfp740	CCCCCCTT	0.495301	“Badis et al., Science 2009” [21]
CCCCTTTAAG	Yer130c	CCCCTTTA	0.474102	“Zhu et al., GR 2009” [22]
CCCGGAAGCA	Gabpa	CCGGAAGC	0.494322	“Badis et al., Science 2009” [21]
CCCTTCCTCC	Sfpi1	CTTCCTCC	0.467427	“Badis et al., Science 2009” [21]
CCCTTTAAGG	Nsy-7	CTTTAAGG	0.454334	“Lesch et al., GD 2009” [23]
CCTTCCTCCC	Sfpi1	CTTCCTCC	0.467427	“Badis et al., Science 2009” [21]
CCTTTAAGAG	Bapx1	TTTAAGAG	0.45061	“Berger et al., Cell 2008” [20]
CGGGGAGGAA	Sp4	GGGGAGGA	0.462762	“Badis et al., Science 2009” [21]
GCCCCTTTAA	Plagl1	GCCCCTTT	0.47783	“Badis et al., Science 2009” [21]
GGAACTACAA	Gat3	GGAACTAC	0.493443	“Zhu et al., GR 2009” [22]
GGCCACACCC	Met32	GCCACACC	0.493529	“Zhu et al., GR 2009” [22]
GGCCCCTTTA	Plagl1	GGCCCCTT	0.494069	“Badis et al., Science 2009” [21]
GGGAAACTGA	Irf3	GGAAACTG	0.473723	“Badis et al., Science 2009” [21]
TGCGCGCGCA	Zfp161	GCGCGCGC	0.498641	“Badis et al., Science 2009” [21]

The significance of overrepresentation was assessed with a Bonferroni corrected P-value where the total number of tests was the total number of GO categories in the set of target genes of the motif being tested. Figure 4 in the main text, and Figure S14 report the highest enriched functional category for the set of genes that are the target of each TF.

5.5 Co-occurrence analysis of binding locations for each TF.

For each motif, we explored whether there was significant co-localization with other motifs that might indicate the presence of regulatory modules. For each pairwise comparison of motif-types i and j, we estimated the expected number of bound sites of j in the non-overlapping 200 bp region immediately outside of the i motifs. This expectation was controlled for the specific overrepresentation of most motifs at TSSs, and we used a two-sample Poisson test to test for significant overrepresentation compared to this expectation. The TSS-controlled expectation for each pair of i and j was calculated as:

$$E[j\epsilon i] = \sum_{w=1}^{K} j_w \left(\frac{\{J_w \epsilon i_w\}}{J_w} \right) \tag{27}$$

where $E[j\epsilon i]$ is the expected number of bound sites of i in the non-overlapping 200 bp region immediately outside of the j motifs. w indexes non-overlapping windows of distance to the TSS that each high-posterior binding site was assigned to. Thus, j_w refers to the number of high posterior binding sites of motif-type j that fall within the distance to the nearest TSS defined by w. $\{J_w \epsilon i_w\}$ refers to the total number of all motif types within a TSS window that fall into the regions surrounding i and J_w refers to the total number of all motif-types that fall into the window defined by w. The choices for non-overlapping windows were the regions between the following breakpoints: 0 bp, 250 bp, 500 bp, 750 bp, 1 kb, 2kb, 3kb, 4kb, 5kb, 6kb, 7kb, 8kb, 9kb, 10kb, and >10kb from TSS. Motif j was considered significantly over-represented in the windows surrounding i if the actual count of these co-occurrences was significantly greater than the expectation, using an FDR threshold of 5% [26].

5.6 Linear modeling of steady state mRNA levels.

Gene regulation is governed through combinatorial action of many transcription factors and it is expected to be a non-linear highly complex function of the regulatory elements nearby the TSS. However, we hypothesized that simple linear model could explain a significant portion of the of the gene expression levels using the TF binding sites as explanatory variables. In order to test this hypothesis, absolute measures of steady state mRNA in lymphoblasts, averaged across 69 Yuruban individuals, were obtained for each of 33,502 Ensembl genes from [27]. We defined 288 indicator variables, one for each motif. For each gene (i), an indicator variable, I_motif ,took on a value of 1 if there was a high posterior site within 5 kb of any of the annotated TSSs for that gene and took on a value of 0 otherwise. We then modeled log2-transformed gene expression (measured as average RNA-seq reads per base per individual) as a linear combination of these variables. That is, we fit the linear model:

$$Y_i = \beta_0 + \beta_1 I_{motif1,i} + \beta_1 I_{motif2,i} \cdots + \beta_{288} I_{motif288,i} + \epsilon_i \tag{28}$$

where i indexes each gene. We used stepwise addition (implemented in the R package 'lars') coupled with Bayesian Information Criterion (BIC) to perform variable selection. Ultimately, we found that the top 96 indicator variables gave the minimum BIC. With this relatively simple model, the linear model explains 38% of the total variance in gene expression (Figure S7). Interestingly, the binding sites corresponding to a small number of TF could explain a large portion of the total variance in expression across genes. For example, the SP1 TRANSFAC motif M00196 explains 15% of the variance in gene expression, and the top 20 variables explain 32% of the total variance.

Since many of our predictions are correlated with DNaseI sensitivity, we asked whether these predictions had explanatory power beyond the general level of DNaseI sensitivity in the 5 kb window considered. For this analysis, we fit three measures of DNase sensitivity (sensitivity within 500 bp, sensitivity between 500 and 1000 bp, and sensitivity between 1000 and 5000 bp) together with the original 288 indicator variables described above. i.e.,

$$\begin{aligned} Y_i = \beta_0 + \delta_{DNaseA,i} + \delta_{DNaseB,i} + \delta_{DNaseC,i} + \beta_1 I_{motif1,i} + \\ \beta_1 I_{motif2,i} \cdots \beta_{288} I_{motif288,i} + \epsilon_i \end{aligned} \tag{29}$$

When fitting this model, we find that the three DNaseI measures explain 37% of the total variance in gene expression while the remaining 62 significant factors explain an additional 8%. Thus, we find that there is additional information in the TF binding site predictions that is lost when treating DNaseI only as a general measure of active regulatory elements. For example, while NRSF and CTCF sites both fall into DNaseI hypersensitive sites, each has a repressive effect on nearby genes even though a typical hypersensitivity site is associated with an increase in gene expression. For the remaining activating motifs, it is unclear whether TF binding or chromatin changes are the root cause of expression differences in nearby genes.

The effect sizes of the multiple regressions, presented before, are difficult to interpret given that many of the 288 factors partially overlap each other (i.e., multicolinearity). To alleviate this problem we estimated the effect size individually for each factor in a smaller linear model that only adds the DNaseI sensitivity of the promoter region:

$$Y_i = \beta_0 + \delta_{DNaseA,i} + \delta_{DNaseB,i} + \delta_{DNaseC,i} + \beta_j I_{\text{motif}j,i} + \epsilon_i \tag{30}$$

We report the marginal effect sizes for this final model in Figure 6 of the main text and in Figure S14.

5.7 Analysis of multi-tissue expression for TF target genes.

We extracted gene expression values for 16,077 Ensembl genes from the tissue gene expression atlas (`http://biogps.gnf.org/downloads/`) corresponding to 86 tissues / cell-lines. The gene expression for each tissue was normalized using RMA [28], and then log-2 transformed. Using the same definition of target genes as in the previous section, we calculated for each tissue the average gene expression across all the genes that a particular TF targets. The resulting average gene expression per TF and tissue can be visualized as an image, see Figure S13. TFs that also affect the gene expression for tissues other than LCLs, appear as long stretches of red, implying that core promoter elements like Sp1/GC affect in a similar manner a large number of genes. On the other hand, TFs that are important in B-cell development such as Pax5, PU.1 only appear on LCLs and other closely related cell-lines. For the Figure 6 of the main part of the paper we selected a subset of TFs and tissues that best represent this in a compact way.

5.8 Estimating enrichment/depletion of histone modifications at TF binding locations.

We estimated the enrichment/depletion of histone modification at TF binding locations as in Section 4.3. For each histone modification H_k, separately, we extracted the number of reads within a 400bp window of each motif instance l. Then, for each histone modification k, we fit the following logistic model:

$$log(p_l/(1-p_l)) = \beta_0 + \beta_k H_{l,k} \tag{31}$$

where p_l is the CENTIPEDE posterior probability for motif instance l. For each β_k we can calculate a z-score using the second derivative of the maximum likelihood estimate (i.e. the Hessian). This is equivalent to the way statistical significance is calculated for a generalized linear model (GLM) in R.

6 Resource availability

The regulatory map for LCLs and the source code for CENTIPEDE will be released at http://centipede.uchicago.edu.

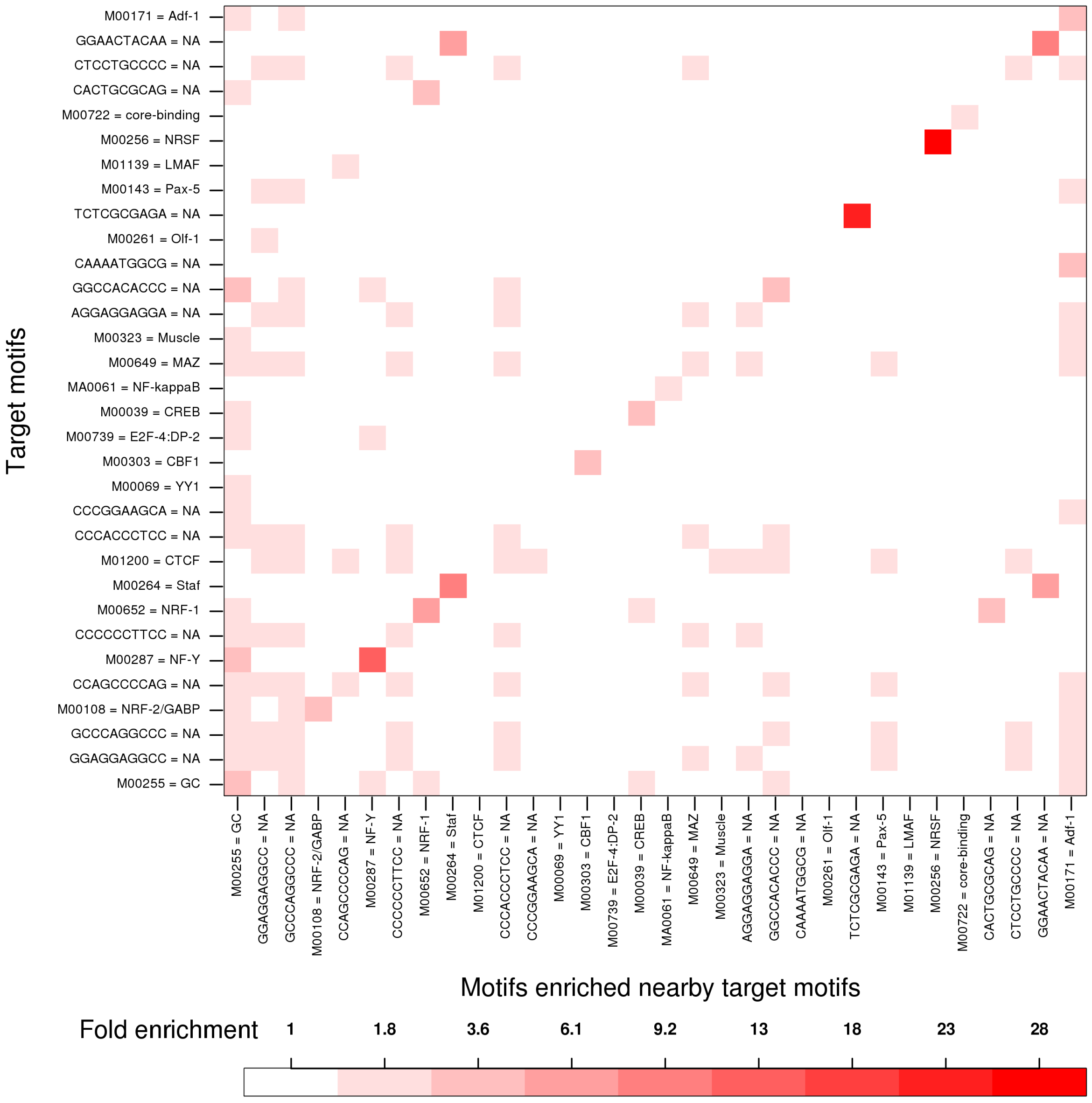


Figure S5: **Examples of motifs from Figure 6 of the main text with evidence of significant co-localization.** For each motif included in Figure 6 of the main text, we plot the fold enrichment of other motifs (also limited to those in Figure 6) in the 200 bp non-overlapping window surrounding the motif. Only those motifs pairs with significant enrichment are colored, and the color scale gives the mean fold enrichment for pairs of motifs with the corresponding color. The full table giving all significant pairwise overlaps is provided at http://centipede.uchicago.edu.

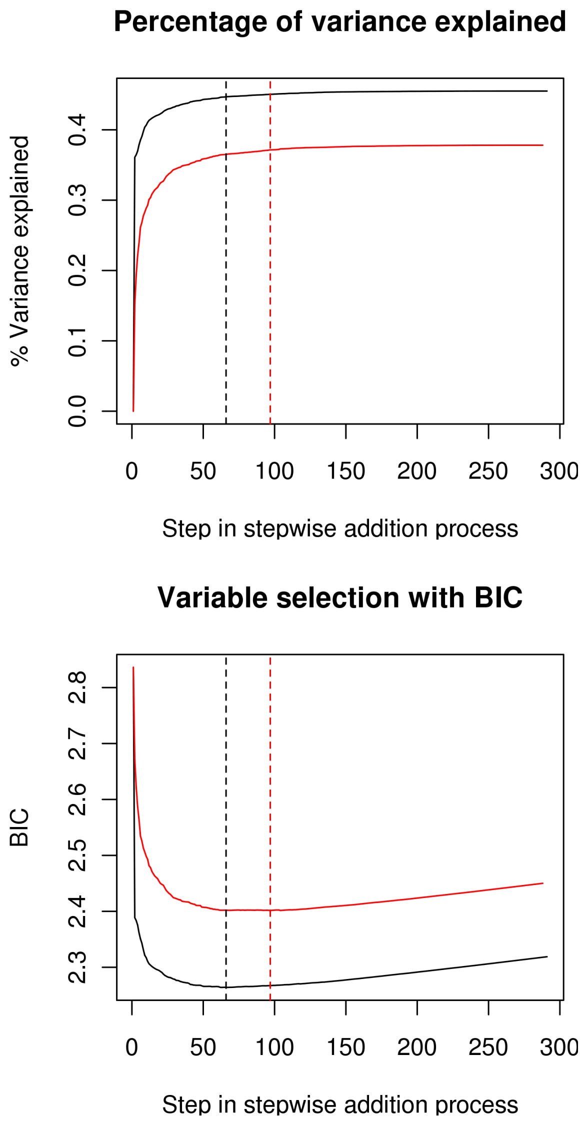


Figure S6: **The percentage of variance explained by linear models with and without DNaseI sensitivity and variable selection with BIC.** The top panel shows the increase in R^2 for each variable added to the model using forward stepwise addition; the black/(red) solid lines corresponds to the model with/without the DNaseI sensitivity variables in addition to the indicator variables. The bottom panel shows the value of the Bayesian Information Criterion (BIC). The dashed lines mark the location of the minimum BIC for the models.

7 References

[1] Gupta, S., Stamatoyannopoulos, J., Bailey, T. & Noble, W. Quantifying similarity between motifs. *Genome Biol* **8**, R24 (2007).

[2] Stormo, G. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).

[3] Wingender, E., Dietze, P., Karas, H. & Knüppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**, 238–41 (1996).

[4] Sandelin, A., Alkema, W., Engström, P., Wasserman, W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**, D91–4 (2004).

[5] Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034–50 (2005).

[6] Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res* **30**, 38–41 (2002).

[7] McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* (2010).

[8] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).

[9] Boyle, A. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–22 (2008).

[10] Fu, Y., Sinha, M., Peterson, C. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* **4**, e1000138 (2008).

[11] Hesselberth, J. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* **6**, 283–9 (2009).

[12] Heintzman, N. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311–8 (2007).

[13] Degner, J. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–12 (2009).

[14] Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

[15] Ernst, J., Plasterer, H., Simon, I. & Bar-Joseph, Z. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res* (2010).

[16] Xie, X., Rigor, P. & Baldi, P. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics* **25**, 167–74 (2009).

[17] Wang, T. *et al.* A general integrative genomic feature transcription factor binding site prediction method applied to analysis of USF1 binding in cardiovascular disease. *Hum Genomics* **3**, 221–35 (2009).

[18] Won, K., Ren, B. & Wang, W. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol* **11**, R7 (2010).

[19] Bulyk, M., Huang, X., Choo, Y. & Church, G. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci U S A* **98**, 7158–63 (2001).

[20] Berger, M. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–76 (2008).

[21] Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–3 (2009).

[22] Zhu, C. *et al.* High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* **19**, 556–66 (2009).

[23] Lesch, B., Gehrke, A., Bulyk, M. & Bargmann, C. Transcriptional regulation and stabilization of left-right neuronal identity in C. elegans. *Genes Dev* **23**, 345–58 (2009).

[24] Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–7 (2006).

[25] Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–8 (2007).

[26] Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, pp. 1165–1188 (2001).

[27] Pickrell, J. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* (2010).

[28] Irizarry, R. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–64 (2003).

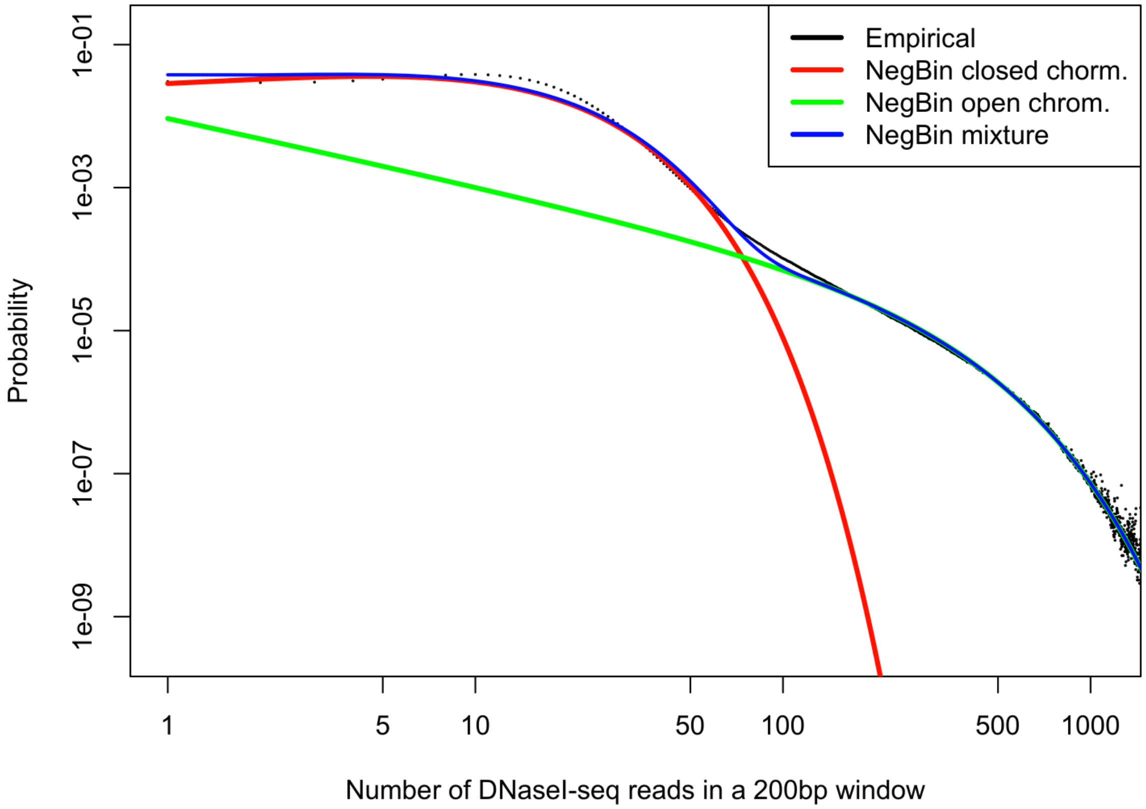


Figure S7: **Negative Binomial modeling of DNase-seq reads.** Genome-wide distribution of the total number of DNase-seq reads for every 200bp window in the genome. The black dots are the empirical probabilities calculated as the fraction of times a window with that number of reads is seen in the genome. The blue line corresponds to the probabilites of a mixture of two negative binomail distributions that are represeted with a green and a red line. The two components, green and red, can be attributed to the DNase-seq reads coming from accessible (open) chromatin and inaccessible (closed) chromatin, respectively.

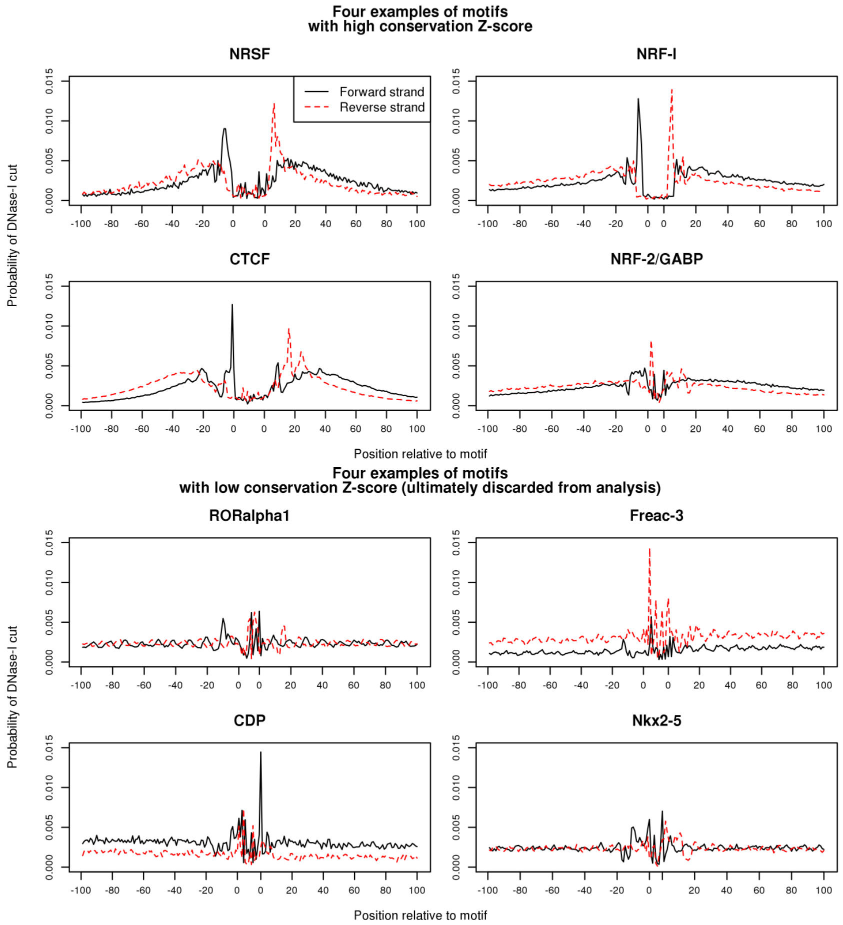


Figure S8: **Examples of estimated footprints for motifs that pass our conservation z-score threshold (top panel), and for which we discard the motif from further analysis (bottom panel)**. In the CENTIPEDE model, we have fit different multinomial parameters for positions on the forward (black solid lines) and reverse DNA-strands (red dashed lines). Due to the particular chemistry of the DNase-seq protocol, the positions of reads mapping to each strand is not necessarily expected to be symmetric (e.g., CTCF; [9]). For motifs that pass the conservation z-score threshold, we find considerable variation in the shape of the distribution of cutsite locations, and this difference in shapes can add to the TF-specificity of our model. There are unifying aspects to the conserved class of motifs that give us confidence they represent actual protein binding and can be distinguished from cases where the CENTIPEDE model might pick out differences in sequence-dependent DNaseI sensitivity. First, for motifs with high conservation z-scores, there is usually evidence of a protected site directly over the informative bases of the PWM, indicating that DNA-TF interactions produce steric hindrance to the DNaseI enzyme. In contrast, this central protected region is not usually observed for motifs with low conservation z-scores. Furthermore, for motifs with high conservation z-scores there are frequently spikes of sensitivity, on opposite DNA strands, on either side of the central protected region. These spikes are consistent with DNA conformational changes induced by TF-binding that sensitize particular positions to cleavage (e.g., widening of the minor groove). The opposing spikes on either strand are caused when two single stranded nicks on opposite sides of the bound protein lead to a double strand break with a single strand overhang. All footprints can be found at http://centipede.uchicago.edu.

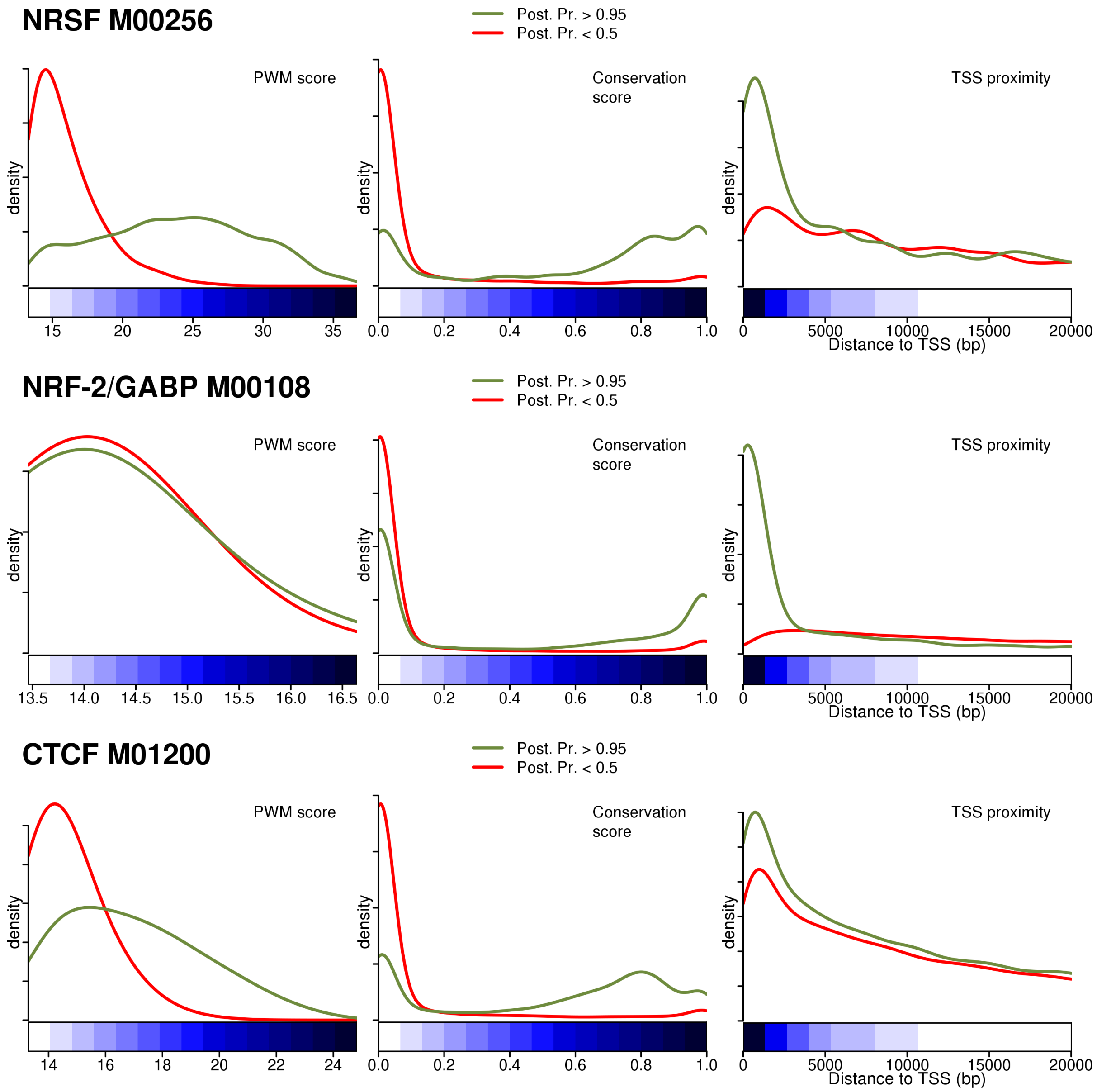


Figure S9: **Prior information for several TFs**. CENTIPEDE can adjust the logistic model paramters to learn the different distributions on the different types of prior information for different TFs (NRSF also known as REST, GABPA and CTCF).

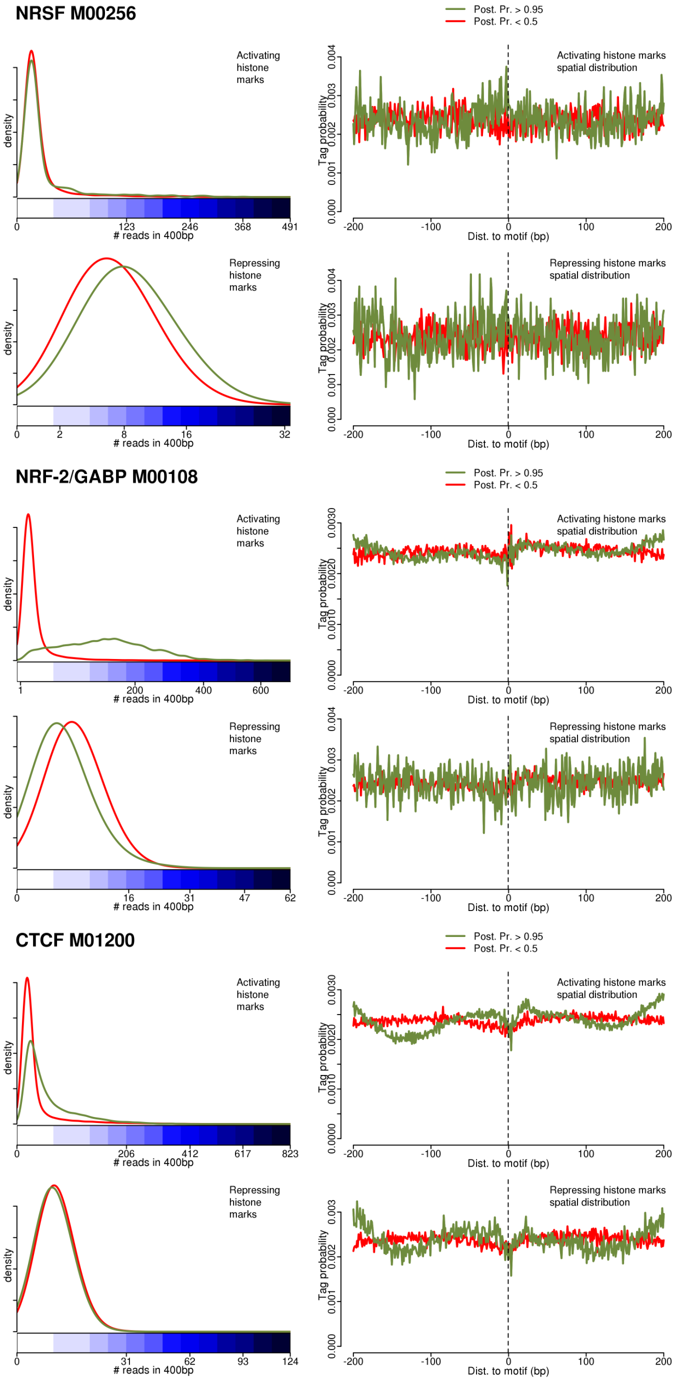


Figure S10: **Histone modification ChIP-seq profile for several TFs**. NRSF (REST) shows a small enrichment of repressing histone marks and no enrichment in activating histone marks. GABPA, on the other hand shows a strong enrichment of activating histone marks and a small depletion of repressing histone marks. For CTCF, the difference on the histone mark changes between bound and unbound states is smaller than in GABPA. For the three motifs, the spatial distribution of the histone marks ChIP-seq reads do not seem very informative for classifying the binding state.

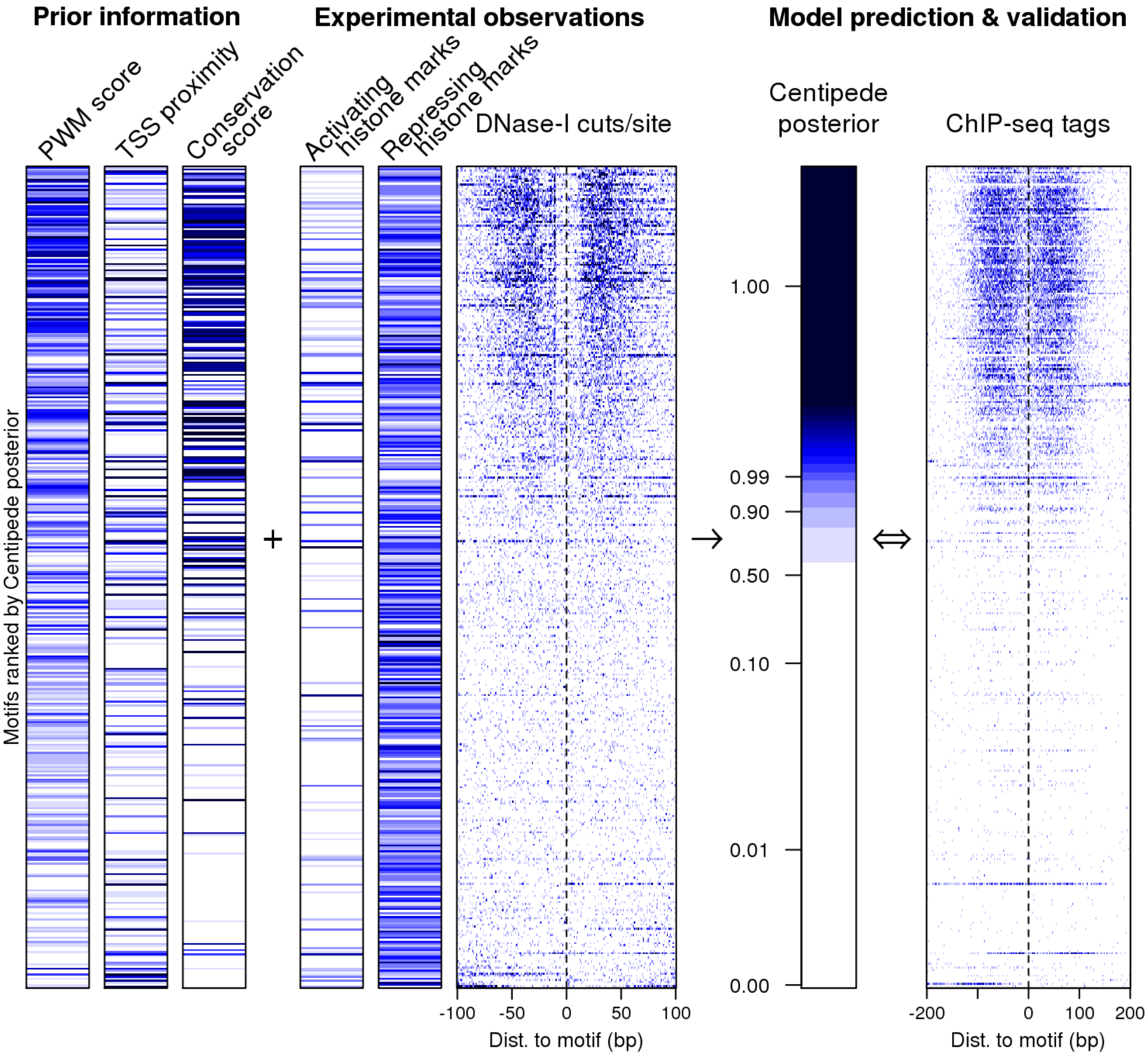


Figure S11: **Equivalent to Figure 1 for the CTCF motif**

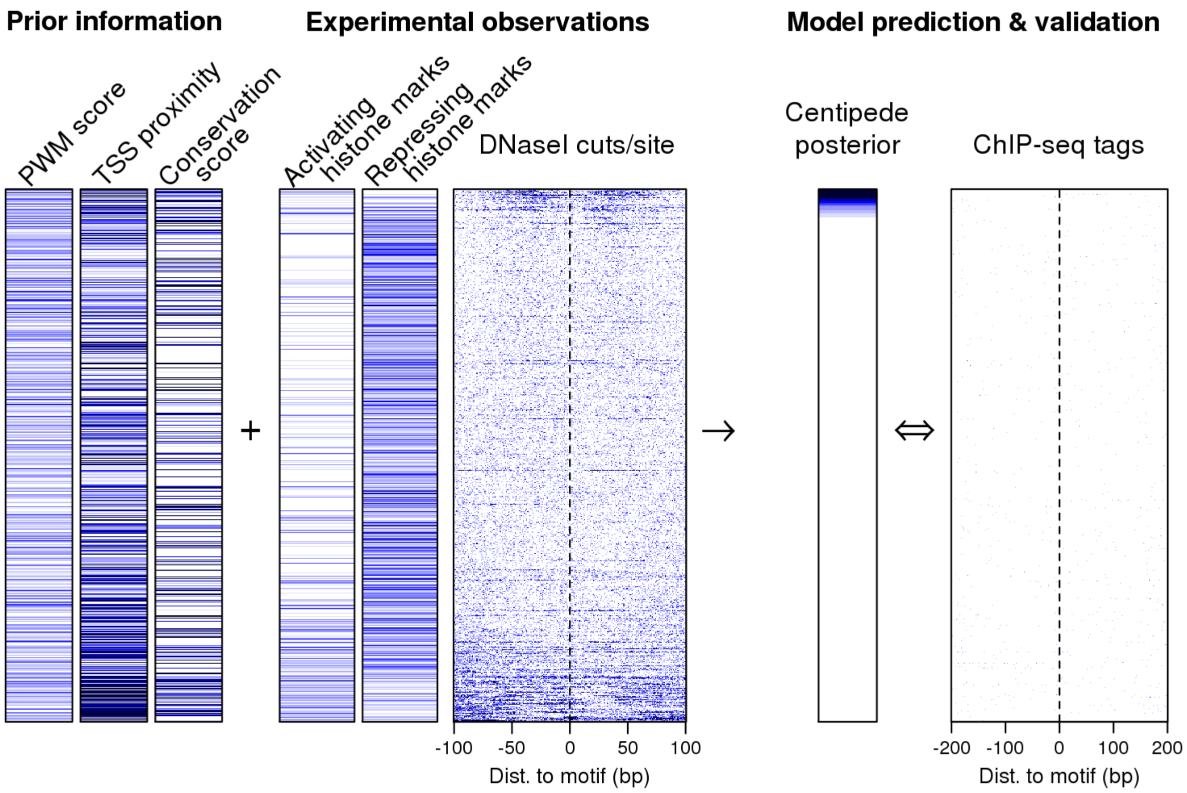


Figure S12: **CTCF motifs in DNaseI hypersensitive sites that are not bound according to ChIP-seq.** In contrast to Figure 1 and S11, here for each row we randomly selected CTCF motif instances that are in DNaseI hypersensitive sites (> 50 reads in a 200bp window) but are ChIP-seq negatives (TF not bound). This Figure illustrates that CENTIPEDE, in most cases, can effectively use the footprint information to report the motif site as not bound.

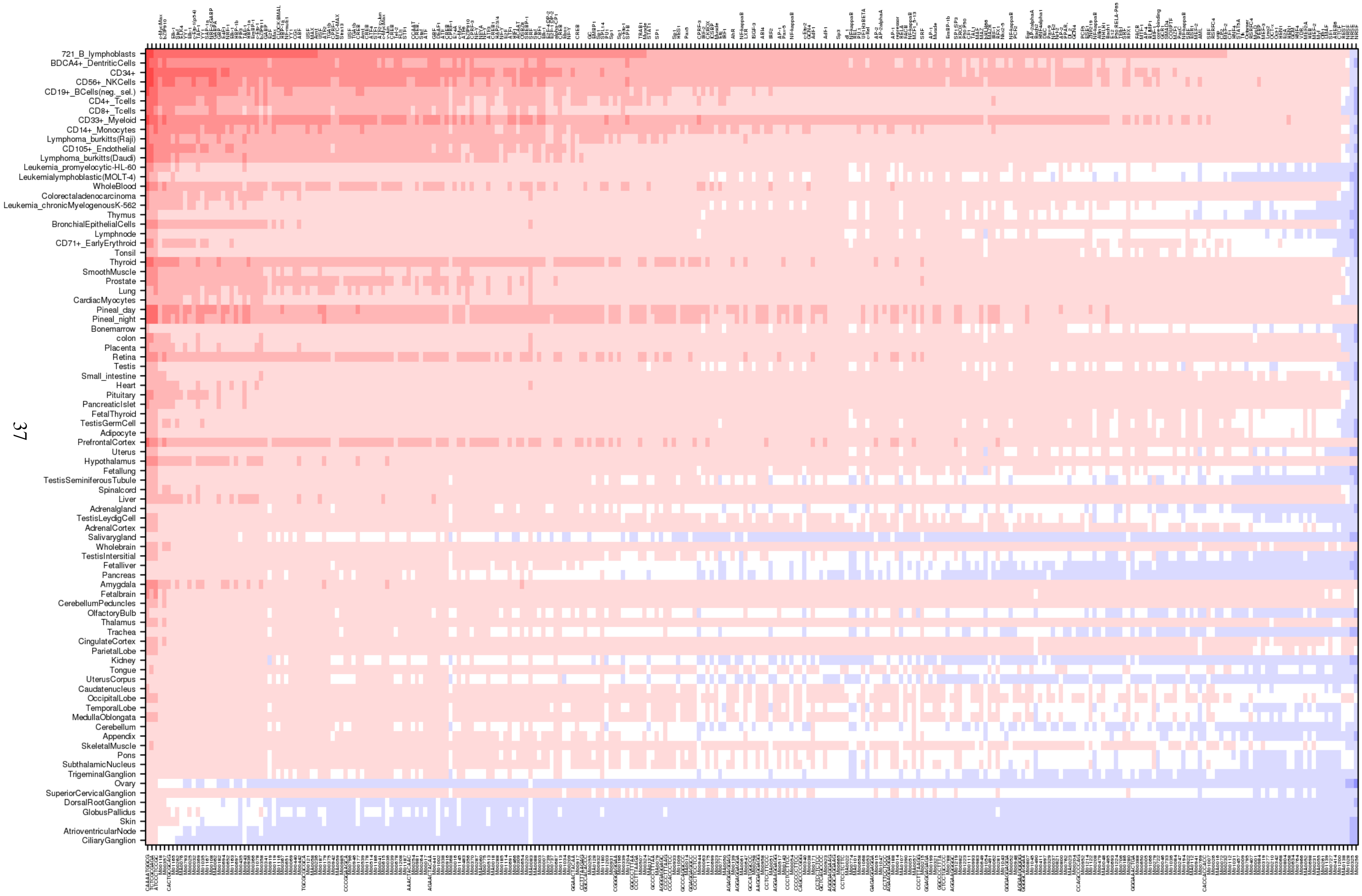


Figure S13: **Average gene expression across 86 tissues for the targets of 288 motifs predicted binding sites.** Rows correspond to 86 different tissues ordered by global gene expression similarity to LCLs. Columns correspond to each motif ordered by mean gene expression of the motif binding site targets in LCLs. A darker tone of red/blue represents a higher/lower average gene expression.

TF name/10-mer seed	TANSFAC/JASPAR motif	Number of sites predicted	Percent within 1kb of TSS	Top significant GO category	Most enriched nearby motif
CAAAATGGCG	-	715	52	gene expression	TAF-1
TCTCGCGAGA	-	574	93	translational elongation	YY1
GABPA	MA0062	3173	53	cellular metabolism	YY1
NRF-2/GABP	M00108	3173	53	cellular metabolism	YY1
YY1	M01035	305	93	ubiquitin-dependent protein catabolism	CPRF-3
SAP-1a	M01167	1220	73	cellular biopolymer metabolism	HBP-1b
ELK4	MA0076	1232	75	cellular biopolymer metabolism	CCCGGAAGCA
Elk-1	M00025	525	81	cellular biopolymer metabolism	HBP-1b
YY1	M00793	564	58	regulation of protein sumoylation	CPRF-3
c-Ets-1	M01078	818	54	cellular biopolymer metabolism	TCTCGCGAGA
FLI1	M01208	3639	38	cellular macromolecule metabolism	YY1
PU1	M01203	7059	17	cellular biopolymer metabolism	TCTCGCGAGA
Tel-2	M00678	3634	31	cellular macromolecule metabolism	YY1
CLOCK:BMAL	M01116	2227	46	nucleic acid metabolism	CACTGCGCAG
c-Ets-1(p54)	M00032	451	74	cellular biopolymer metabolism	bZIP910
GABP	M00341	2342	60	RNA splicing, via transesterification reactions	YY1
GBP	M00182	575	60	reg. of T-helper cell differentiation	TGCGCGCGCA
STAT1	MA0137	6286	13	response to virus	CPRF-3
HBP-1b	M00697	491	53	nucleic acid metabolism	Elk-1
E74A	M00016	3391	33	metabolism	CPRF-3
GBF	M00441	712	55	vesicle-mediated transport	AhR
Elk-1	M01163	297	77	cellular biopolymer metabolism	YY1
Elk-1	M01165	474	74	cellular biopolymer metabolism	NRF-1
MAX	MA0058	1592	44	nucleic acid metabolism	CAAAATGGCG
NF-muE1	M00651	726	70	ubiquitin-dependent protein catabolism	CPRF-3
c-Ets-1	M00339	7324	23	cellular process	ATCCCTCCGC
YY1	M00069	1280	47	intracellular transport	CPRF-3
USF	M00121	1646	61	cellular response to DNA damage stimulus	YY1
SPI-B	M01204	6239	14	cytokine biosynthesis	CCTTTAAGAG
ICSBP	M00699	8022	10	biological regulation	AAACTACAAC
Elk-1	M00007	5063	24	cellular macromolecule metabolism	MYC-MAX
PIF3	M00435	877	51	water-soluble vitamin metabolism	HIF1
Ets	M00771	4979	18	negative regulation of transcription	TCTCGCGAGA
ISGF-3	M00258	5440	12	regulation of RNA metabolism	GGCCCCTTTA
c-Myc:Max	M00615	1318	52	response to DNA damage stimulus	AhR
GGAGGAGGCC	-	9400	55	regulation of cellular metabolism	AGGGAGGAGC
IRF2	MA0051	3219	11	regulation of RNA metabolism	AGAACTACAA
E2F	M00919	654	73	DNA replication	ATCCCTCCGC
Sp1	M00932	10931	71	nucleic acid metabolism	CPRF-3
Stra13	M00985	1286	68	lysosome organization	YY1
c-Myc:Max	M00118	1004	37	-	YY1
MYC-MAX	MA0059	1130	39	nucleic acid metabolism	AhR
USF	M00796	801	55	cellular biopolymer biosynthesis	YY1
E4F1	M00694	601	54	nucleic acid metabolism	CAAAATGGCG
IRF1	MA0050	7859	12	transcription	AGAACTACAA
ATCCCTCCGC	-	254	73	regulation of primary metabolism	CPRF-3
Staf	M00262	949	68	cellular macromolecule metabolism	AGAACTACAA
CBF1	M00303	2566	39	lysosome organization	CAAAATGGCG

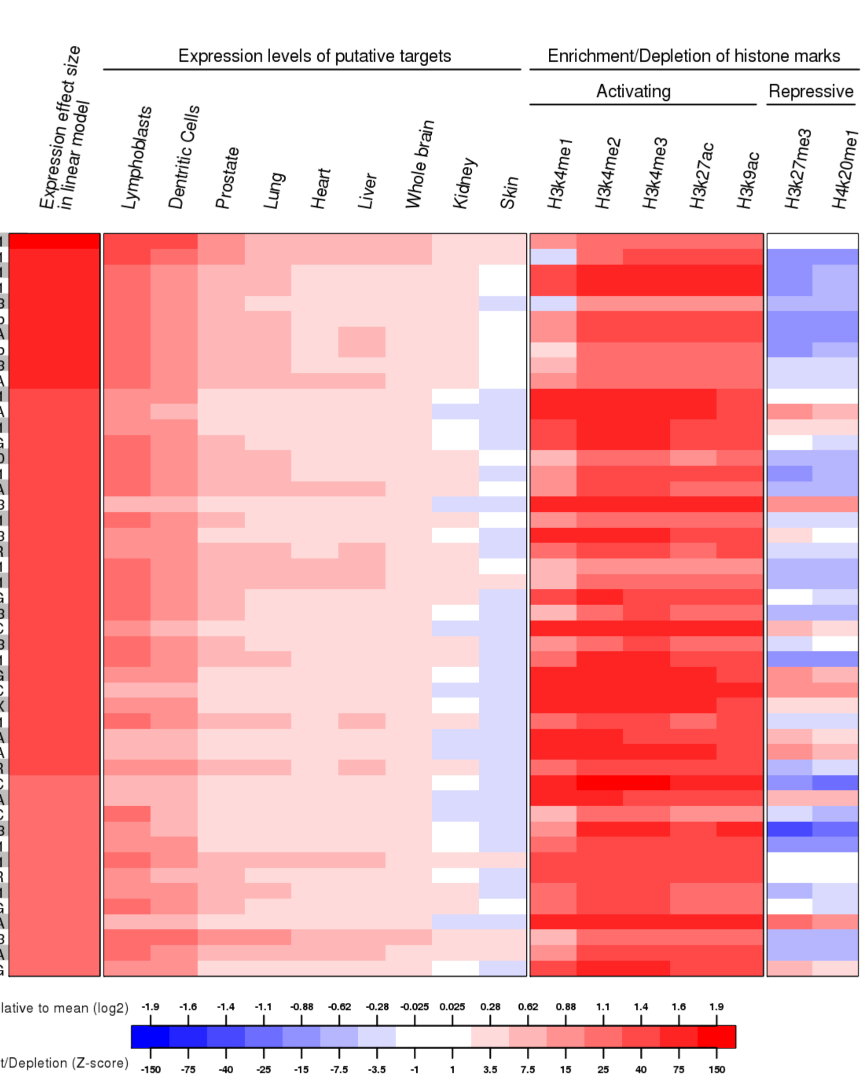
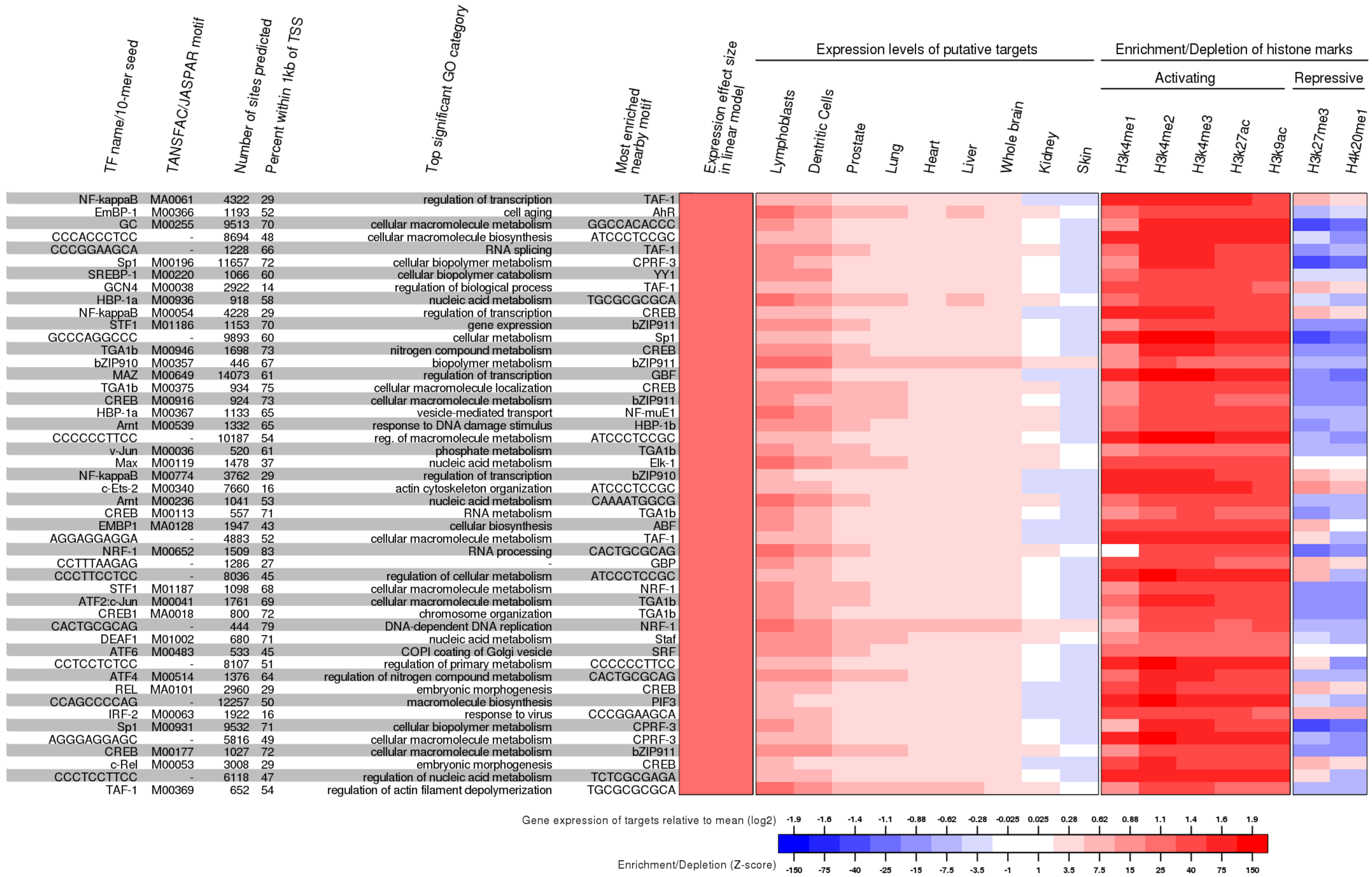


Figure S14: **Characteristics of the binding sites for 288 motifs.** See main text Figure 6 for description.*Continues...*



TF name/10-mer seed	TANSFAC/JASPAR motif	Number of sites predicted	Percent within 1kb of TSS	Top significant GO category	Most enriched nearby motif
NF-kappaB	MA0061	4322	29	regulation of transcription	TAF-1
EmBP-1	M00366	1193	52	cell aging	AhR
GC	M00255	9513	70	cellular macromolecule metabolism	GGCCACACCC
CCCACCCTCC	-	8694	48	cellular macromolecule biosynthesis	ATCCCTCCGC
CCCGGAAGCA	-	1228	66	RNA splicing	TAF-1
Sp1	M00196	11657	72	cellular biopolymer metabolism	CPRF-3
SREBP-1	M00220	1066	60	cellular biopolymer catabolism	YY1
GCN4	M00038	2922	14	regulation of biological process	TAF-1
HBP-1a	M00936	918	58	nucleic acid metabolism	TGCGCGCGCA
NF-kappaB	M00054	4228	29	regulation of transcription	CREB
STF1	M01186	1153	70	gene expression	bZIP911
GCCCAGGCCC	-	9893	60	cellular metabolism	Sp1
TGA1b	M00946	1698	73	nitrogen compound metabolism	CREB
bZIP910	M00357	446	67	biopolymer metabolism	bZIP911
MAZ	M00649	14073	61	regulation of transcription	GBF
TGA1b	M00375	934	75	cellular macromolecule localization	CREB
CREB	M00916	924	73	cellular macromolecule metabolism	bZIP911
HBP-1a	M00367	1133	65	vesicle-mediated transport	NF-muE1
Arnt	M00539	1332	65	response to DNA damage stimulus	HBP-1b
CCCCCCTTCC	-	10187	54	reg. of macromolecule metabolism	ATCCCTCCGC
v-Jun	M00036	520	61	phosphate metabolism	TGA1b
Max	M00119	1478	37	nucleic acid metabolism	Elk-1
NF-kappaB	M00774	3762	29	regulation of transcription	bZIP910
c-Ets-2	M00340	7660	16	actin cytoskeleton organization	ATCCCTCCGC
Arnt	M00236	1041	53	nucleic acid metabolism	CAAAATGGCG
CREB	M00113	557	71	RNA metabolism	TGA1b
EMBP1	MA0128	1947	43	cellular biosynthesis	ABF
AGGAGGAGGA	-	4883	52	cellular macromolecule metabolism	TAF-1
NRF-1	M00652	1509	83	RNA processing	CACTGCGCAG
CCTTTAAGAG	-	1286	27	-	GBP
CCCTTCCTCC	-	8036	45	regulation of cellular metabolism	ATCCCTCCGC
STF1	M01187	1098	68	cellular macromolecule metabolism	NRF-1
ATF2:c-Jun	M00041	1761	69	cellular macromolecule metabolism	TGA1b
CREB1	MA0018	800	72	chromosome organization	TGA1b
CACTGCGCAG	-	444	79	DNA-dependent DNA replication	NRF-1
DEAF1	M01002	680	71	nucleic acid metabolism	Staf
ATF6	M00483	533	45	COPI coating of Golgi vesicle	SRF
CCTCCTCTCC	-	8107	51	regulation of primary metabolism	CCCCCCTTCC
ATF4	M00514	1376	64	regulation of nitrogen compound metabolism	CACTGCGCAG
REL	MA0101	2960	29	embryonic morphogenesis	CREB
CCAGCCCCAG	-	12257	50	macromolecule biosynthesis	PIF3
IRF-2	M00063	1922	16	response to virus	CCCGGAAGCA
Sp1	M00931	9532	71	cellular biopolymer metabolism	CPRF-3
AGGGAGGAGC	-	5816	49	cellular macromolecule metabolism	CPRF-3
CREB	M00177	1027	72	cellular macromolecule metabolism	bZIP911
c-Rel	M00053	3008	29	embryonic morphogenesis	CREB
CCCTCCTTCC	-	6118	47	regulation of nucleic acid metabolism	TCTCGCGAGA
TAF-1	M00369	652	54	regulation of actin filament depolymerization	TGCGCGCGCA

Figure S14: ... **Characteristics of the binding sites for 288 motifs.**Continues...

TF name/10-mer seed	TANSFAC/JASPAR motif	Number of sites predicted	Percent within 1kb of TSS	Top significant GO category	Most enriched nearby motif
GCTCAGCCCC	-	7998	56	cellular macromolecule metabolism	CPRF-3
CCCTTTAAGA	-	1530	28	protein export from nucleus	ATF2
SRF	MA0083	639	21	organ development	CREB
c-Myc	M01145	601	38	negative regulation of cellular process	Elk-1
CKROX	M01175	12734	60	cellular macromolecule metabolism	CPRF-3
OSBZ8	M00654	897	73	cellular macromolecule metabolism	CPRF-3
USF	M00187	685	70	protein sumoylation	NF-muE1
HIF1	M00466	643	69	cellular biopolymer metabolism	c-Myc
CCTCCTCTTC	-	7337	45	cellular biopolymer biosynthesis	TAF-1
GGGGCAGGAA	-	11612	37	regulation of cellular metabolism	CPRF-1
CPRF-3	M00944	272	82	activation of caspase activity by cyt. c	bZIP911
ATF	M00338	1305	68	cellular macromolecule metabolism	GGCCCCTTTA
TAF-1	M00943	255	49	polyamine biosynthesis	CAAAATGGCG
CAGCCCCGGG	-	7366	64	cellular biopolymer metabolism	CREB
CGGGGAGGAA	-	3673	57	cellular biopolymer metabolism	ATCCCTCCGC
ATF3	M00513	1477	61	cellular macromolecule metabolism	TGA1b
AAACTACAAC	-	1177	28	embryonic development	Staf
Staf	MA0088	1877	52	cellular macromolecule metabolism	AGAACTACAA
CPRF-1	M00368	1226	39	regulation of epidermis development	AhR
CREB	M00178	1340	75	nucleic acid metabolism	TGCGCGCGCA
AAGGAGGAGG	-	7081	45	regulation of metabolism	ATCCCTCCGC
CCTCCTTCCC	-	7345	48	cellular macromolecule metabolism	ATCCCTCCGC
PU.1	M00658	8899	24	regulation of metabolism	CPRF-3
AGGAAGGGGC	-	11071	50	cellular biopolymer biosynthesis	GC
Sp1	M00933	10365	65	regulation of cellular metabolism	CPRF-3
AGAGGAGGAG	-	6628	46	cellular process	TCTCGCGAGA
CREB	M00039	1932	67	cellular macromolecule metabolism	GGCCCCTTTA
CPRF-3	M00370	616	72	-	NF-muE1
GAGAGGAGGA	-	6194	43	reg. of macromolecule biosynthesis	bZIP910
NF-kappaB	M00051	4086	45	transcription	CPRF-3
IRF-1	M00062	3135	13	regulation of transcription	CCCGGAAGCA
CTCCTGCCCC	-	5299	48	regulation of primary metabolism	CACTGCGCAG
E2F-4:DP-2	M00739	1704	75	cellular biopolymer metabolism	TGCGCGCGCA
CCCCTTTAAG	-	1279	35	regulation of transcription	TAF-1
CG1	M00440	758	59	regulation of cellular protein metabolism	bZIP911
AGGAAGGAGG	-	4496	40	regulation of cellular process	TCTCGCGAGA
GGAACTACAA	-	1034	41	RNA metabolism	Staf
AP-1	M00199	6000	15	regulation of biological process	TAF-1
E2F	M01114	2077	73	cellular biopolymer metabolism	TGCGCGCGCA
AGGAAGGGGG	-	8798	44	transcription	ATCCCTCCGC
Staf	M00264	1971	56	cellular biopolymer metabolism	GGAACTACAA
ATF	M00017	922	71	nucleic acid metabolism	TGA1b
CREB	M00917	974	70	nitrogen compound metabolism	TGA1b
TRAB1	M00507	1206	74	cellular biopolymer metabolism	OSBZ8
GGCCACACCC	-	3402	48	cellular metabolism	GC
NF-kappaB	M00194	4135	29	regulation of cellular process	bZIP910
SPL14	M01180	910	52	cellular process	YY1
ABF	M00442	567	51	water-soluble vitamin metabolism	bZIP910

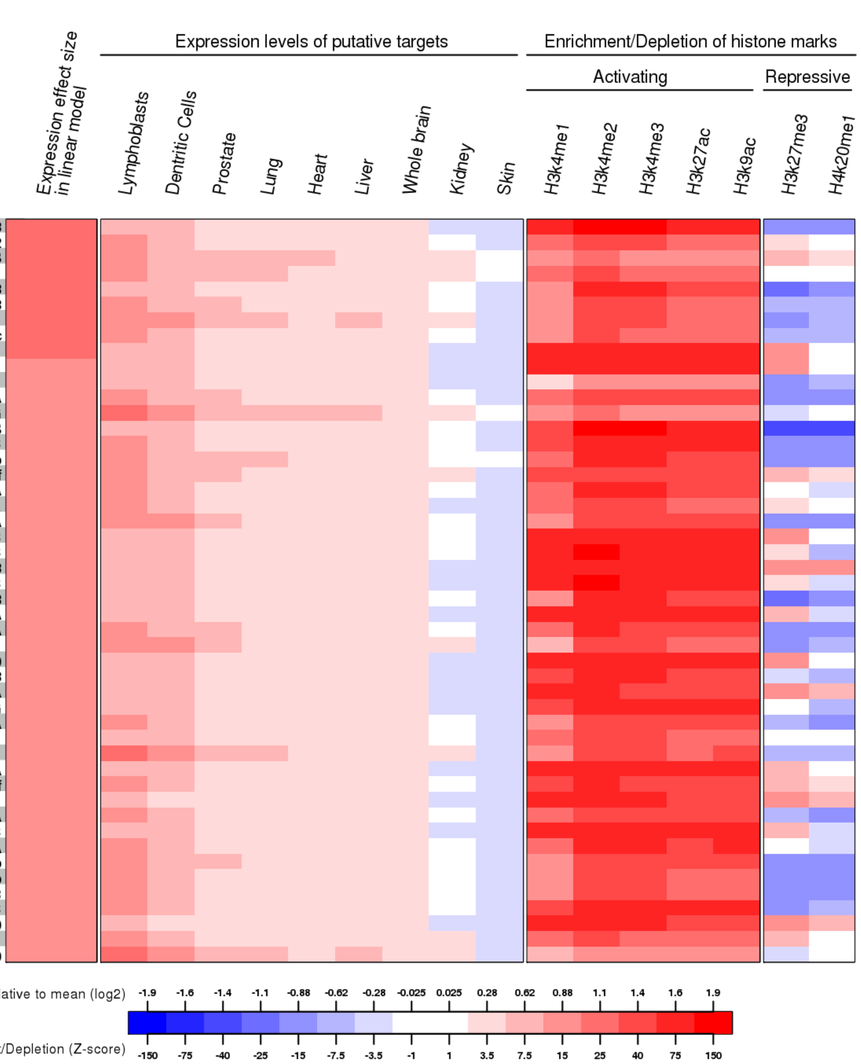


Figure S14: ... **Characteristics of the binding sites for 288 motifs.**Continues...

TF name/10-mer seed	TANSFAC/JASPAR motif	Number of sites predicted	Percent within 1kb of TSS	Top significant GO category	Most enriched nearby motif
GCCATGGCCC	-	4463	51	cellular macromolecule metabolism	GBP
ATF2	M00179	1199	72	nucleic acid metabolism	GGCCCCTTTA
Bach2	M00490	2269	15	GTPase mediated signal transduction	ATF3
AP-1	M00517	1994	13	biological regulation	TAF-1
RELA	MA0107	5240	30	gene expression	bZIP910
AP-1	M00188	2249	13	-	v-Jun
CREB,	M00981	831	74	nucleic acid metabolism	GCCCCTTTAA
AGGGAGGAAG	-	5070	36	cellular macromolecule metabolism	TAF-1
GGAGGAGAGA	-	5740	43	nucleic acid metabolism	TAF-1
NF-kappaB	M00208	3203	29	ureteric bud development	DEAF1
bZIP910	M00356	392	39	DNA metabolism	bZIP910
ESE1	M01214	9023	14	regulation of biological process	bZIP910
AGGCAGGAAG	-	4680	29	regulation of biological process	v-Jun
ATF1	M00691	939	60	cellular biopolymer metabolism	TGA1b
CPRF-1	M00942	926	41	cell volume homeostasis	AhR
Bach1	M00495	2848	9	GTPase mediated signal transduction	c-Myc
Dl_2	MA0023	4712	30	regulation of transcription	TAF-1
EmBP-1b	M00788	1291	38	reg. of macromolecule metabolism	CPRF-3
NF-kappaB	M00052	5901	31	regulation of gene expression	CREB
E2F-1:DP-2	M00737	1411	72	cellular biopolymer metabolism	TAF-1
P50:P50	M01223	3361	31	regulation of biological process	bZIP910
HES1	M01009	1474	57	nucleic acid metabolism	TGCGCGCGCA
Pax-5	M00143	3219	59	cortical actin cytoskeleton organization	HIF1
CCTTCCTCCC	-	6224	47	regulation of primary metabolism	ATCCCTCCGC
bZIP911	M00358	301	51	dNMP metabolism	bZIP910
AGAAGGAGGA	-	5076	37	cellular macromolecule metabolism	TCTCGCGAGA
P50:RELA-P65	M01224	3982	29	RNA biosynthesis	AhR
AGAACTACAA	-	999	32	-	Staf
GCCCCTTTAA	-	1122	36	regulation of metabolism	CREB
Nrf-2	M00821	4049	16	ER-nuclear signaling pathway	CPRF-3
MAF	M00983	1976	15	cell aging	CPRF-3
GGCCCTTTAA	-	1073	32	regulation of nucleic acid metabolism	CREB
GCN4	M00204	7578	13	Ras protein signal transduction	TAF-1
LXR	M00647	1018	42	response to retinoic acid	Elk-1
Pax5	MA0014	2028	62	macromolecule metabolism	bZIP910
CCCTTTAAGG	-	1232	33	regulation of metabolism	TAF-1
AML	M00769	4848	13	muscle contraction	CPRF-1
alpha-CP1	M00687	2034	46	reg. of macromolecule biosynthesis	CREB
Ebox	M01034	300	67	nucleic acid metabolism	YY1
HAP2/3/4	M00288	2085	51	nucleosome assembly	CACTGCGCAG
core-binding	M00722	9798	11	regulation of metabolism	bZIP911
Sp3	M00665	4464	58	regulation of metabolism	CPRF-3
AP-2alphaA	M01047	2790	64	regulation of transcription	TAF-1
BLIMP1	M01066	8298	24	regulation of cellular biosynthesis	TCTCGCGAGA
HEN1	M00068	3539	43	phasic smooth muscle contraction	CF1
NF-Y	M00209	2070	37	DNA packaging	CPRF-3
dl_1	MA0022	2911	33	nucleic acid metabolism	bZIP910
Pax-9	M00329	2834	45	negative regulation of cellular biosynthesis	v-Jun

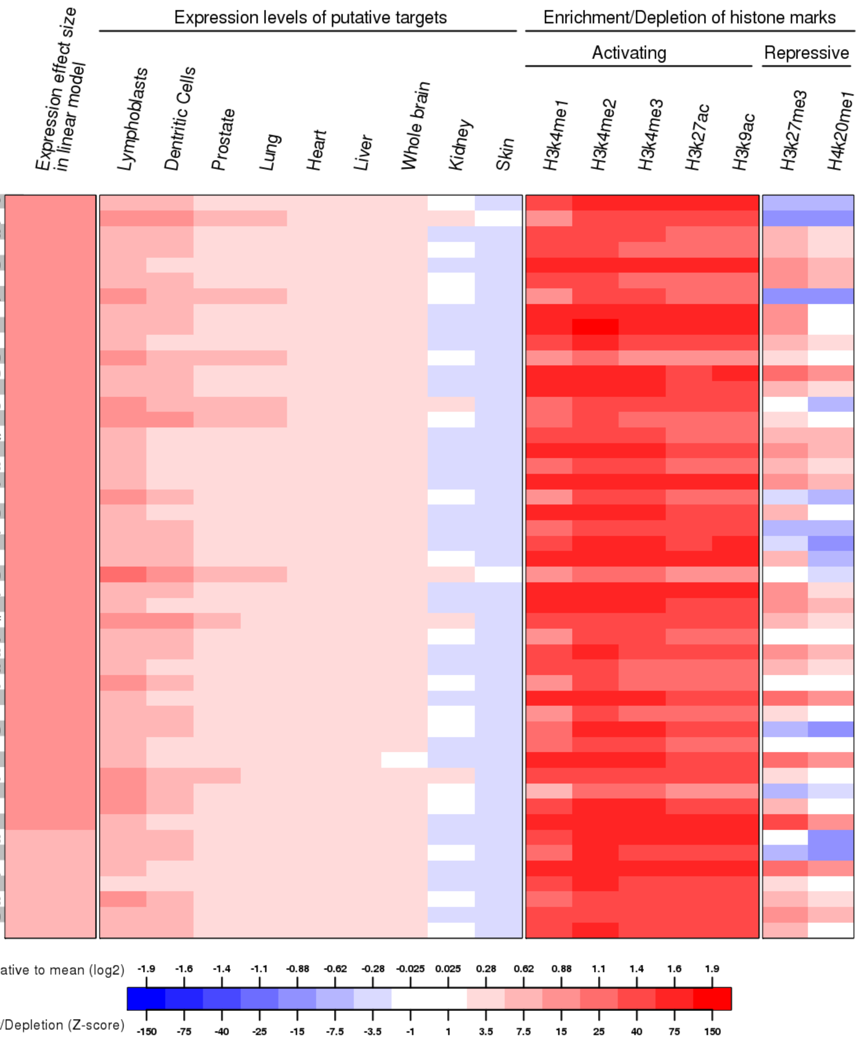


Figure S14: ... **Characteristics of the binding sites for 288 motifs.**Continues...

TF name/10-mer seed	TANSFAC/JASPAR motif	Number of sites predicted	Percent within 1kb of TSS	Top significant GO category	Most enriched nearby motif
Nkx2-5	M01043	6463	22	transcription	TAF-1
KROX	M00982	8797	61	regulation of primary metabolism	CPRF-3
HEN1	M00058	3389	45	regulation of transcription	TAF-1
NF-E2	M00037	3671	13	endoderm development	TAF-1
UF1H3BETA	M01068	3606	45	cellular macromolecule metabolism	CPRF-3
PEBP	M00984	4316	15	cell motion	bZIP910
Adf-1	M00171	6741	78	reg. of macromolecule metabolism	CPRF-3
TGCGCGCGCA	-	664	80	macromolecule metabolism	bZIP911
Octamer	M00795	4775	12	chromatin assembly or disassembly	ATCCCTCCGC
MZF1_5-13	MA0057	1060	44	RNA biosynthesis	ATF2
SRF	M00186	1983	23	regulation of RNA metabolism	CPRF-3
MZF1	M00084	4243	47	regulation of gene expression	ATF2
Adf-1	M00923	4969	81	cellular biopolymer metabolism	CPRF-3
DEC	M00997	3060	39	reg. of macromolecule metabolism	CPRF-3
HNF4alpha1	M00411	2803	35	regulation of cellular process	ATF6
PCF2	M00948	3094	48	transcription	c-Myc
NHLH1	MA0048	4728	50	cellular biopolymer metabolism	TAF-1
SRF	M00922	2160	21	actin filament-based movement	CACTGCGCAG
GGCCCCTTTA	-	883	41	microtubule cytoskeleton polarity	ATF2
COUPTF	M01036	6482	46	regulation of RNA metabolism	CPRF-1
Muscle	M00323	3303	48	regulation of transcription	ATF2
SRF	M00215	2117	19	actin filament-based process	CPRF-3
Ttk	M00009	16109	44	regulation of metabolism	Pax-5
FACB	M00390	1970	62	negative regulation of transcription	TGCGCGCGCA
NFYA	MA0060	3684	64	reg. of macromolecule metabolism	YY1
AhR	M00139	713	61	RNA metabolism	GBP
AP-2alphaA	M01045	3764	58	cellular process	Ebox
SP1:SP3	M01219	10641	65	regulation of primary metabolism	CPRF-3
AP-4	M00005	5594	37	biological regulation	TCTCGCGAGA
SRF	M01007	2426	20	actin filament-based movement	TAF-1
PPAR,	M00762	2919	31	-	TCTCGCGAGA
MEF-2	M00941	8487	12	regulation of nucleic acid metabolism	bZIP911
ACAAT	M00309	1446	45	reg. of macromolecule biosynthesis	ATCCCTCCGC
NF-Y	M00287	3737	64	reg. of macromolecule metabolism	YY1
MTF-1	M00650	3027	60	RNA biosynthesis	TGCGCGCGCA
RFX1	M00280	1524	32	negative regulation of cellular biosynthesis	RFX1
GGGAGGAGGA	-	3278	34	regulation of primary metabolism	Adf-1
MAZR	M00491	3930	40	regulation of metabolism	YY1
ABI4	M00958	1938	70	cellular biopolymer metabolism	CPRF-3
SMAD4	M00733	4790	40	regulation of metabolism	TGCGCGCGCA
CCAAT	M00254	1810	46	nucleosome assembly	YY1
OCT-x	M00210	5696	12	regulation of transcription	ATCCCTCCGC
GGGAAACTGA	-	2867	17	peptide hormone processing	TCTCGCGAGA
Muscle	M00324	2640	45	cellular macromolecule biosynthesis	GBP
FACB	M00388	2944	47	regulation of transcription	TAF-1
TFII-I	M00706	5900	29	regulation of cellular metabolism	YY1
HEB	M00698	17067	32	regulation of RNA metabolism	bZIP911
PacC	M00247	5481	49	biological regulation	TGCGCGCGCA

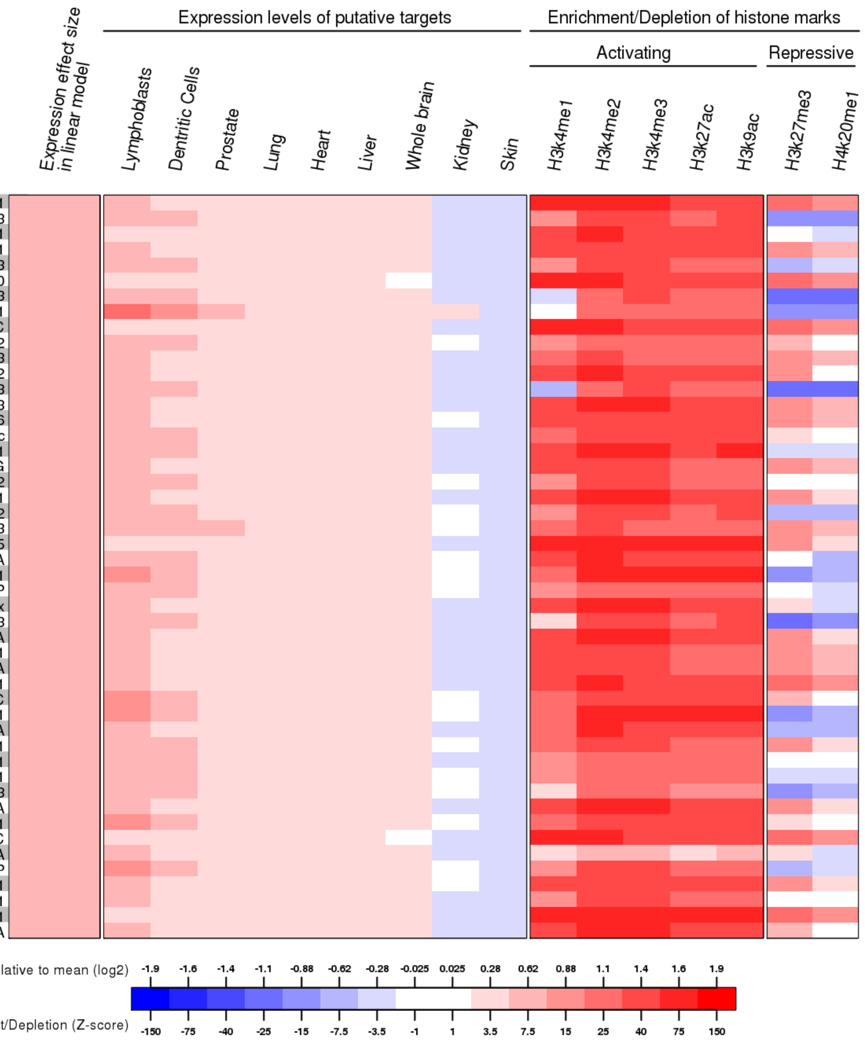


Figure S14: ... **Characteristics of the binding sites for 288 motifs.**Continues...

TF name/10-mer seed	TANSFAC/JASPAR motif	Number of sites predicted	Percent within 1kb of TSS	Top significant GO category	Most enriched nearby motif
Muscle	M00321	3689	56	regulation of cellular metabolism	TGCGCGCGCA
E2A	M00804	7150	37	nervous system development	CPRF-3
TAL1	M00993	6088	17	regulation of phosphorus metabolism	CPRF-3
repressor	M00014	1503	70	regulation of gene expression	TGCGCGCGCA
Oct-1	M00342	7104	12	regulation of transcription	ATCCCTCCGC
Egr	M00807	3894	57	cellular biopolymer metabolism	ATF2
GCR1	M00046	4763	34	regulation of transcription	Elk-1
CP2	M00072	5919	40	reg. of cyclin-dependent protein kinase activity	CPRF-3
HNF4	M01031	4118	32	regulation of RNA metabolism	ATCCCTCCGC
CF1	M00111	857	43	regulation of metabolism	bZIP911
RSRFC4	M00407	4310	11	regulation of transcription	HBP-1b
ESR1	MA0112	3961	42	intracellular signaling cascade	TGA1b
CACCACCACC	-	10458	32	regulation of transcription	Elk-1
AP-2	M00800	2802	64	regulation of RNA metabolism	Adf-1
NF-Y	M00775	2414	51	nucleosome assembly	RFX1
bHLH66	M01054	1082	46	ear development	AhR
MEF-3	M00319	3918	25	RNA biosynthesis	TAF-1
Alfin1	M00479	4958	33	regulation of nucleic acid metabolism	CACTGCGCAG
Lmo2	M00277	3890	43	transmembrane receptor PTK signaling	GBP
RSRFC4	M00026	4184	11	regulation of transcription	E4F1
GCM	M00634	1863	39	ear development	bZIP911
MIF-1	M00279	909	41	oxaloacetate metabolism	RFX1
MEF-2	M00232	3511	10	energy reserve metabolism	TCTCGCGAGA
HNF4	M00764	4979	33	macromolecular complex disassembly	NRSE
E12	M00693	3604	20	MAPKKK cascade	TAF-1
MEF2A	MA0052	5152	11	regulation of transcription	HBP-1b
Olf-1	M00261	5312	38	regulation of cellular metabolism	AhR
SP1	MA0079	1229	54	cellular process	TAF-1
Myf	MA0055	13685	36	negative regulation of cellular process	CPRF-3
ZNF219	M01122	5464	46	regulation of primary metabolism	CPRF-3
AP-2	M00915	3112	59	Ras protein signal transduction	TAF-1
STAT5A	M00457	7003	25	regulation of transcription	Staf
usp	MA0016	5229	43	regulation of cellular metabolism	bZIP911
MEF-2	M00231	5273	13	muscle contraction	HBP-1b
LMAF	M01139	9390	25	embryonic development	c-Myc:Max
NF-Y	M00185	2197	35	DNA packaging	HBP-1b
MyoD	M00001	3399	20	phosphagen biosynthesis	CREB1
PCF5	M00952	2088	60	regulation of metabolism	bZIP911
VDR,	M00966	3544	39	lipid biosynthesis	Elk-1
AREB6	M00414	6119	37	regulation of cellular process	LXR
CF1	M00112	4221	38	regulation of primary metabolism	bZIP911
RFX1	M00281	1839	38	-	CACTGCGCAG
SF1	M00727	9842	29	dendrite development	Elk-1
CTCF	M01200	14895	8	multicellular organismal development	ATCCCTCCGC
REST	MA0138	1337	24	secretion	bZIP910
NRSF	M01028	1103	29	neurological system process	CPRF-3
NRSE	M00325	655	18	synaptic transmission	bZIP910
NRSF	M00256	721	15	synaptic transmission	CPRF-3

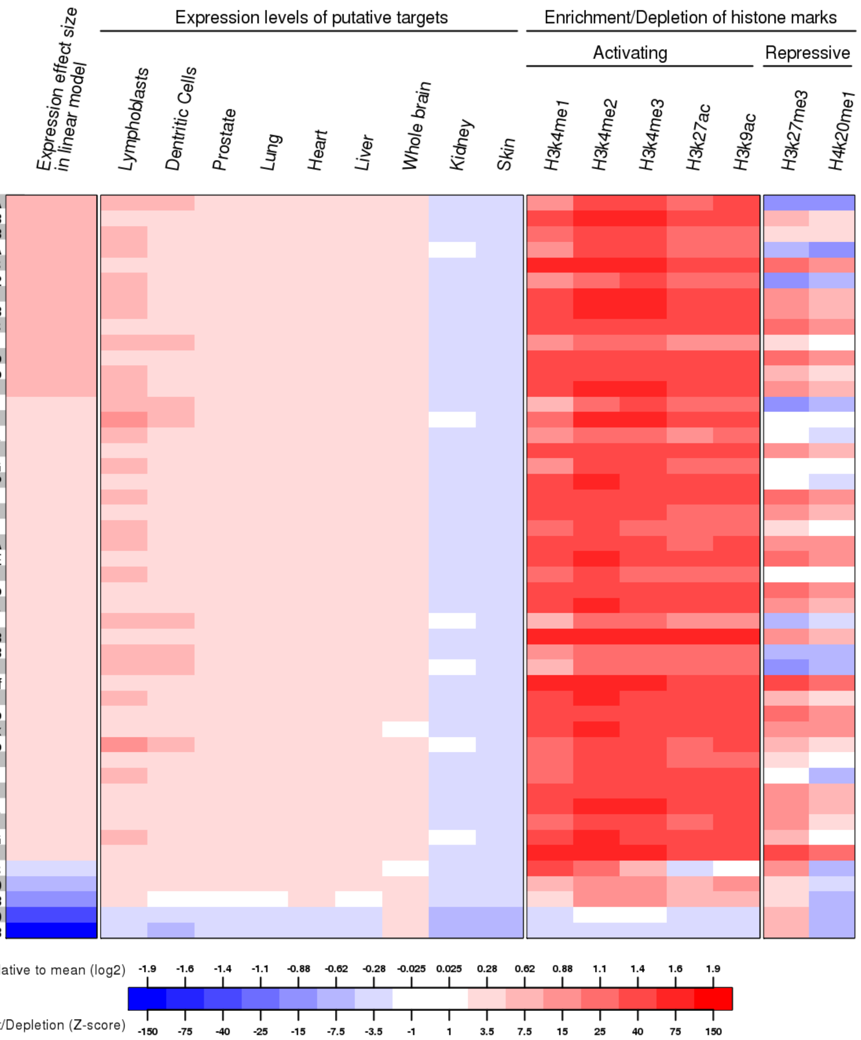
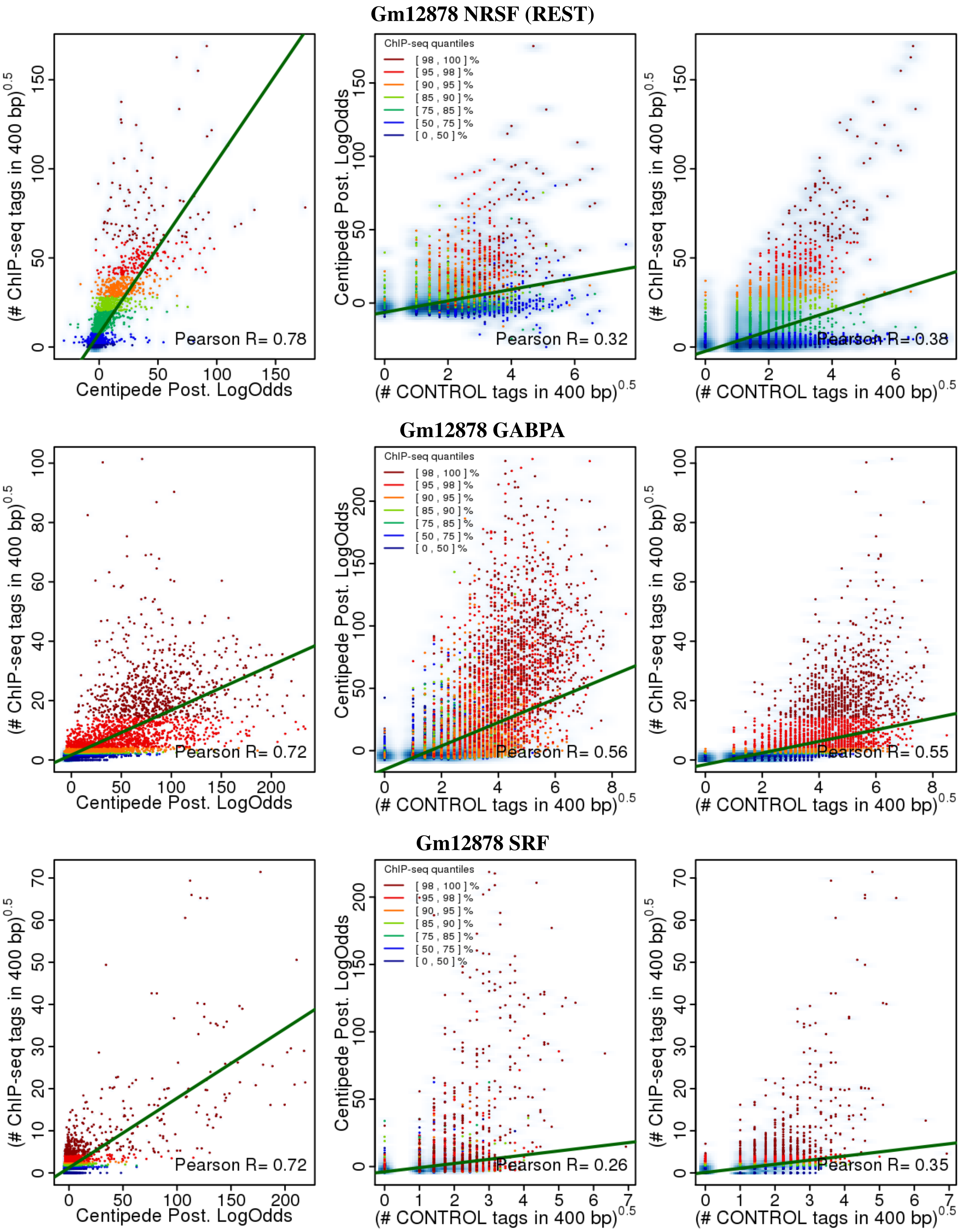


Figure S14: ... **Characteristics of the binding sites for 288 motifs.**

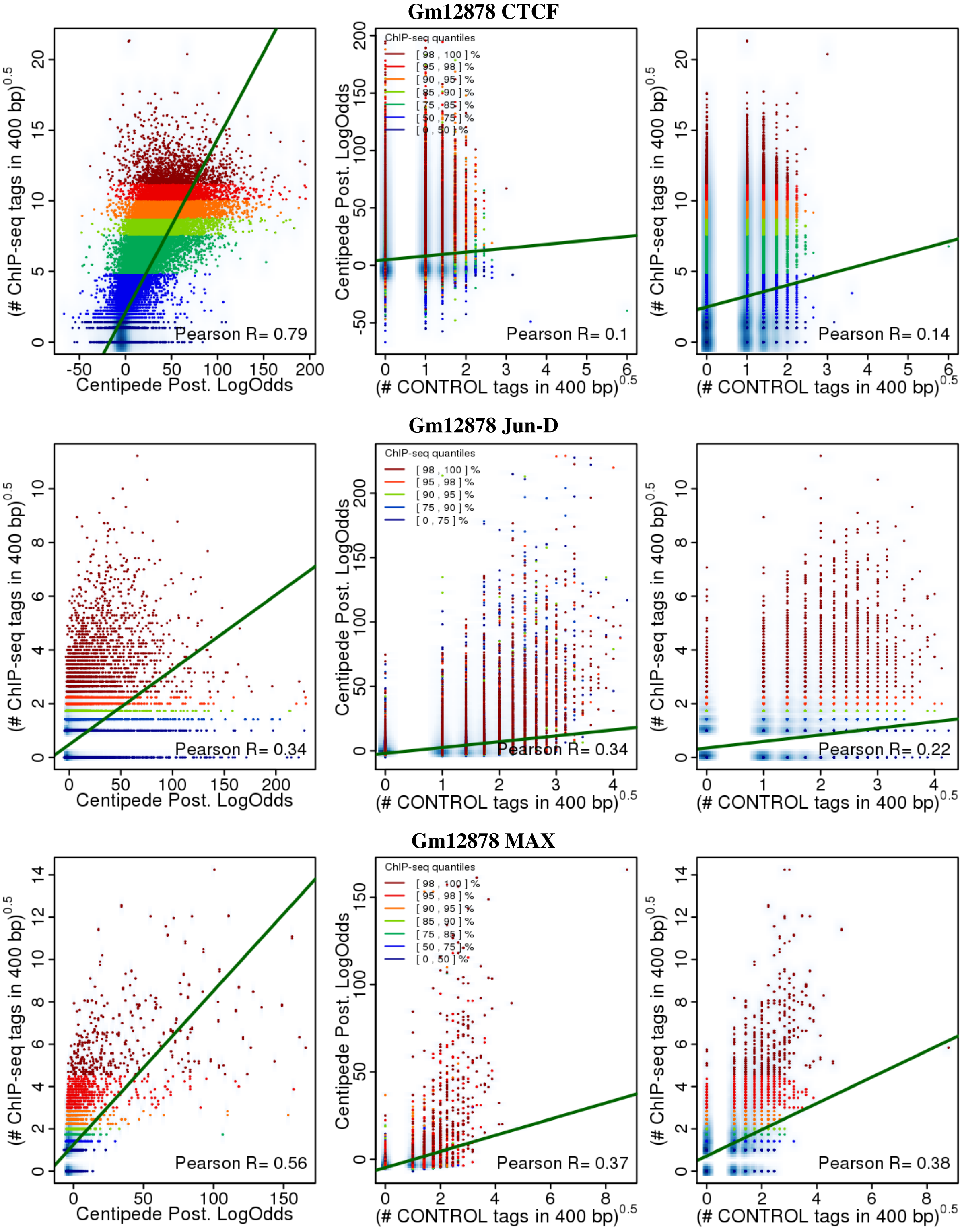


Gm12878 NRSF (REST)

Gm12878 GABPA

Gm12878 SRF

Figure S15: **Correlation of the CENTIPEDE posterior log-odds with ChIP-seq**. *Continues...*



Gm12878 CTCF

Gm12878 Jun-D

Gm12878 MAX

Figure S15: ... **Correlation of the CENTIPEDE posterior log-odds with ChIP-seq**