An Interaction-Dependent Model for Transcription Factor Binding

Li-San Wang[1], Shane T. Jensen[2], and Sridhar Hannenhalli[3,*]

[1] Department of Biology
lswang@mail.med.upenn.edu
[2] Department of Statistics, The Wharton School
stjensen@wharton.upenn.edu
[3] Penn Center for Bioinformatics and Department of Genetics
sridharh@pcbi.upenn.edu
University of Pennsylvania
Philadelphia, PA 19104-6021

Abstract. Transcriptional regulation is accomplished by several transcription factor proteins that bind to specific DNA elements in the relative vicinity of the gene, and interact with each other and with Polymerase enzyme. Thus the determination of transcription factor-DNA binding is an important step toward understanding transcriptional regulation. An effective way to experimentally determine the genomic regions bound by a transcription factor is by a ChIP-on-chip assay. Then, given the putative genomic regions, computational motif finding algorithms are applied to estimate the DNA binding motif or positional weight matrix for the TF. The *a priori* expectation is that the presence or absence of the estimated motif in a promoter should be a good indicator of the binding of the TF to that promoter. This association between the presence of the transcription factor motif and its binding is however weak in a majority of cases where the whole genome ChIP experiments have been performed. One possible reason for this is that the DNA binding of a particular transcription factor depends not only on its own motif, but also on synergistic or antagonistic action of neighboring motifs for other transcription factors. We believe that modeling this interaction-dependent binding with linear regression can better explain the observed binding data. We assess this hypothesis based on the whole genome ChIP-on-chip data for Yeast. The derived interactions are largely consistent with previous results that combine ChIP-on-chip data with expression data. We additionally apply our method to determine interacting partners for CREB and validate our findings based on published experimental results.

1 Introduction

Gene transcription in eukaryotes is tightly regulated, both spatially and temporally, by several transcription factors (TF) [1, 2]. Thus, a first step in computational analysis of transcription is to model the DNA binding preference of the TFs, which can be done using a set of experimentally determined DNA sequences bound by a TF. Traditionally this was either done using *in vitro* SELEX experiment [3] or

---

* Corresponding author.

footprinting assay [4] to determine binding sites followed by specific mutagenesis of the binding site to determine the acceptable bases. Either of these technologies result in short sequences, which can be aligned to derive the binding specificity as a *positional weight matrix (PWM)* [5]. A recent high throughput technique to determine the genomic regions bound by a specific transcription factor is *Chromatin Immunoprecipitation (ChIP)* experiments [6]. Although these regions are large (few hundred bases), computational motif discovery algorithms can be applied to the bound regions in order to detect the most likely PWM for the TF [5].

Critical to transcriptional regulation, is the interaction among the TFs, which in turn depends on TFs binding to specific *cis* elements in relative vicinity. This interaction-dependent functionality is the basis for *transcriptional modules* [7-9]. In some cases, this dependency is involved in protein modifications, like phosphorylation. However, there are other cases where the binding of the TF itself is interaction-dependent [10, 11]. This interaction-dependent binding often results in inconsistency between binding specificity of a factor and experimentally determined binding sites for the factor. For example, transcription factor CREB is believed to bind to the cAMP response element (CRE), which has motif CGTCA. However, based on genome wide Chromatin Immunoprecipitation (ChIP) assay, 70% of CREB-bound regions do not contain the CGTCA motif in the ~800 bps region [12]. More generally, we have found that for a majority of factors for which genome-wide ChIP-chip assay has been performed, the derived PWM does not sufficiently explain the experimental binding data.

We investigate whether, modelling the binding of a TF as an interaction-dependent mechanism can lead to better predictions of the binding of a TF to a promoter region. We model the binding probability of a TF as a linear combination of PWM scores corresponding to that TF as well as other potentially interacting TF. Similar linear model was used to model the regulation of a gene by the promoter motifs in [13]. We optimize the coefficients using linear regression. The estimated coefficients of this linear combination indicate the degree of dependence between the TFs and the sign of the coefficients indicate whether the interaction is synergistic or antagonistic. Our interaction-dependent binding model better estimates the experimentally determined binding probabilities and also reveals specific synergistic and antagonistic interactions among factors in yeast. We compare these detected interactions with previous methods [9, 14] and discuss the novel ones. We further apply our method to rat genomic regions bound by CREB to determine its interacting factors. Most of the detected CREB partners are validated using published experimental literature.

2   Results

We use the ChIP-chip data for 204 transcription factors on 6229 genes in yeast [15]. Of these 204 TFs, motif or a PWM derived from the bound promoters is provided for 102 TFs. We evaluate our method on 90 of these transcription factors (see methods). Using these motifs, we estimate the dependence of a factor on other factors for its binding to a promoter. The dependence coefficients are computed using linear regression so as to minimize the overall difference between the experimental binding probability and estimated binding probability. We use the average sum of squared errors (SSE – see methods) to compare the relative

advantage of using interaction-dependent model in estimating the binding probabilities. For each TF, we consider r genes for which the binding p-value is less than 0.001, as suggested in [15]. We also include r genes with the largest p-values as well as *2r* randomly selected genes from among the remaining genes. This set of *4r* genes is split into *training*, *model selection, and test* sets described later.

Improved estimate of binding probability in yeast using pair-wise interactions. The best pair-wise interaction regression model is created for each of the 58 TFs where the number of binding genes is at least 20 ($r = 20$), so that the regression model has at least 80 observations (r bound, r unbound and *2r* random genes; see methods). For several TFs, a model without any interaction terms (the *interaction-free* model) leads to small prediction errors on its own. These TFs may bind in an interaction-independent fashion and thus do not require a more detailed model. Thus, if we consider only the TFs which have "high" sum of squared errors (SSE) assuming interaction-independent binding, we would expect a greater fraction of these TFs to have an improved SSE when interaction is used. As shown in Fig. 1, the binding data for many TFs is better modelled by including interactions between TFs and the fraction of improved TFs increases with the SSE of interaction-independent binding. As a negative control we calculated the interaction-dependent SSE for a randomly chosen interacting TF (shown by the striped bars in Fig. 1). The fractions of TFs that are helped by including interactions is much greater than the fraction when we randomly select the interacting partner instead of the best possible interacting partner (based on our model selection).
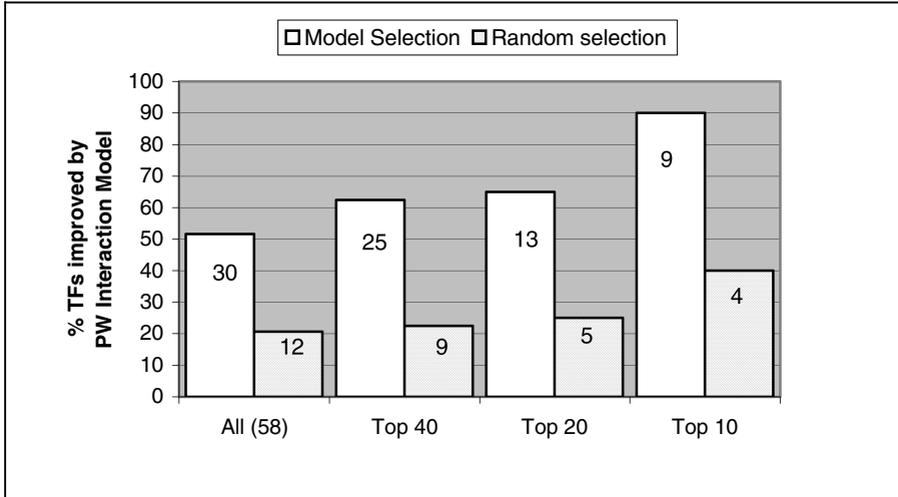

Fig. 1. Percent improved in bins using pair-wise interactions. Using the top X (X=58, 40, 20,10) of all 58 TFs sorted in decreasing order of their SSE with non-interacting models, we calculate the fraction of TFs where the SSE is smaller using the pair-wise interaction model than that for the non-interaction model. We use either model selection or random selection as a control to determine which pair of TFs interact (see methods section). On each bar is the number of TFs improved.

Using three-way interactions. We also investigated whether the model for some TFs can be additionally improved by extending the number of allowed interactions and the se results are shown in Fig. 2. The fraction of TFs for which the three-way interaction model is a better fit to the experimental data is reduced compared with the pair-wise case. It is however, still significantly greater relative to the case when the interacting partners are selected randomly.
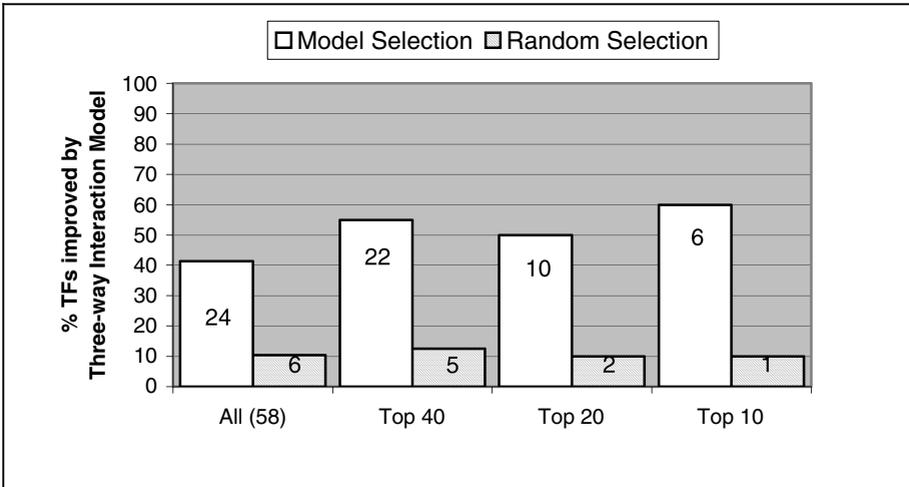

Fig. 2. Percent improved in bins using three-way interactions. Using the top X (X=58, 40, 20,10) of all 58 TFs sorted in decreasing order of their SSE with non-interacting models, we calculate the fraction of TFs where the SSE is smaller using the pair-wise interaction model than that for the non-interaction model. We use either model selection or random selection as a control to determine which pair of TFs interact (see methods section). On each bar is the number of TFs improved.

Identification of interactions using entire data. The interaction between transcription factors are represented in our model by regression coefficients. Positive coefficients imply *synergistic* interactions and the negative coefficients imply *antagonistic* interactions. By applying a FDR threshold of 0.1 to these coefficients, we obtained 377 synergistic interactions and 68 antagonistic interactions.

Yeast cell cycle. We compared our finding with the combinations of TFs regulating the stages of cell cycle that was reported in [9], based on a combination of ChIP-on-chip and conditional expression data. Excluding SKN4, which was excluded from our analysis because of a small number of gene it bind to, there are 18 known interactions among 11 factors. Using our 377 predicted synergistic interactions, which represents 9.4% of all possible interactions among 90 TFs used in our analysis, we predict 34 (62%) of the possible 55 interactions among TFs in the cell cycle, and 16 (89%) of the 18 known interactions. The p-values of our interaction coefficients are significantly lower for the 18 know interactions compared with the other pairs of cell cycle TFs. Our approach accurately identifies (with ~89% sensitivity) the known interactions in

cell cycle regulation at a low overall prediction rate of 9.4%. Cell cycle factors were not included in any of the 68 antagonistic interactions predicted by our method.
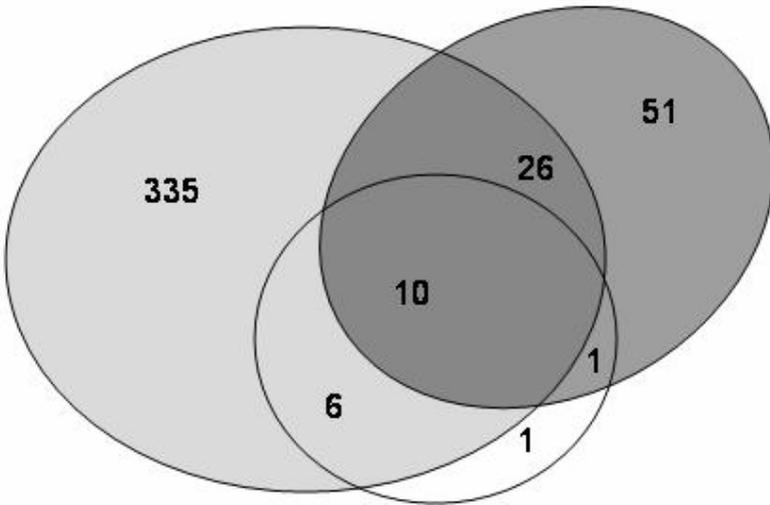

Fig. 3. Comparing three sets of interactions. White ellipse: 18 cell cycle interactions, gray ellipse: 377 regression based predictions, Dark ellipse: 98 interactions from [14].

Comparison with other previously predicted interactions. Banerjee et al. used *expression coherence* to predict cooperativity between transcription factors [14]. The hypothesis being: if the genes targeted (based to ChiP data) both by factors x and y have more similar expression profiles relative to the genes targeted by either x or y, but not both, then x and y are likely to cooperate. The same idea was earlier used in [16]. We took the 193 predicted pairs at p-value of 0.01 in [14]. Filtering out the TFs not represented in our set of 90 factors results in 98 predicted pairs. Fig. 3 shows the comparison of three sets, (1) Our 377 predicted synergistic interactions, (2) 98 interactions from [14], and (3) 18 cell cycle interactions from [9]. Banerjee et al. identify 11 (61%) of the cell cycle interactions, compared with our 89%, while predicting far fewer interaction (98 versus 377). However it should be noted that they used both the ChIP data and the expression data as did Lee et al. in inferring the cell cycle interactions while we have relied solely on ChIP data. Clearly, it is desirable to use additional, functional and expression information, to predict TF-TF interactions. Our aim however, is to predict the dependencies strictly in TF-DNA binding. TF-DNA binding may not correspond to expression of the target gene since it depends on post-translational modification of the transcription factors. Our result, based solely on the ChIP data, is largely consistent with previously known interactions and proposes several novel interactions.

Antagonism. An important aspect of our regression-modeling approach is that we detected 68 pairs of antagonistic interactions, in addition to our predicted synergistic interactions. Unfortunately, there is no experimental data on TF antagonism that we

could use to validate our findings, so we instead used the following approach to indirectly test our findings. If TF x antagonizes TF y, then for all the promoters which contain a motif for y, we should see an inverse correlation between the presence of motif x and the binding of TF y. In other words, the presence of motif x should repress the binding of TF y. For each such pair (x y), we first selected the top 100 genes with strongest motif hits for TF y. For these 100 genes we computed the Pearson's correlation coefficient between 100 occupancy probabilities for TF x and the 100 ChIP binding values for TF y. We computed the correlation for the antagonistic pairs and 500 randomly selected pairs. The average correlation coefficient for the random pairs is 0.004, whereas that for the antagonistic pairs is -0.05. Although this difference is small, it is significant with a p-value of 0.0003 based on t-test and with a p-value of 0.003 based on Kolmogorov-Smirnov test.

Detection of CREB binding partners. As the positive set for CREB bound regions we extracted the ±500 bp regions flanking the 5948 high-confidence CREB bound locations in rat based on SACO technology [17]. We additionally extracted 5000 random 1 kb regions from the rat genome. Similarly to Yeast analysis, we computationally determined the occupancy probability for each of the 546 vertebrate PWMs from TRANSFAC 8.4 against each of these 10948 sequences. Unlike the Yeast ChIP-chip data, in this case, we do not have an experimental binding p-value for CREB against these sequences. Instead we use a value of 1 for the positive sequences and a value of 0 for the background sequences. Based on the pair-wise regression, using FDR threshold of 0.05, ie. 5% false discovery rate, we discovered 18 TFs that synergistically interact with CREB. Using the published literature, 10 of these have evidence for interaction with CREB, majority of which are via the CREB co-activator CBP. These are TBP [18], STAT1 [19], GR [20], GATA-1 [21], p53 [22], Sp-1 [23], AhR (via Arnt and CBP) [24], v-Myb [25], SMAD-3 [26], and Major_T-antigen [27]. Additionally MEIS1 has a CBP-dependent Protein-Kinase-A response domain [28]. For additional 4 factors, we could not find direct evidence in the literature. However the DNA binding specificities for these factors are very similar (details excluded) to another factor for which there was evidence for interaction with CREB. In terms of DNA binding specificity, Ncx is similar to STAT1, GEN_INI to CREB, and MAZR to Sp-1, and AP-4 to MyoD. For 3 of the 18 factors – Pax, ZF5, and XPF-1, we could not find any literature evidence of interaction with CREB.

3   Methods

Data. We use the genome scale ChIP data in Yeast for 204 transcription factors for 6229 genes reported in [15]. The binding data is expressed as a p-value for the null hypothesis of no binding. The lower the p-value for a gene, the more confident we are that there is binding between the gene and the TF. By "gene" we implicitly mean the upstream 700 base pairs of the gene which was used to design the microarray in [15]. Using an extensive set of motif detection algorithms and a supervised integration of their results, the authors also provide the log-odds matrix or positional weight matrix (PWM) for 102 of these factors, some of which are identical (for example, MET31 and MET32). Using identical PWMs in our regression model would lead to identical

coefficients, so our analysis focused on a subset of 90 of these 102 transcription factors that have unique PWMs. For a factor t and gene g represent the binding probability as *ChIP(t,g)*. Accordingly we extracted these regions using the yeast genome sequence as well as the gene annotation from the SGD database at Stanford University (ftp://genome-ftp.stanford.edu/). The ChIP(t,g) p-value was transformed to a larger and more continuous scale by the transformation log(0.001 / ChIP(t,g)). This transformation has been previously shown to be effective in [29].

Occupancy probability of a PWM in a gene promoter. Positional Weight matrix (PWM) is a 4 by w matrix representing the DNA binding specificity of a TF that binds to a w bases long DNA string. For a PWM M, the entry M_{ib} at position i for base b is the $-ln(P_{ib}/Q_b)$, where P_{ib} is the frequency of base b at position i among the known binding sites and Q_b is the background frequency of base b [5]. Given the PWM M and a sequence $s = (s_1, s_2, \ldots, s_w)$, the raw score r of M on s is computed as

$$r(M, s) = \sum_{i=1}^{w} M_{is_i} ,$$ where M_{ib} is the entry in M at position i for base b. The transformed score $x(M,s) = e^{r(M,s)}$ is proportional to the binding energy of the TF to the sequence. The score of M on a 700 bp gene promoter p:

$$x(M, p) = \sum_{s \in P} x(M, s)$$ where s represents all w-long substrings in p in either strand. Let $Z = \sum_{p \in P} x(M, p)$ where P represents the universe of "all" promoters in the genome to which M has access to. The probability that M occupies a promoter p is estimated as *x(M,p)/Z*. However only a small fraction of all promoters are actually accessible to the TF for binding at any given time and we should use only the sum of scores over this subset for normalization. The true occupancy probability is hard to compute as it depends on several other attributes, like the state of the chromatin, concentration of the TF, binding affinity of the TF, genome composition etc. Thus the number we compute serves as a proxy for the occupancy probability.

Modeling the TF-GENE binding using linear regression and model selection. We use the following regression model:

$$Y_{ij} = \mu_j + a_j x_{ij} + \sum_{k \in R_j} b_{jk} x_{ik} + \varepsilon_{ij}$$

Here R_j is a set of TFs (excluding TF j itself) interacting with TF j that will be determined using an iterative model selection procedure detailed below, Y_{ij}=log $(0.001/p_{ij})$, where p_{ij} is the ChIP-chip based p-value of binding for TF j and gene i, x_{ij} is the occupancy probability of TF j to the upstream region of gene i, μ_j, a_j, and b_{jk} are regression coefficients, and ε_{ij} is the normally distributed error term. We use ordinary least squares regression to solve for the coefficients and their significance.

For each TF, we consider r genes for which the binding p-value ≤ 0.001, as suggested in [15]. We also include r genes with largest p-values (least likely to bind) as well as $2r$ randomly selected genes from among the remaining genes. We run validation tests 10 times for each TF. For each validation test, the set of $4r$ genes is randomly split into *training* (50%), *model selection* (25%), and *test sets* (25%). We run the following procedure to determine set R_j, the coefficients, and the error of the regression. Initially we set R_j to be empty. At each iteration we augment R_j by one

new TF as follows. For each TF k' not in R_j, set $R_{jk'} = R_j \cup \{k'\}$. We run the regression using $R_{jk'}$ on the training gene set to obtain the coefficients, and use the model selection gene set to compute the sum-of-squared-errors (SSE) of the coefficients. We find the TF k'' with the smallest SSE on the model selection genes, and add k'' to R_j. We limit the size of R_j to 2 to avoid over-fitting. At the end of the procedure, we compare the square-root of the average SSE (SSE divided by r, the number of test genes) of the newly obtained model with the non-interacting model (assuming R_j is empty) on the set of test genes to see whether the interacting model has smaller error. To assess whether the improvement is due to chance alone, we also incorporated random-selection as the null distribution: we follow all steps as above except when augmenting R_j: instead of selecting the TF with the smallest error to add, we randomly select the TF.

Identification of interactions using entire data. The regression model assuming that TF k affects the binding behavior of TF j is $Y_{ij} = \mu_j + a_j x_{ij} + b_{jk} x_{ik} + \varepsilon_{ij}$ (the model is the same as that in previous section when $R_j = \{k\}$ for some $k \neq j$). The p-value for the assumed interaction is the significance of b_{jk}, a positive b_{jk} represents *synergistic* interaction, and a negative b_{jk} *antagonistic*. To account for multiple-testing, we estimated the false discovery rate (FDR)[1] corresponding to the p-values. Unless otherwise mentioned, we use a FDR threshold of 0.1. By applying these thresholds, we obtain 377 synergistic interactions and 68 antagonistic interactions.

4   Discussion

The importance of combinatorial interaction among transcription factors for proper transcriptional regulation is widely acknowledged. However, the DNA-binding of a factor itself is often treated as an isolated independent event, despite evidence to the contrary. We have shown that by considering dependence between TFs we can better explain the experimental TF-DNA binding data. Unlike the interactions discovered by various other methods [9, 14, 31] the proposed method predicts antagonistic interactions as well as the synergistic ones. Our method accurately identifies most of the TF-TF interactions of the yeast cell cycle. Although a similar validation for antagonistic interactions is currently not possible, but we have shown that the presence of antagonizing factor's motif reduces the binding probability for the antagonized factor. This inverse correlation is modest but significant.

   Our attempts to include more that two additional TFs in the linear regression, does not generally yield an improved model. However, a natural extension of our method is to investigate the optimal number of interacting TFs. This can be done by incrementally adding additional TF in the regression that yields the maximum reduction in SSE. Based on our investigation, we do not expect a majority of factors

---

[1]   FDR control procedure takes as input the p-values of all genes and FDR threshold (eg. 0.1), and finds a maximum p-value threshold for significance, such that the proportion of false positives out of all called positives (significant genes) never exceeds the given FDR threshold. We use the Bioconductor library for R to perform the FDR control, which implements the method described in [30].

to depend on multiple other factors for their binding, and for the subset which might depend on several other factors, perhaps a non-linear model would lead to better predictions. A lack of sufficient data (number of genes bound to a TF) as well as the inherent experimental noise make this a difficult proposition.

Similar to other computational analysis of gene regulation, there are several limitations to our approach. We implicitly assume that all transcription factors are available in the cell and the only determinant of binding is the interaction. We have essentially ignored the dynamic nature of binding owing to presence/absence of factors, co-factors, chromatin structure etc. The available ChIP data presents only a single snap-shot of the cell, and ChIP data under different conditions may reveal slightly different relationships among factors. This general limitation of the available data affects all computational approaches for predicting interactions. Finally, we have assumed a known motif for each factor which was computed based on the bound set of promoters, but an approach that attempts to simultaneously detect the motifs as well as the interactions should be more effective.

References

1. Ptashne, M. A genetic switch, Edn. third. (Cold Spring Harbor Laboratory Press, 2004).
2. Kadonaga, J.T. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 116, 247-257 (2004).
3. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249, 505-510 (1990).
4. Guille, M.J. & Kneale, G.G. Methods for the analysis of DNA-protein interactions. *Mol Biotechnol* 8, 35-52 (1997).
5. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* 16, 16-23 (2000).
6. Horak, C.E. & Snyder, M. ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol* 350, 469-483 (2002).
7. Bolouri, H. & Davidson, E.H. Modeling DNA sequence-based cis-regulatory gene networks. *Dev Biol* 246, 2-13 (2002).
8. Thompson, W., Palumbo, M.J., Wasserman, W.W., Liu, J.S. & Lawrence, C.E. Decoding human regulatory circuits. *Genome Res* 14, 1967-1974 (2004).
9. Lee, T.I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298, 799-804 (2002).
10. Lomvardas, S. & Thanos, D. Nucleosome sliding via TBP DNA binding in vivo. *Cell* 106, 685-696 (2001).
11. Hochschild, A. & Ptashne, M. Cooperative binding of lambda repressors to sites separated by integral turns of the DNA helix. *Cell* 44, 681-687 (1986).
12. Euskirchen, G. et al. CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol* 24, 3804-3814 (2004).
13. Bussemaker, H.J., Li, H. & Siggia, E.D. Regulatory element detection using correlation with expression. *Nat Genet* 27, 167-171. (2001).
14. Banerjee, N. & Zhang, M.Q. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res* 31, 7024-7031 (2003).
15. Harbison, C.T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99-104 (2004).

16. Pilpel, Y., Sudarsanam, P. & Church, G.M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* 29, 153-159 (2001).

17. Impey, S. et al. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* 119, 1041-1054 (2004).

18. Conkright, M.D. et al. Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. *Mol Cell* 11, 1101-1108 (2003).

19. Valineva, T., Yang, J., Palovuori, R. & Silvennoinen, O. The transcriptional co-activator protein p100 recruits histone acetyltransferase activity to STAT6 and mediates interaction between the CREB-binding protein and STAT6. *J Biol Chem* 280, 14989-14996 (2005).

20. Imai, E., Miner, J.N., Mitchell, J.A., Yamamoto, K.R. & Granner, D.K. Glucocorticoid receptor-cAMP response element-binding protein interaction and the response of the phosphoenolpyruvate carboxykinase gene to glucocorticoids. *J Biol Chem* 268, 5353-5356 (1993).

21. Blobel, G.A., Nakajima, T., Eckner, R., Montminy, M. & Orkin, S.H. CREB-binding protein cooperates with transcription factor GATA-1 and is required for erythroid differentiation. *Proc Natl Acad Sci U S A* 95, 2061-2066 (1998).

22. Grossman, S.R. p300/CBP/p53 interaction and regulation of the p53 response. *Eur J Biochem* 268, 2773-2778 (2001).

23. Raychowdhury, R. et al. Interaction of early growth response protein 1 (Egr-1), specificity protein 1 (Sp1), and cyclic adenosine 3'5'-monophosphate response element binding protein (CREB) at a proximal response element is critical for gastrin-dependent activation of the chromogranin A promoter. *Mol Endocrinol* 16, 2802-2818 (2002).

24. Kobayashi, A., Numayama-Tsuruta, K., Sogawa, K. & Fujii-Kuriyama, Y. CBP/p300 functions as a possible transcriptional coactivator of Ah receptor nuclear translocator (Arnt). *J Biochem (Tokyo)* 122, 703-710 (1997).

25. Oelgeschlager, M., Janknecht, R., Krieg, J., Schreek, S. & Luscher, B. Interaction of the co-activator CBP with Myb proteins: effects on Myb-specific transactivation and on the cooperativity with NF-M. *Embo J* 15, 2771-2780 (1996).

26. Pouponnot, C., Jayaraman, L. & Massague, J. Physical and functional interaction of SMADs and p300/CBP. *J Biol Chem* 273, 22865-22868 (1998).

27. Love, T.M. et al. Activation of CREB/ATF sites by polyomavirus large T antigen. *J Virol* 79, 4180-4190 (2005).

28. Huang, H. et al. MEIS C termini harbor transcriptional activation domains that respond to cell signaling. *J Biol Chem* 280, 10119-10127 (2005).

29. Smith, A.D., Sumazin, P., Das, D. & Zhang, M.Q. Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* 21 Suppl 1, i403-i412 (2005).

30. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.* B 85, 289-300 (1995).

31. Hannenhalli, S. & Levy, S. Predicting transcription factor synergism. *Nucleic Acids Res* 30, 4278-4284 (2002).