Genome Biology

Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors

Genome Biology 2011, **12**:R111 doi:10.1186/gb-2011-12-11-r111

Chao Cheng (chao.cheng@yale.edu)
Chong Shou (chong.shou@yale.edu)
Kevin Y Yip (kevinyip@cse.cuhk.edu.hk)
Mark B Gerstein (ccmg@gersteinlab.org)

ISSN 1465-6906

Article type Method

Submission date 24 May 2011

Acceptance date 7 November 2011

Publication date 7 November 2011

Article URL http://genomebiology.com/2011/12/11/R111

Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors

Chao Cheng[1*], Chong Shou[2*], Kevin Y Yip[1,3] and Mark B Gerstein[1,2,4$]

1 Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, New Haven, Connecticut, 06520, USA.
2 Program in Computational Biology and Bioinformatics, Yale University, 260 Whitney Avenue, New Haven, Connecticut, 06520, USA.
3 Department of Computer Science and Engineering, The Chinese University of Hong Kong, Rm. 1006, Ho Sin-Hang Engineering Bldg., Shatin, New Territories, Hong Kong.
4 Department of Computer Science, Yale University, 260 Whitney Avenue, New Haven, Connecticut, 06520, USA.

* These authors contributed equally to this work
$ Correspondence: mark.gerstein@yale.edu

Abstract

We propose a method to predict yeast transcription factor targets by integrating histone modification profiles with transcription factor binding motif information. It shows improved predictive power compared to a binding motif-only method. We find that transcription factors cluster into histone-sensitive and -insensitive classes. The target genes of histone-sensitive transcription factors have stronger histone modification signals than those of histone-insensitive ones. The two classes also differ in tendency to interact with histone modifiers, degree of connectivity in protein-protein interaction networks, position in the transcriptional regulation hierarchy, and in a number of additional features, indicating possible differences in their transcriptional regulation mechanisms.

Introduction

Transcription factors regulate target gene expression through binding to specific genomic regions. In *Saccharomyces cerevisiae*, transcription factor binding sites (TFBSs) are often adjacent to and upstream of target loci due to the compact nature of the yeast genome [1, 2]. Upon binding, TFs interact with RNA polymerase II to activate or repress transcription. TFs also recruit chromatin modification enzymes to induce chromatin structure changes, which in turn affect the accessibility of factors to genomic DNA regions [3, 4]. The target genes of a TF change according to developmental, physiological and extra-cellular environmental conditions [5]. In addition, TFs interact with each other through combinatorial binding [6]. Uncovering TF target genes and inter-relationships between TFs for all different conditions is thus important for understanding gene expression regulation, but it is also a difficult task due to the scale of the problem.

Several different experimental methods have been developed to identify TFBSs. Chromatin immuno-precipitation followed by tiling array (ChIP-chip) has been widely used to identify TFBSs in the genomic scale [7-9]. More recently, high-throughput sequencing after chromatin immuno-precipitation (ChIP-Seq) has been shown promising in getting higher resolution of the TFBSs [10, 11]. With these methods, an increasing amount of TFBS data have been accumulated for different TFs in different species, cell types, conditions, and so on, which have started to unravel the global picture of gene expression regulation. In yeast, the TFBSs and target genes for an almost complete set of TFs have been mapped in common YPD medium using ChIP-chip [5]. However, resources are still too limited to support a complete exploration of TF binding for all the combinations of cell types and conditions.

Many computational methods have also been proposed to predict TFBSs [12-19]. These methods are mostly based on the idea that the binding of a TF is mediated by the recognition of its binding motif represented as a position specific scoring matrix (PSSM). PSSMs are usually discovered as those enriched motifs from TFBSs in ChIP-chip or ChIP-seq experiments, or *de novo* from non-coding genomic sequences [5, 20]. Scanning and matching PSSMs in the genome constitute the core of these methods, which are then improved by incorporating information of motif conservation and TFBSs co-localization. Nevertheless, these methods often lead to considerably high rate of false positives. Furthermore, most of these methods are not condition specific and thus could not reflect the dynamic nature of TF binding under different conditions.

Chromatin modifications could modulate the accessibility of DNA regions and affect the recruitment of TFs [3, 4, 21]. Both functions directly relate to transcription regulation by TFs. Genomic mapping of chromatin modifications in yeast using ChIP-chip has provided the opportunity to investigate their underlying relationships with TFBSs [22, 23]. Many chromatin modifications have been shown to be associated with transcription activation and repression [3, 4]. Recent studies have shown that incorporating histone modification data improves prediction of TFBS in mouse and human [24, 25]. In these models, chromatin modifications generally provide non-TF-specific chromatin accessibility, while PSSMs determine TF-specific bindings.

Here we propose a method that integrates PSSMs and chromatin modifications to improve TF target gene predictions in yeast. Specifically, we trained individual support vector machine (SVM) models [26] for 203 yeast TFs using 2 types of features: the existence of PSSM upstream of genes and chromatin modifications adjacent to the ATG start codons. The models were trained and tested using TF target genes from ChIP-chip experiments. Interestingly, we found that some yeast TFs were more sensitive to histone modifications than others: their target genes had relatively higher histone modification signals, and were better predicted when these signals were included in the models. We termed these two TF classes as histone-sensitive and histone-insensitive TFs for simplicity. We show that histone-sensitive and histone-insensitive TFs differ in many biological characteristics found in large-scale and small-scale experiments. Furthermore, we used the model to investigate condition specificity and TF-specificity of chromatin modifications as well as TF-TF co-operation. Our analysis helps understand the mechanism of gene expression regulation by TFs and chromatin modifications.

Results

Differential histone modifications between ChIP-chip verified and non-verified TFBSs

In order to examine whether chromatin modifications are predictive features for functional TF binding sites, we first investigated chromatin modification signals at ChIP-chip verified and non-verified TFBSs defined as follows. Based on previous TFBS prediction models, we denoted the TFBSs of a factor as the local genomic sequences that match its PSSM. We then used ChIP-chip experimental data to distinguish ChIP-chip verified and non-verified TFBSs based on the existence of actual binding peak signals. Although both verified and non-verified TFBSs contain TF binding PSSMs, they were found to have differential chromatin modification signals. Here, we use the factor SWI4, a component of the SBF complex regulating cell cycle gene expressions, as an example (Figure 1). We observed that individual histone modifications varied significantly between ChIP-chip verified and non-verified TFBSs. Among the 14 different histone modifications under 2 conditions (YPD and H_2O_2), 11 were significantly different ($p\text{-value}<0.01$) in their signals between the verified and non-verified TFBSs of SWI4 (Figure 1). Among them, H3 and H4 signatures were particularly strong features for distinguishing between the two types of TFBSs, as they showed significantly lower signal in verified TFBSs than in non-verified TFBSs ($p\text{-value}<10^{-20}$). Consistent with previous studies, this indicates that verified binding sites of factors in regulatory regions are typically depleted of nucleosomes [27-30]. Encouraged by the observed differential histone modifications between verified and non-verified TFBSs, we then constructed a model that combines histone features with binding motif information for target gene prediction in yeast.

Improving target gene prediction by combining histone modifications and PSSMs

Since the *S. cerevisiae* genome is quite compact with respect to other higher eukaryotic species, it is reasonable to define the target genes of a TF as those with one or more upstream TFBSs. We combined chromatin modifications and PSSM data and used them as input features to a SVM learning model for predicting TF target genes. The model was compared with two other models that are based only on histone modifications or PSSM. The prediction accuracy of the models was tested using a gold standard dataset from ChIP-chip experiments, which provided target genes of 203 yeast TFs [5]. Specifically, we choose 0.01 as the P-value cutoff for target gene calling from ChIP-chip, which provides us with enough number of high confidence positive target genes for model training (Figure 3A). The data set was separated into training and testing data, and the performance of the models was assessed by cross-validation (see Materials and Methods).

For chromatin modifications, we used 11 histone modifications that covered most yeast ORF regions from ChIP-chip experiments [23]. Since TFs bind to the upstream sequence of ORFs, we focused on histone modification signals 1kb flanking translation start sites (ATGs), because TFBSs were enriched in these regions. For TF PSSMs, 2 independent sets were obtained from previous studies. One set of PSSMs were discovered using sequence analysis based method, basically looking for enriched motifs in the DNA regions upstream of all yeast ORFs [20]. From ~5,600 upstream sequences a total of 666 motifs have been discovered, among which 48 could be associated with known transcription factors according to literature. The other set of PSSMs were based on ChIP-chip data [5]. For each TF a target gene set was determined and then the binding motif for these TF was identified from the DNA region upstream of these genes.

Our results indicate that for almost all TFs, the model using both histone modification and PSSM data achieves the best or nearly the best performance (measured in AUC, area under the receiver operator characteristic curve) than the other 2 models using histone modification data or PSSM data alone (Figure 2, Additional file 1-Table S1, Additional file 2-Table S2). For example, we obtained an AUC of 0.89 for predicting target genes of the factor SUM1 when both the histone modification and PSSM information were used. However, if only PSSM or histone modification information was used, the models resulted in lower AUCs (0.78 for PSSM-only and 0.86 for histone modification-only) (Figure 2A, Additional file 1-Table S1). The improved performance of the combined model was observed for both of the TF PSSM sets (Figure 2B, 2C, Additional file 1-Table S1, Additional file 2- Table S2), indicating that the improvement does not rely on particular source or quality of PSSMs. Interestingly, for some TFs, histone-only model performed better than PSSM-only model; while for some other TFs, the opposite was observed.

In order to examine whether we could achieve better TF target prediction when we have more chromatin modification data, we used histone modification datasets from two independent experiments, performed by Pokholok *et al* [23] and Kurdistani *et al*, respectively [22]. We found that we achieved higher prediction accuracy by using dataset from Pokholok *et al* than from Kurdistani *et al*. This is probably due to the fact that the latter contains only histone acetylation data, while the former contains both histone methylation and acetylation data, which might provide complementary information for regulating chromatin structure and recruiting TF binding. We then combined the two datasets for predicting TF target genes, and found that for most TFs the performance was only slightly better than using the Pokholok dataset alone (Figure 3B).

Nevertheless, for some TFs we observed substantial improvement by including the Kurdistani dataset (Table 1). It is thus promising that we could improve our chromatin model performance by incorporating more histone modification data in the future.

We have also investigated the positional effect of histone modification signals for target gene prediction. First, we observed that signals of different types of histone modifications showed different patterns at DNA regions around the ATG, suggesting that they might affect TF binding in different manners. Second, histone modification signals from the upstream of ATG are generally more predictive than those from the downstream, as we have observed for both 500bp and 1000bp flanking region sizes (Figure 3C, 3D). This is somewhat expected, because TF binding sites are more enriched in the upstream regions of ORFs for transcriptional regulation. It is also interesting to see that ATG flanking regions of 500bp and 1000bp result in almost the same performance (Figure 3C, 3D). Given the compact nature of the yeast genome, transcription start sites for most ORFs are located within 1000bp region upstream of ATG [31]. The comparable performance by using 500bp flanking ATG region indicates that most discriminative histone modification signals for TF binding were embedded in this region.

The number of PSSM-containing genes is another factor that might also affect the prediction by histone modifications. We counted the number of genes containing PSSM in their promoters for each TF, and found no obvious relationship between it and the prediction power (AUC) by histone modifications (Additional file 3-Table S3, Additional file 4-S4). In fact, the number of PSSM-containing genes varies extensively from 79 to 5810 depending on the information content of the PSSMs, but its correlation with the prediction power is neglectable (with R= 0.1 for Frankel *et al* dataset).

Both histone modifications and TF binding are involved in gene expression regulations. To rule out the possibility that the capability of histone modifications for predicting TF target genes is actually mediated by expression level of genes, we examined the marginal effect of gene expression levels on TF target prediction. Including expression levels as an additional predictor in the histone modification based models does not significantly change the prediction accuracy for any TF. Furthermore, SVM models based on expression level alone can hardly predict targets for any of the TFs (Additional file 5-Table S5). The only exception is FHL1 (prediction accuracy AUC=0.79), which predominantly regulates the transcription of ribosomal protein genes. It is the extreme abundance of these target gene proteins that enables accurate prediction for the TF. Thus, the prediction power of histone modification models is not likely to be mediated by gene expression levels overall.

Condition specificity of the chromatin model

TFs bind to and regulate target gene expression in a complex and dynamic manner to coordinate biological processes [5, 32]. Chromatin modifications also change rapidly in response to stimulus from extra-cellular environment [3]. Therefore, chromatin modifications in one condition should match TF target binding in that condition but not other conditions.

We investigated condition specificity of our chromatin model in two conditions, YPD and H_2O_2. We tested a total of 12 TFs, for which we had the PSSM, histone modification and TF target binding data under both YPD and H_2O_2 conditions. For each of the TFs, we constructed two separate chromatin models: one model (Model A) used PSSM and histone modifications under YPD condition as features, while the other (Model B) used PSSM from YPD condition but histone modifications under H_2O_2 condition. The two models were then used to predict TF target binding under H_2O_2 conditions. It is generally believed that TFs keep their binding specificity PSSMs in different conditions and even over large evolutionary distances. Therefore, we use PSSMs from YPD condition as a close approximation in Model B, where no PSSM information is available under H_2O_2 condition.

For TFs that are known to be functional under H_2O_2 condition, Model B achieved better performance than Model A. For example, HSF1, a heat shock TF that activates genes in response to stresses, is more active under H_2O_2 condition with 326 target genes, than in YPD medium with only 123 target genes. Using condition-matched histone modification data (Model B), our chromatin model achieved an AUC of 0.77. In contrast, using non-condition-matched data (Model A), the chromatin model only achieved 0.56 AUC (Figure 4A). Similar results were observed for another TF, MSN2, which is activated along with MSN4 to regulate stress response genes. MSN2 is more active under H_2O_2 condition, and our chromatin model performed better with condition-matched data (Figure 4B). These results indicate that histone modifications are actually dynamic and function in a condition specific manner. Thus, target genes of TFs under a certain condition can be best predicted by using histone modification data from the same condition. In practice, this enables us to predict conditional specific target genes of TFs, which cannot be achieved by using PSSM based method.

Relative importance of different histone modifications for target prediction

TFs bind to the upstream of target genes through recognizing their specific binding motif PSSMs. We then asked an analogous question: Do TFs have specific histone modification profiles at binding sites of their targets? To address this question, we calculated a histone modification profile for each TF by averaging upstream-ATG histone modification signals over all its target genes. In our analysis, we included 25 different histone modifications from the two studies mentioned above [22, 23].

We found that different TFs have distinct target histone modification profiles. A histone modification high in one TF's profile could be low in another TF's profile (Figure 5A). We performed unsupervised clustering analysis for the histone modification profiles of all TFs, and detected two TF clusters (Figure 5A). One of the two clusters showed generally larger variations (more high and low signals) among histone modifications in the upstream of their target genes, while the signals in the other cluster are more often around the mean, ($P<10^{-16}$, t-test). We thus refer the 68 TFs in the former cluster as "histone modification sensitive" (histone-sensitive) TFs, and the 135 TFs in the latter cluster as "histone modification insensitive" (histone-insensitive) TFs (Additional file 6-Table S6).

The correlations between pairs of histone modifications are shown in Figure 5C, based on their signals in histone modification profile over all TFs. Only pairs with strong correlation ($r > 0.5$ or < -0.5) were connected in the form of correlation network. The dense connectivity in this network reveals strong pairwise redundancy of histone modification signals, which is also indicative of redundancy for predicting target genes.

We next examined the relative importance of each histone modification for predicting target genes of all TFs. Given a histone modification, we compared its signal difference between target and non-target genes of a TF. The signal difference was represented as the t-statistics (see "Method" section for detail), which indicated the relative importance of that histone modification for predicting target genes of a TF. A larger absolute value of t-statistic indicated more importance. The t-statistics for all histone modifications forms a TF specific profile, denoted as differential modification profiles of the TF. Interestingly, histone sensitive TFs and histone insensitive TFs defined based on target histone modification profiles were also distinct in their differential modification profiles (Figure 5B). This suggests that histone sensitive and insensitive

TFs are actually robust clusters with different patterns of histone modifications in their target genes.

Histone modification sensitivity of transcription factors

To understand the biological nature of the histone sensitive and insensitive TFs, we explored the feature differences of these two TF classes under different biological “context”. First, we observed different predictive power of histone modifications for target gene prediction between the two TF classes. As shown in Figure 6A, histone modifications are generally more predictive of the target genes for histone sensitive TFs than for histone insensitive TFs. This is due to the fact that target genes of histone sensitive TFs have stronger histone modification signals, which substantially improve the performance of our chromatin model.

Histone sensitive and insensitive TFs also show distinct topological characteristics in biological networks. In general, histone sensitive TFs have less target genes than histone insensitive TFs (Figure 6B). Yu et al has constructed a hierarchical network in yeast based on the TF-TF regulation relationships identified by the ChIP-chip [33]. We mapped the histone sensitive and insensitive TFs onto the hierarchical network, and found that histone sensitive TFs were enriched in the upper layers. This suggests that histone sensitive TFs are more likely to act as “managers” that regulate other TFs, while histone insensitive TFs tend to be “workers” at bottom layer in the hierarchy (Table 2). We have also examined the “degrees” of these TFs in the protein-protein interaction networks [34]. Our results indicate that histone sensitive TFs tend to have more physical interacting partners than those histone insensitive TFs (Figure 6C). The high connectivity of histone sensitive TFs further implies their functional importance.

Interestingly, histone sensitivity of TFs also indicates distinct co-regulation relationships. Two TFs are said to co-regulate if their sets of targets significantly overlap. Among the ~14,000 possible TF pairs, we found 1,440 significant co-regulatory relationships ($P<0.05$, Fisher’s exact test). Among the TFs involved in co-regulatory relationships, 64 are histone sensitive and 95 are histone insensitive. In the 1,440 significant co-regulation pairs, 447 are between two histone sensitive TFs, 437 between two histone insensitive TFs, and 556 between one histone sensitive TF and one histone insensitive TF. Fisher’s Exact test showed that histone sensitive TFs are more likely to be involved in a co-regulatory relationship than histone insensitive TFs ($P<10^{-16}$). In summary, the histone sensitive TFs reside mostly in the upper layers of the regulatory network, tend to work and communicate with other TFs during transcriptional regulations.

Furthermore, we found that the expression levels of histone sensitive TFs were higher than those of the histone insensitive TFs (Figure 6D). It seems that histone sensitivity of TFs was also related to their biological functions. For example, TFs involved in cell cycle regulation predominantly belong to histone sensitive class (Table 3). Out of 17 cell cycle TFs reported in a previously study [35], only CST6 was classified to be histone insensitive. We also examined TFs that were specific to certain conditions, e.g. heat shock or oxidative stress. Conditional specific TFs were classified into both histone sensitive and insensitive classes, thus it was not obvious whether the histone sensitivity and condition specificity of TFs are related.

PSSM predictability and cooperativity of transcription factors

TFs exhibit different PSSM predictability in that the targets of some TFs are well predicted using its PSSM alone but others are not. PSSM predictability reflects the extent to which TFs recognize its binding site through motif matching. Since the PSSMs are available for only 50 TFs, accounting for about a quarter of 203 TFs with histone modification profiles, it is difficult to perform systematic identification and classification to categorize them into PSSM well-predictable and weakly-predictable. However, we have identified 10 TFs with target prediction AUC>0.7 using their PSSMs alone, providing a confident subset of PSSM well-predictable TFs (Table 4). To check whether the high predictability is attributed to PSSM specificity, we calculated the information content of these PSSMs. We found that the information content of the 10 well-predictable TFs' PSSMs is not different from other TFs (P=0.4, Wilcoxon test). We investigated the expression level, number of target genes, hierarchy in regulatory network of these 10 TFs, and found no significant difference from the other TFs. We will be able to make more confident conclusions when more PSSMs for TFs become available in the future.

For the TFs that are weakly predictable using PSSM information, we hypothesized that these TFs may bind to their targets indirectly by cooperating with other TFs. If the hypothesis were true, we would expect to predict its target genes accurately by using the PSSM of its cooperative TF. We tested this by using each PSSM to predict targets of all TFs (Additional file 1-Table S1, Additional file 2-Table S2). The targets of most TFs were best predicted by their own PSSMs, but some TFs have their targets better predicted using other TFs' PSSMs. For example, YAP1's target genes were better predicted using CAD1's PSSM, with AUC increased from 0.71 to 0.75. Similarly, using INO2's PSSM, INO4 was better predicted of its target genes with AUC increased from 0.78 to 0.81. In fact, YAP1 and CAD1 work together in stress induced transcriptional responses, and INO2 and INO4 form heteromeric complexes involved in phospholipids biosynthesis [36, 37]. We also found that the PSSMs of the cooperative TFs are actually quite similar, measured by a similarity score range from 0 to 1. The PSSMs of CAD1 and YAP1 render a similarity score of 0.72 (top 1% among all pairs), and those of INO2 and INO4 render 0.55 (top 5%). This further indicates the cooperation between the two TF pairs through indirect binding. Therefore, TF target gene prediction using cross PSSMs could help identify cooperations between TFs. On the other hand, this suggests that using TF's own PSSM may not always be the best for predicting its target genes, when there is evidence of TF cooperations.

Chromatin model improves prediction of TF binding sites

We have examined our chromatin model using the TF target data from large-scale ChIP-Chip experiments [5], and shown its effectiveness for predicting target genes. The data from Ref. [5] investigated TF binding within yeast promoter regions only. However, the technical advancement of ChIP-Chip and ChIP-seq has enabled us to obtain the binding sites of a TF across the whole genome. Given these more high-resolution data, our chromatin model can also be used to predict TF binding sites. For demonstration, we would use the ChIP-seq data for STE12 as the example [38].

We examined chromatin signals around the STE12 binding peaks. We found many histone marks such as H3K9ac were enriched in the peak regions (Additional file 7-Figure S1), implying that they are informative for predicting TF binding sites. We separated yeast genome into 100-nucleotide bins, and divided them into positive bins and negative bins according their overlap with STE12 binding peaks. Then we constructed SVM classification models (histone modification model, PSSM model and combined model) to predict positive bins in a similar way as predicting target genes. We estimated their prediction accuracy by using cross-validation method, which verified the effectiveness of the chromatin model for binding site prediction (Figure 7). The prediction power of using both histone modification and PSSM information is roughly the same as using only histone modification (AUC=0.73), whereas using PSSM only is close to random (AUC=0.5). The advantage of combining histone modifications with PSSM information, however, is clearly demonstrated when the PPV (Positive Predictive Value, the fraction of positive predictions being true positives) is concerned (Figure 7). This is of particular importance in TF binding site prediction because reducing false positive predictions is the major challenge of model improvement.

Comparison with previous methods

We compared our SVM based method with several previously published approaches including Cluster-Buster [13], MCAST [39], EEL [40], and Stubb [16]. We calculated the prediction accuracy of each method by applying it to 10 TFs with more than 200 target genes under YPD condition. As shown in Table 5, our method that integrates histone modification and PSSM data sets achieves the best prediction for most factors. For those histone-senstive TFs such as SWI4 and SWI6, including histone modification data can improve target prediction accuracy substantially, and PSSM alone gives relatively poor predictions no matter what algorithms are used to search for TF binding sites.

Among those previously published methods, EEL and Stubb take advantage of conservation of TF binding motifs between related species, and as shown they achieve relatively more accurate prediction results than FIMO, Cluster-Buster and MCAST. We also tried the “Chromia” method proposed by Won et al [24]. Similar to our method, Chromia integrates histone modification and PSSM data sets but using a hidden Markov model. The method has shown impressive performance when applied to genome-wide ChIP-seq data in mouse for predicting TF binding sites. However, when applied to the yeast data in our case, it does not result in good prediction due to the low coverage of the histone modification and TF binding data sets [5, 23]. For example, the arrays used for Pokholok ChIP-chip data contains ~42,000 probes (60-mers), representing only about 20% of the yeast genome [23]. The arrays used for identifying yeast TF binding sites are essentially promoter arrays, covering only DNA regions around the translation start site of yeast ORFs [5]. In practice, our method only requires data for interested regions, e.g. promoter regions, and thereby is more flexible and can be applied to a wide range of data sets.

We examine the effectiveness of these methods for predicting STE12 binding sites (see Section “Chromatin model improves prediction of TF binding sites”). Chromia predicts STE12 binding sites with AUC=0.66 and PPV=5.6%, presumably also due to the low resolution of the histone modification data. Those motif-based methods perform similarly with our PSSM-based methods (Figure 7), and taking into account conservation does not lead to significant improvement for the STE12 case.

Discussion

Histone-sensitive and insensitive TFs

For the 203 yeast TFs used in our study, we classified them into 68 histone-sensitive and 135 histone-insensitive TFs based on the upstream histone modification signals of their target genes. The two classes have generally opposite characteristics in histone modification signals, expression levels, topology in regulatory networks and other biological features.

Steinfeld et al. have discovered a list of TF-chromatin modifier interactions in yeast from genome-wide analysis of high-throughput experiments [41]. Among the 35 TFs that interact with chromatin modifiers, 20 are classified as histone-sensitive and 15 as histone-insensitive according to our analysis. Namely, there is a significant enrichment of histone-sensitive TFs (P=0.001, Fisher exact test) in the chromatin modifier interacting TFs. This suggests that histone-sensitive TFs are more likely to interact with chromatin modifiers, consistent with the observation that their target genes tend to have stronger histone modification signals.

Hitone sensitive TFs might target highly regulated genes. It is known that gene expression is regulated by specific TFs and their orchestrating chromatin modification enzymes. Thus, stronger histone modification signals upstream of the target genes of the histone sensitive TFs might imply more intensive transcriptional regulation. Our results showed that cell cycle TFs were mostly histone-sensitive TFs, consistent with the previous knowledge that cell cycle is highly-regulated to achieve cyclical expression of genes.

The majority of histone modification data used in this study is based on YPD medium. We found that histone sensitive TFs tend to be active under YPD medium, as indicated by larger number of target genes and higher expression levels with respect to those insensitive TFs. It is possible that histones upstream of the target genes of the histone sensitive TFs have more chance to be modified by histone modification enzymes, because they are recruited by these more active TFs.

We found that histone sensitive TFs were enriched in higher layers of the hierarchical regulatory network. This suggests that histone sensitive TFs tend to be “managers” that regulate other TFs and for such a reason their binding to target genes is highly regulated through histone modifications. In consistent with this hypothesis, we have observed stronger histone modification signals in the upstream regions of their target genes. We caution that here the classification of the histone-sensitive and insensitive TFs is only based on the histone modification signals of their target genes and the contribution of these signals for target predictions. Further experiments might be worthwhile to investigate the relationship between the target selection of the TF classes and histone modifications in more details.

A more recent paper in human showed the capability of using unique chromatin signatures to identify two distinct classes of genomic elements, active and poised enhancer [19]. Consistently, here we find in yeast that the histone modifications are also informative for distinguishing ChIP-chip verified and non-verified binding sites. As we demonstrated, the method could be applied to identify target genes and the genome-wide binding sites of TFs.

Contribution of histone modifications to condition specific target prediction

It is widely known that transcriptional regulation is condition specific in that TFs change their binding sites under different conditions. We showed here that histone modification data is most predictive of TF target binding under the same condition. This is true especially for those TFs that are mostly active in specific stress conditions.

Because of limited resources, it is impossible to perform exhaustive experiments for every TFs, cell types, species and all possible conditions. Meanwhile, the TF binding recognition motifs, PSSMs, are generally thought to be non-condition-specific which do not change under different conditions [3]. Thus, prediction based on motif information alone cannot provide condition-specific target genes for a TF. As an alternative method, we proposed the feasibility to predict target genes of a TF under an interested condition by combining histone modification data under that condition with its PSSM. In this way, we can achieve much higher results than using PSSM alone. More importantly, since histone modifications reflect the chromatin states in a specific condition or tissue, the predicted targets are also condition and tissue specific.

Combinatorial interaction of TFs: direct and indirect binding

When a TF binds directly to the promoter regions of its target genes, the enriched motifs identified from its binding sites can be regarded as its own PSSM. However, TFs do not always act individually; sometimes they cooperate with (physically bind to) each other to form regulatory functional units, such as yeast cell cycle complexes SBF (SWI4-SWI6) and MBF (MBP1-SWI6) [30]. In these indirect binding cases, it is important to distinguish the TFs that are motif-recognizing and the ones that are not.

By examining the PSSM sensitivity of TFs, we were able to infer some possible combinatorial interactions between TFs. If TF A's targets are better predicted by using another TF B's PSSM instead of its own PSSM, then this is an indication of potential cooperation of the two TFs. In particular, TF B directly binds to promoter regions through its PSSM, and TF A indirectly binds to promoter regions through physical binding to TF B [42]. This is also referred to as indirect piggy-back binding [6].

PSSM sensitivity under indirect TF binding is important for our target gene prediction model. Instead of using a TF's own PSSM, PSSM of another TF through which the TF binds should be used for more accurate predictions. Therefore, identifying those cases before using our model will be necessary to achieve better results.

Implications on gene expression regulation

In this study, we showed that incorporating chromatin modification information could substantially improve the prediction of TF target genes. In fact, chromatin modifications relate to gene expression regulation in two layers [3]. First, chromatin is modified to form structure as euchromatin, within which genes could be turned on and off, or heterochromatin, within which genes are silenced. Second, euchromatin is further modified by enzymes recruited by specific TFs to mark the "on and off" status of transcription. We examined the target vs. non-target differential histone modification profiles for each individual TFs, and observed TF specific chromatin modifications marked in the target genes. Therefore, we suggest that chromatin modifications might function as both non-specific euchromatin marks and TF specific regulatory marks. And our model takes advantage of the chromatin information from both of the two layers.

However, it is still not quite clear the sequential order in terms of time and causality of chromatin modification and TF binding. It is possible that one of them happens first which then drives the occurrence of the other. The other possibility is that the two events might be interactive in a feedback manner to regulate gene expression. More fine-tuned experiments in the future would be helpful for unraveling the time-dependent interaction between chromatin modification and TF binding.

Materials and methods

Chromatin modification data

The yeast histone modification data sets used in this study are basically from two sources. The first data set is available from Pokholok et al [23], which contains the profiles of 14 chromatin features under YPD or H_2O_2 condition. These chromatin features include histone H3 and H4 occupation, H3K9ac, H3K14ac, H4K5ac8ac12ac16ac, H3K4me1, H3K4me2, H3K4me3, H3K36me3, and H3K79me3. The profiles of these features are measured by ChIP-chip experiments using over 40,000 probes, which cover 85% of the yeast genome. We calculated the signal of each chromatin feature in the 1kb upstream region of each open reading frame (ORF) by averaging signals of all the probes within this region. Similarly, for each ORF the average signal of each feature in the 1kb region downstream of the start codon was also calculated. We named them as the upstream chromatin signal and downstream chromatin signal for ORFs, respectively.

The second data set is available from Kurdistani et al [22]. The data contains levels of acetylation of 11 lysines in intergenic regions (IR) as well as ORF regions. These profiles were also measured by using ChIP-chip experiments. These 11 histone acetylations are H2AK7ac, H2BK11ac, H2bK16ac, H3K9ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H4K8ac, H4K12ac and H4K16ac. We named the signal in IR and ORF as the IR chromatin signal and coding region (CR) chromatin signal for ORFs, respectively.

Target genes of yeast transcription factors

Target genes for 203 yeast transcription factors under various conditions (including YPD and H_2O_2) have been identified using the ChIP-chip experiment by Harbison et al [5]. For each binding interaction, a probability score (P-value) was calculated, measuring the binding potential of a TF with the promoter region of a gene.

When TF target genes are determined according to ChIP-chip data, one needs to set a cutoff for P-values, which indicates the confidence of regulation of genes by TFs. A small (strict) P-value cutoff would result in fewer but more confident target genes, while a large (loose) P-value cutoff would do the opposite. For instance, there are 159 target genes for RAP1 using a cutoff value of 0.001, while the target gene number increases to 581 when the cutoff value 0.05 is used. We therefore tested the influence of P-value cutoff on our model performance. As shown in Figure 3A, our results indicate that a more stringent P-value cut-off, i.e. smaller target gene set, improves the prediction accuracy of our model. Moreover, at all cutoff values the models combining histone modification and PSSM data outperform the models using either of them alone. On the other hand, a more stringent cutoff results in less target genes. To ensure enough positive target genes for model training, we decided to use 0.01 as the P-value cutoff in our analysis.

Position-specific scoring matrices of transcription factors

Two sets of position-specific scoring matrices (PSSMs) for yeast transcription factors have been identified previously using different strategies [10,18]. The first set was downloaded from [20], which was based on *de novo* motif finding in all yeast promoter sequences. The promoter DNA sequences (from start codon of a ORF to 800bp upstream) of all yeast ORFs were analyzed to identify enriched motifs by using the AlignACE program [43]. A total of 666 motifs have been found, among which 51 can be associated with known yeast transcription factors. The occurrences and matching scores of these motifs in the promoter regions of all yeast genes were also provided by Beer et al [20].

The second set of PSSMs was available from [5], which is based on motif analysis of target promoters identified by the ChIP-chip experiment. The detail about motif discovery procedure can be found in [5]. In brief, for a transcription factor the motifs were discovered by applying a suite of motif discovery programs to the intergenic sequences identified by the binding data for this factor. The resulting motifs were subsequently clustered, filtered and selected to give rise to a single PSSM that can best represent the motif of a factor. For some factors the above procedure failed to identify their motifs and in such cases motifs were derived from literature or databases.

The information content (IC) of a PSSM is calculated as $IC = -\sum_{i,j} p_{i,j} \times \log\left(p_{i,j} / p_b\right)$, where i and j represent positions in PSSMs and four nucleotides, respectively. $p_{i,j}$ is then the weight at each PSSM position of each nucleotide, and p_b is the background nucleotide frequency of *S. cerevisiae* genome. Specifically, we use 37% as the GC content to calculate p_b for each nucleotide.

The similarity between two PSSMs is calculated as the averaged dot product at each PSSM positions, $Similarity = \frac{1}{n}\sum_{i,j} p_{1,i,j} \times p_{2,i,j}$, where n is the length of PSSM. If two PSSMs are of different lengths, we compare each possible alignment of the two PSSMs with no gap, and keep the maximum similarity from each alignment. The similarity score has the range from 0 to 1.

Searching promoters for known motifs

Given the list of PSSMs for transcription factors, we searched the promoters of all yeast genes for the occurrences of these motifs using FIMO of the MEME suite [44]. The promoter region was defined the DNA region from the start codon to 800bp upstream of an ORF. The cumulative matching score (CMS) of the occurrences of a motif in the promoter region of a gene was calculated, which was subsequently used as features for predicting TF target genes.

Comparison of chromatin modifications between ChIP-chip verified TFBSs and non-verified motif matching sites

We performed comparative analysis of chromatin modification differences between TFBSs verified by ChIP-chip experiment and those non-verified motif matching sites for TFs with available PSSMs. SWI4 was shown here as an example. A list of binding sites (TFBS) of the factor SWI4 was downloaded from the Saccharomyces Genome Database (SGD) [45]. This list contains 99 binding sites that were targeted by the SWI4 under YPD medium according to the ChIP-chip data. We also collected a list of non-TFBSs by selecting DNA regions that was consistent with the SWI4 motif but not targeted by SWI4 under YPD medium as indicated by the ChIP-chip results (P>0.4). These non-TFBSs were further filtered to ensure that there is no SWI4 TFBSs within the nearby 2kb region, which ultimately resulted in 485 non-TFBSs for SWI4. All of the TFBSs and non-TFBSs are less than 20bp in size. The levels of the chromatin features on these TFBSs and non-TFBSs were calculated based on the intensities of probes covering them. Finally, the signal of the 14 chromatin features was compared between the TFBS and non-TFBS groups using the t-test.

Support vector machine model for transcription factor target prediction

For a transcription factor in each gene we have obtained the following features: cumulative matching score from motif searching, the upstream and downstream signal of 14 histone or histone modification profiles, the IR and CR signal of 11 histone acetylation profiles. Based on the ChIP-chip data from [5], genes were classified into target and non-target sets for a TF. All or subsets of these features were integrated using a support vector machine (SVM) model [26] for predicting target genes of a TF. The radial kernel was used for training and predicting in the SVM classification model.

We evaluated the performances of the models using two-fold cross validation. Specifically, we randomly split the data into two sets of equal sizes, a training set and a testing set. The model was then trained using the training set and applied to the testing set to predict target genes. The prediction power of the model was estimated based on the testing set. In general, the SVM model outputs a probability indicating how likely a gene is the target of a TF. By setting different cut-off values, we can balance the sensitivity (true positive rate) and specificity (true negative rate) of predictions of the model. The plot of the sensitivity versus 1-specificity is called receiver operating characteristic (ROC), which can be used to show the classification accuracy of the SVM model. AUC, the area under the ROC curve can be used to summarize the prediction power of the model. For each TF, we repeated this process 50 times and the average of the SUC values was calculated to represent the prediction accuracy.

Based on the above-described SVM classification method, we constructed and compared several different models for TF target prediction, each taking advantage of different features (i.e. histone modifications, PSSM information, gene expression levels) or a combination of them. First, we prepared several groups of histone modification features. Based on the Pokholok data [23], we calculated the signals of 14 histone modification types in the upstream (500bp, 1000bp) and the downstream (500bp, 1000bp) of the translational start sites of all yeast genes. From the Kurdistani data [22], we collected the signals of 11 histone modification types in the intergenic regions and in the ORF regions. Second, we calculated the cumulative matching score of the occurrences of the PSSM of each TF in the promoter region of genes, resulting in the PSSM features. Third, these features together with the expression levels of genes were selected and used as the predictors to classify targets versus non-targets of TFs. For instance, in the histone+PSSM model for factor RAP1, the histone modification features and the PSSM feature of RAP1 motif were selected as the predictors of the SVM. In "Results", we mostly used the 1000bp upstream and the 1000bp downstream histone modification features as the predictors, but the other groups of histone modification features were also investigated for comparison purposes. The SVM function in the R package "e1071" was utilized to implement the models with default parameter settings. For SVM the radial basis function (RBF) kernel was used since it achieved the best performance according to the cross-validation results.

Clustering of TFs using target chromatin modification profile

For each TF, target histone modification profile was calculated by averaging histone modification signals among all its targets. 1kb upstream chromatin modifications from Pokholok et al [20] and intergenic chromatin modifications from Kurdistani et al [22] were used. Unsupervised k-means clustering algorithm was performed to generate two TF clusters, histone sensitive and insensitive TFs, by their target histone modification profiles.

To understand the relative importance of each chromatin modification to target prediction, target-nontarget differential histone modification profiles for TFs were calculated based on t-statistic. For each chromatin modification in differential modification profile for a TF, modification signals for target genes and non-target genes were collected and t-statistic calculated. The t-statistics in differential modification profiles indicated the directional significance of chromatin modifications to distinguish target genes.

Inferring interactions between transcription factors

The target genes identified by ChIP-chip experiment could be wither direct or indirect targets of a TF. For example if two transcription factors A and B that are interacted with each other, the ChIP-chip for A can potentially identify target genes of B as well. Conversely, the existence of B's motif would be informative for predicting target genes of factor A. We used the TF target prediction model with chromatin modifications and the TF' own PSSM, and then compared the model's AUC performance to the models with chromatin modifications and other TFs' PSSMs. Models with improved AUC performances suggest better predictive power of other PSSMs than the TF's own. These cases might indicate interactions between the TFs.

Application of previously reported methods

We run all methods with their default parameter settings. Internal thresholding is turned off in all cases to report a full list of predictions with scores. PSSMs of 10 TFs and upstream 1kb DNA sequences for all annotated yeast *S. cerevisiae* ORFs are used as inputs to MCAST [22] and Cluster-Buster [22]. Pairwise pre-aligned upstream 1kb DNA sequences of all annotated *S. cerevisiae* and *S. paradoxus* orthologous ORFs are used instead for running EEL [22] and Stubb [22]. Predicted binding targets with respective scoring systems from the programs are collected for all 10 TFs. ROC curves and AUCs are calculated based on the same thresholding scheme for all methods.

Acknowledgements

We thank Dr. Won, Kyoung-Jae and Dr. Wang, Wei for providing "Chromia" source codes and suggestions about running the program. We also thank Robert Bjornson and Yale High Performance Computing for technology support. We acknowledge support from the NIH and from the AL Williams Professorship funds.

Abbreviations

TF: transcription factor; TFBS: transcription factor binding site; ChIP-chip: chromatin immunoprecipitation with microarray hybridization; ChIP-seq: chromatin immunoprecipitation with massively parallel DNA sequencing; SVM: support vector machine; PSSM: position specific scoring matrix; ORF: open reading frame; AUC: area under ROC curve; PPV: positive predictive value; IR: intergenic region.

Authors' contributions

CC conceived of the study and participated in the analysis. CS participated in the analysis. KY participated in the data analysis. MG participated in its design and coordination. All authors read and approved the final manuscript.

References

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al*: **Life with 6000 genes**. *Science* 1996, **274**:546, 563-547.
2. Goffeau A: **Four years of post-genomic life with 6,000 yeast genes**. *FEBS Lett* 2000, **480**:37-41.
3. Kouzarides T: **Chromatin modifications and their function**. *Cell* 2007, **128**:693-705.
4. Li B, Carey M, Workman JL: **The role of chromatin during transcription**. *Cell* 2007, **128**:707-719.
5. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome**. *Nature* 2004, **431**:99-104.
6. Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ: **Identifying combinatorial regulation of transcription factors and binding motifs**. *Genome Biol* 2004, **5**:R56.
7. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: **Genome-wide location and function of DNA binding proteins**. *Science* 2000, **290**:2306-2309.
8. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF**. *Nature* 2001, **409**:533-538.
9. Horak CE, Snyder M: **ChIP-chip: a genomic approach for identifying transcription factor binding sites**. *Methods Enzymol* 2002, **350**:469-483.
10. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing**. *Nat Methods* 2007, **4**:651-657.
11. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions**. *Science* 2007, **316**:1497-1502.
12. Stormo GD: **DNA binding sites: representation and discovery**. *Bioinformatics* 2000, **16**:16-23.
13. Frith MC, Li MC, Weng Z: **Cluster-Buster: Finding dense clusters of motifs in DNA sequences.** *Nucleic Acids Res* 2003, **31**:3666-3668.

14. Zhou Q, Wong WH: **CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling**. *Proc Natl Acad Sci U S A* 2004, **101**:12114-12119.

15. Gupta M, Liu JS: **De novo cis-regulatory module elicitation for eukaryotic genomes**. *Proc Natl Acad Sci U S A* 2005, **102**:7079-7084.

16. Sinha S, Liang Y, Siggia E: **Stubb: a program for discovery and analysis of cis-regulatory modules**. *Nucleic Acids Res* 2006, **34**:W555-559.

17. Van Loo P, Aerts S, Thienpont B, De Moor B, Moreau Y, Marynen P: **ModuleMiner - improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues?** *Genome Biol* 2008, **9**:R66.

18. Xie D, Cai J, Chia NY, Ng HH, Zhong S: **Cross-species de novo identification of cis-regulatory modules with GibbsModule: application to gene regulation in embryonic stem cells**. *Genome Res* 2008, **18**:1325-1335.

19. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J: **A unique chromatin signature uncovers early developmental enhancers in humans**. *Nature* 2011, **470**:279-283.

20. Beer MA, Tavazoie S: **Predicting gene expression from sequence**. *Cell* 2004, **117**:185-198.

21. Berger SL: **The complex language of chromatin regulation during transcription**. *Nature* 2007, **447**:407-412.

22. Kurdistani SK, Tavazoie S, Grunstein M: **Mapping global histone acetylation patterns to gene expression**. *Cell* 2004, **117**:721-733.

23. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK, Young RA: **Genome-wide map of nucleosome acetylation and methylation in yeast**. *Cell* 2005, **122**:517-527.

24. Won KJ, Ren B, Wang W: **Genome-wide prediction of transcription factor binding sites using an integrated model**. *Genome Biol* 2010, **11**:R7.

25. Ernst J, Plasterer HL, Simon I, Bar-Joseph Z: **Integrating multiple evidence sources to predict transcription factor binding in the human genome**. *Genome Res* 2010, **20**:526-536.

26. Cristianini N, Shawe-Taylor J: **An Introduction to Support Vector Machines and other kernel-based learning methods**; 2000.

27. Pina B, Bruggemeier U, Beato M: **Nucleosome positioning modulates accessibility of regulatory proteins to the mouse mammary tumor virus promoter**. *Cell* 1990, **60**:719-731.

28. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD: **Evidence for nucleosome depletion at active regulatory regions genome-wide**. *Nat Genet* 2004, **36**:900-905.

29. Sekinger EA, Moqtaderi Z, Struhl K: **Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast**. *Mol Cell* 2005, **18**:735-748.

30. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: **Genome-scale identification of nucleosome positions in S. cerevisiae**. *Science* 2005, **309**:626-630.

31. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing**. *Science* 2008, **320**:1344-1349.
32. Ni L, Bruce C, Hart C, Leigh-Bell J, Gelperin D, Umansky L, Gerstein MB, Snyder M: **Dynamic and complex transcription factor binding during an inducible response in yeast**. *Genes Dev* 2009, **23**:1351-1363.
33. Yu H, Gerstein M: **Genomic analysis of the hierarchical structure of regulatory networks**. *Proc Natl Acad Sci U S A* 2006, **103**:14724-14731.
34. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Res* 2006, **34**:D535-539.
35. Wu WS, Li WH: **Systematic identification of yeast cell cycle transcription factors using multiple data sources**. *BMC Bioinformatics* 2008, **9**:522.
36. Hirata D, Yano K, Miyakawa T: **Stress-induced transcriptional activation mediated by YAP1 and YAP2 genes that encode the Jun family of transcriptional activators in Saccharomyces cerevisiae**. *Mol Gen Genet* 1994, **242**:250-256.
37. Schuller HJ, Richter K, Hoffmann B, Ebbert R, Schweizer E: **DNA binding site of the yeast heteromeric Ino2p/Ino4p basic helix-loop-helix transcription factor: structural requirements as defined by saturation mutagenesis**. *FEBS Lett* 1995, **370**:149-152.
38. Lefrancois P, Euskirchen GM, Auerbach RK, Rozowsky J, Gibson T, Yellman CM, Gerstein M, Snyder M: **Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing**. *BMC Genomics* 2009, **10**:37.
39. Bailey TL, Noble WS: **Searching for statistically significant regulatory modules**. *Bioinformatics* 2003, **19 Suppl 2**:ii16-25.
40. Palin K, Taipale J, Ukkonen E: **Locating potential enhancer elements by comparative genomics using the EEL software**. *Nat Protoc* 2006, **1**:368-374.
41. Steinfeld I, Shamir R, Kupiec M: **A genome-wide analysis in Saccharomyces cerevisiae demonstrates the influence of chromatin modifiers on transcription**. *Nat Genet* 2007, **39**:303-309.
42. Gordan R, Hartemink AJ, Bulyk ML: **Distinguishing direct versus indirect transcription factor-DNA interactions**. *Genome Res* 2009, **19**:2090-2100.
43. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.
44. **The MEME Suite**. [http://meme.sdsc.edu/meme/intro.html].
45. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M *et al*: **SGD: Saccharomyces Genome Database**. *Nucleic Acids Res* 1998, **26**:73-79.

Figure legends

Figure 1. Differential histone occupation and modifications between ChIP-chip verified TFBSs and non-verified motif matching sites.

Showing SWI4 as an example, most histone modifications (in different colors) are significantly different between ChIP-chip verified TFBSs (left boxes), which have binding motifs and are bound by TFs in ChIP-chip experiments, and non-verified motif matching sites (right boxes), which have matching motifs but are not bound by TFs.

Figure 2. Chromatin modifications substantially improve TF target gene predictions.

(A) ROC curves show improved TF target gene predictions using histone modifications. (B) Performance of prediction models for individual TFs, with PSSMs from Beer et al and (C) from Harbison et al. TFs are sorted by prediction performance using histone modifications and PSSMs (red bars).

Figure 3. Model parameters

(A) Stricter thresholds of target gene call result in better predictions. (B) Combine independent histone modification datasets can improve target predictions. Predictions for 203 TFs are evaluated by AUC. (C) Using histone modifications within 500bp and (D) 1000bp window upstream and downstream of ATG sites of target genes achieve similar performance to using only upstream signals, and better performance to using only downstream signals. No significant performance difference between using 500bp and 1000bp window sizes. (E) Using histone modifications in intergenic regions has better predictive power than that in ORF coding regions.

Figure 4. Conditional specificity of chromatin model for TF target prediction

(A) ROC curves showing performance of two chromatin models to predict target genes of HSF1 and (B) MSN2/4 complex. The model (Model B) using histone modifications in H_2O_2 condition performs better to predict target genes also in H_2O_2 condition (blue curve) than to predict target genes in YPD medium (Model A) (red curve), which indicates conditional specificity of chromatin modifications and TF target genes.

Figure 5. Target histone modification profiles and target-nontarget differential modification profiles of TFs

(A) Target histone modification profiles comprising different normalized histone modification signals (columns) of TFs (rows). Target histone modification profile is the averaged histone modification signals of the TF's targets. TFs are clustered into histone sensitive (blue bar) and insensitive TFs (orange bar) using their target histone modification profiles. Histone sensitive TFs have stronger histone modification signals. (B) Target-nontarget differential modification profiles of TFs showing discriminating power (t-statistic) of histone modifications to TF targets and non-targets. TFs are ordered the same as (A). Histone sensitive and insensitive TFs have distinct differential modification profiles, indicating preferential histone modifications of TFs targets. (C) Correlation network of histone modifications in terms of TF differential modification profile. Histone modification pairs with correlation coefficient larger than 0.5 (red edges) or smaller than -0.5 (green edges) are connected. The network shows high level of redundancy of histone modifications in differential modification profiles.

Figure 6. Distinctions between histone sensitive and insensitive TFs

(A) Target genes of histone sensitive TFs are better predicted using intergenic or upstream histone modifications than histone insensitive TFs. (B) Histone sensitive TFs show smaller number of target genes in regulatory network than histone insensitive TFs. (C) Histone sensitive TFs have larger number of interacting partners in protein interaction network. (D) Higher mRNA expression levels of histone sensitive TFs.

Figure 7. Chromatin modifications improve STE12 binding site prediction.

(A) and (B) show the AUC and positive predictive value (PPV) of the histone-only model (HM), the PSSM-only model (PWM) and the combined model (HM+PWM) for STE12 binding site prediction.

Tables:

Table 1. TFs with improved prediction by including multiple histone modification datasets

TF	AUC (Pokholok et al)	AUC (Pokholok et al & Kurdistani et al)
CST6	0.70	0.80
IFH1	0.53	0.64
KSS1	0.53	0.65
RDS1	0.50	0.67
SFP1	0.59	0.74
SWI5	0.62	0.74
TYE7	0.53	0.64
YKL222C	0.57	0.69
YKR064W	0.49	0.59

Table 2. TF histone sensitivity relates to hierarchical level in regulatory network

Hierarchical levels in regulatory network		Number of histone sensitive TFs		Number of histone insensitive TFs	
1	Bottom level	15	15	52	52
2		43		36	
3	Upper levels	6	51	8	48
4		2		4	

Fisher exact test p-value = 0.0007 for separate levels, p-value = 0.0002 for combined upper levels.

Table 3. TF histone sensitivity relates to cellular functions

Cellular functions	Histone sensitive TFs	Histone insensitive TFs
Cell cycle	ACE2, ASH1, CIN5, FKH1, FKH2, MBP1, MCM1, NDD1, RLM1, STB1, STE12, STP1, SWI4, SWI5, SWI6, TEC1	CST6
Heat shock or stress conditions	GAT1, MSN4, SKN7, YAP1	GLN3, HAL9, HMS2, HSF1, MGA1, MSN2, WAR1, USV1

Table 4. Top 10 PSSM well-predictable TFs

PSSM sensitive TFs	AUC
REB1	0.87
ABF1	0.86
CBF1	0.84
FHL1	0.83
RAP1	0.79
TYE7	0.77
SUM1	0.76
UME6	0.76
MBP1	0.72
GCN4	0.71

Table 5: Comparison of several computational methods for target gene prediction.

		ROC AUC						
	Number of Target genes	**HIS+PSSM**	**HIS alone**	**PSSM alone (FIMO)**	**Cluster-Buster**	**MCAST**	**EEL**	**Stubb**
ABF1	549	0.830	0.736	0.781	0.776	0.676	0.807	0.893
FHL1	207	0.957	0.963	0.827	0.855	0.874	0.852	0.887
FKH1	284	0.656	0.625	0.606	0.680	0.546	0.661	0.725
FKH2	216	0.723	0.694	0.664	0.698	0.566	0.688	0.735
HAP1	215	0.738	0.711	0.635	0.675	0.624	0.676	0.663
RAP1	408	0.865	0.818	0.805	0.752	0.774	0.802	0.811
REB1	278	0.773	0.623	0.774	0.727	0.818	0.765	0.758
SWI4	252	0.831	0.790	0.634	0.680	0.626	0.664	0.651
SWI6	230	0.809	0.768	0.719	0.720	0.629	0.742	0.665
UME6	298	0.854	0.767	0.831	0.774	0.783	0.814	0.815

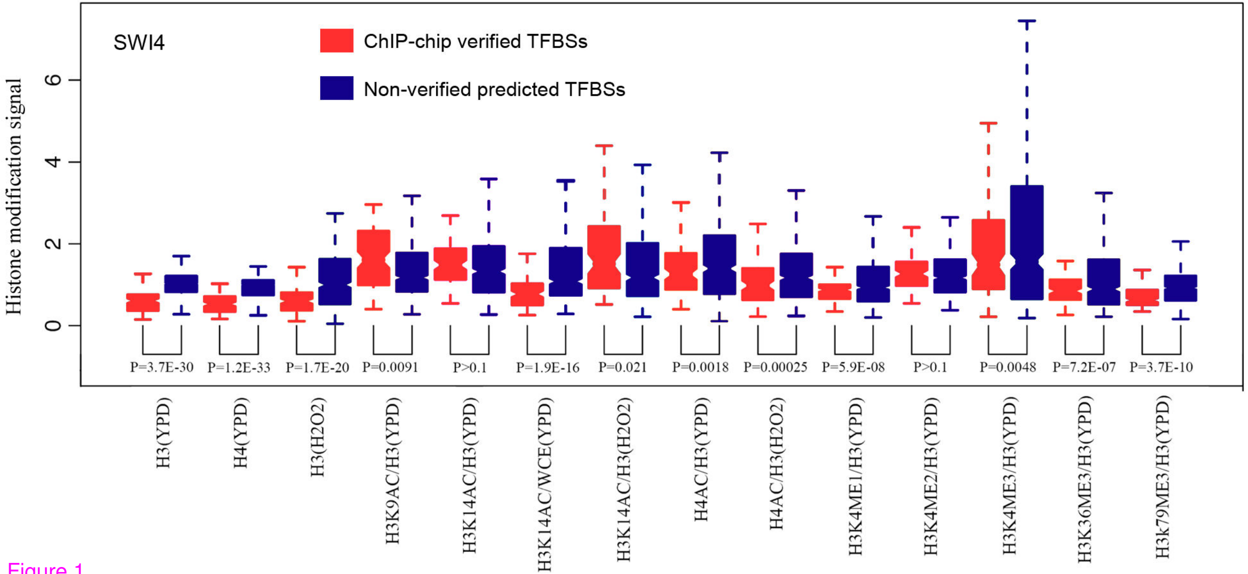
SWI4
ChIP-chip verified TFBSs
Non-verified predicted TFBSs
Histone modification signal
0
2
4
6
P=3.7E-30
P=1.2E-33
P=1.7E-20
P=0.0091
P>0.1
P=1.9E-16
P=0.021
P=0.0018
P=0.00025
P=5.9E-08
P>0.1
P=0.0048
P=7.2E-07
P=3.7E-10
H3(YPD)
H4(YPD)
H3(H2O2)
H3K9AC/H3(YPD)
H3K14AC/H3(YPD)
H3K14AC/WCE(YPD)
H3K14AC/H3(H2O2)
H4AC/H3(YPD)
H4AC/H3(H2O2)
H3K4ME1/H3(YPD)
H3K4ME2/H3(YPD)
H3K4ME3/H3(YPD)
H3K36ME3/H3(YPD)
H3k79ME3/H3(YPD)

Figure 1

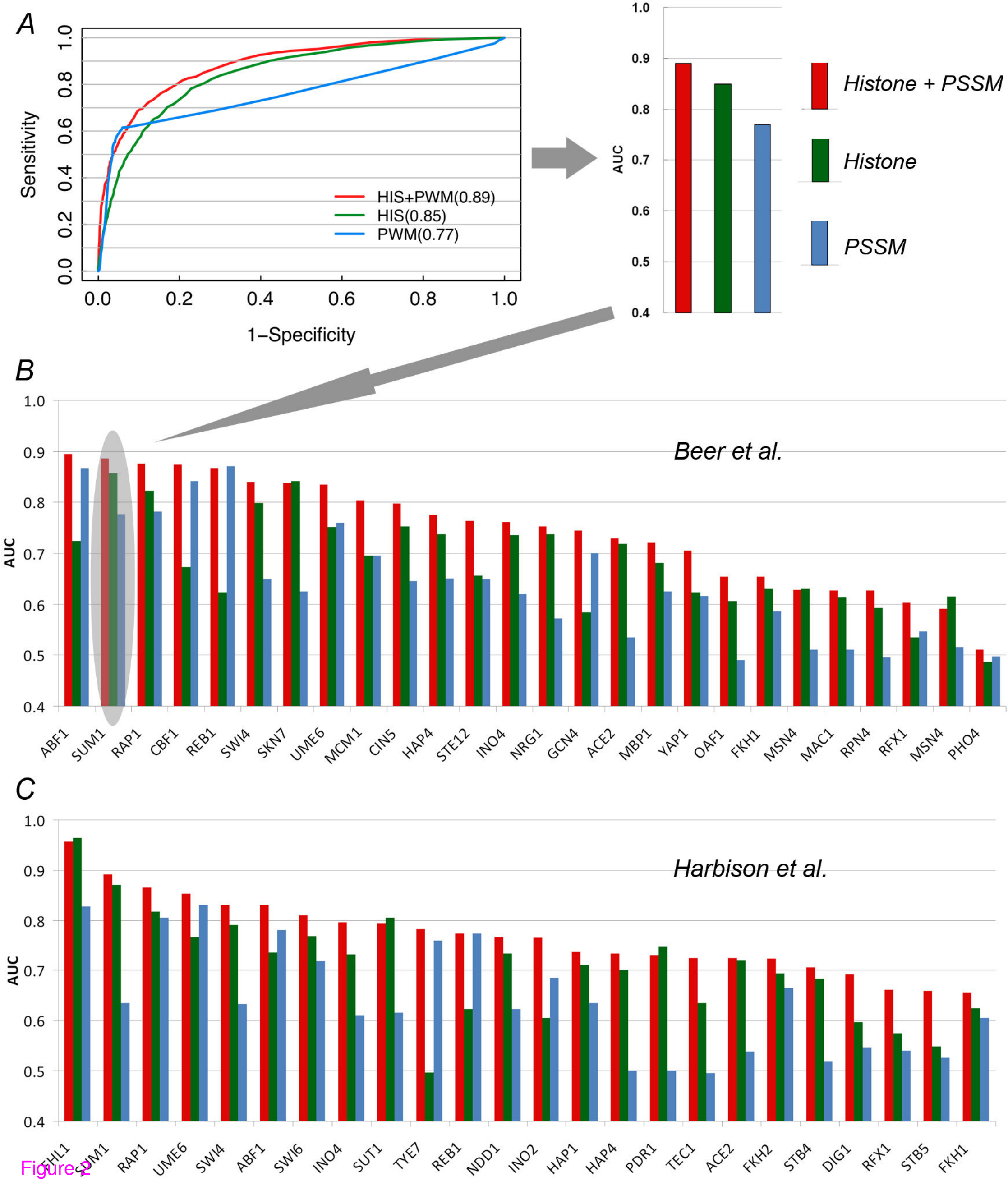
A
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0
HIS+PWM(0.89)
HIS(0.85)
PWM(0.77)
AUC
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Histone + PSSM
Histone
PSSM
B
Beer et al.
ABF1
SUM1
RAP1
CBF1
REB1
SWI4
SKN7
UME6
MCM1
CIN5
HAP4
STE12
INO4
NRG1
GCN4
ACE2
MBP1
YAP1
OAF1
FKH1
MSN4
MAC1
RPN4
RFX1
MSN4
PHO4
C
Harbison et al.
FHL1
SUM1
RAP1
UME6
SWI4
ABF1
SWI6
INO4
SUT1
TYE7
REB1
NDD1
INO2
HAP1
HAP4
PDR1
TEC1
ACE2
FKH2
STB4
DIG1
RFX1
STB5
FKH1

Figure 2

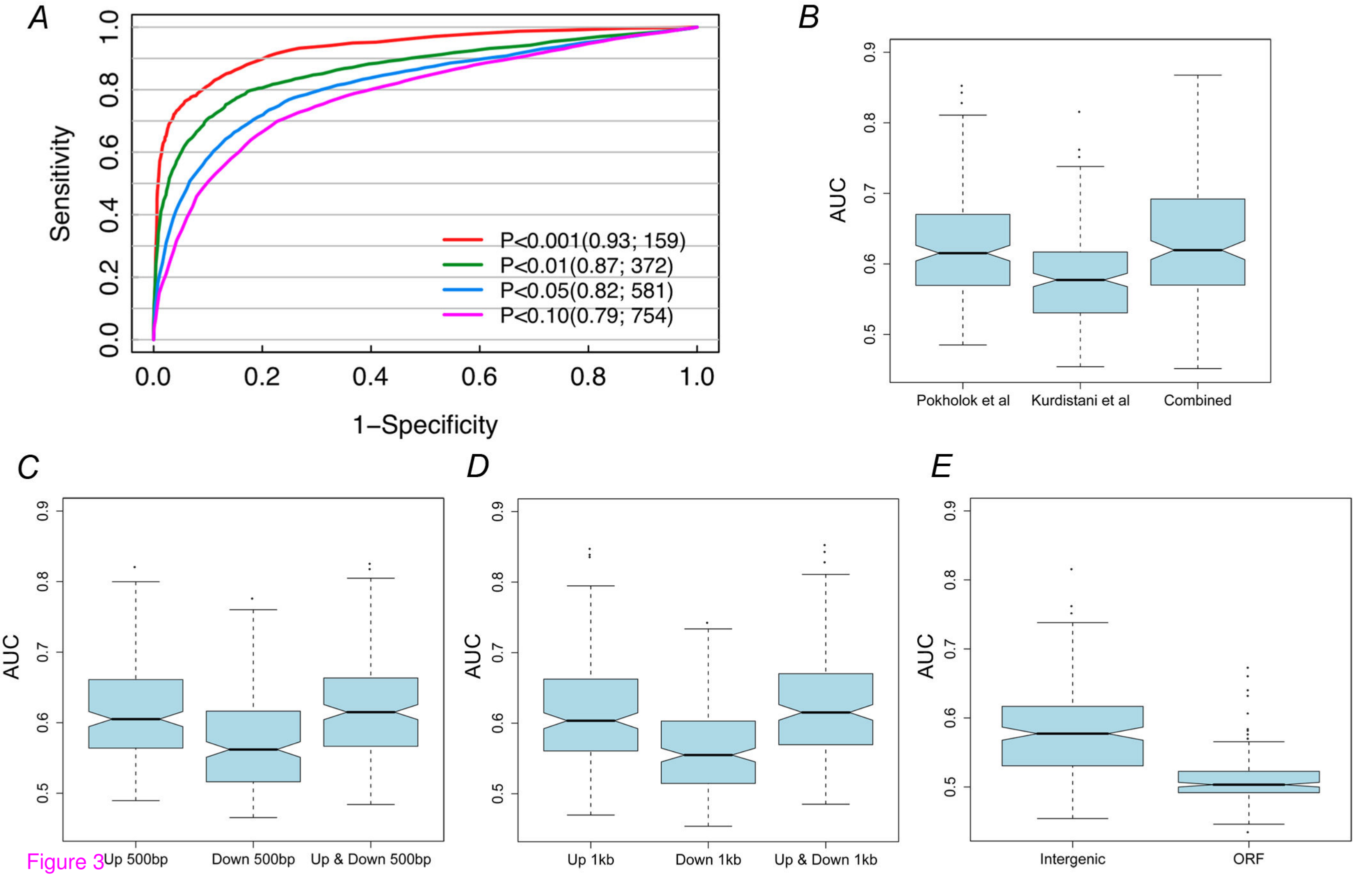
A
Sensitivity
1–Specificity
0.0
0.2
0.4
0.6
0.8
1.0
P<0.001(0.93; 159)
P<0.01(0.87; 372)
P<0.05(0.82; 581)
P<0.10(0.79; 754)
B
AUC
0.5
0.6
0.7
0.8
0.9
Pokholok et al
Kurdistani et al
Combined
C
AUC
Up 500bp
Down 500bp
Up & Down 500bp
D
AUC
Up 1kb
Down 1kb
Up & Down 1kb
E
AUC
Intergenic
ORF

Figure 3

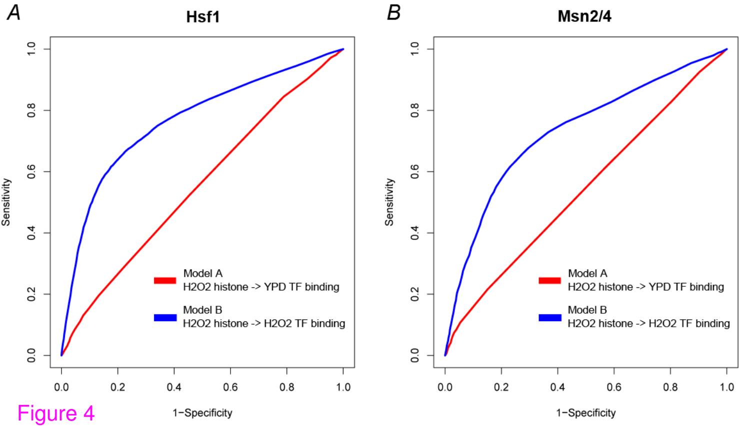
A
Hsf1
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0
🟥 Model A
H2O2 histone -> YPD TF binding
🟦 Model B
H2O2 histone -> H2O2 TF binding
B
Msn2/4
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0
🟥 Model A
H2O2 histone -> YPD TF binding
🟦 Model B
H2O2 histone -> H2O2 TF binding

Figure 4

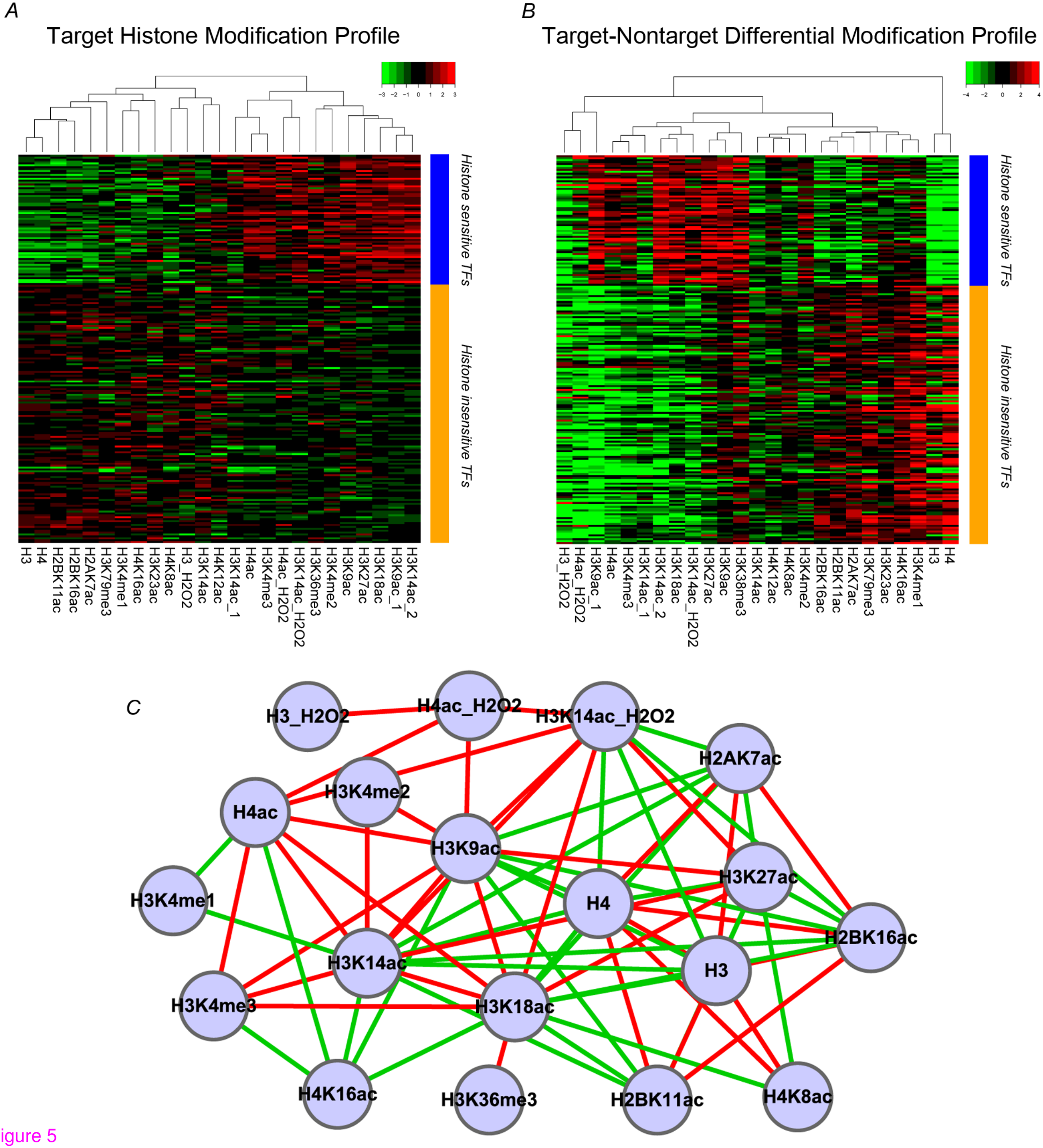
A
Target Histone Modification Profile
−3 −2 −1 0 1 2 3
Histone sensitive TFs
Histone insensitive TFs
H3
H4
H2BK11ac
H2BK16ac
H2AK7ac
H3K79me3
H3K4me1
H4K16ac
H3K23ac
H4K8ac
H3_H2O2
H3K14ac
H4K12ac
H3K14ac_1
H4ac
H3K4me3
H4ac_H2O2
H3K14ac_H2O2
H3K36me3
H3K4me2
H3K9ac
H3K27ac
H3K18ac
H3K9ac_1
H3K14ac_2
B
Target-Nontarget Differential Modification Profile
−4 −2 0 2 4
Histone sensitive TFs
Histone insensitive TFs
H3_H2O2
H4ac_H2O2
H3K9ac_1
H4ac
H3K4me3
H3K14ac_1
H3K14ac_2
H3K18ac
H3K14ac_H2O2
H3K27ac
H3K9ac
H3K38me3
H3K14ac
H4K12ac
H4K8ac
H3K4me2
H2BK16ac
H2BK11ac
H2AK7ac
H3K79me3
H3K23ac
H4K16ac
H3K4me1
H3
H4
C
H3_H2O2
H4ac_H2O2
H3K14ac_H2O2
H2AK7ac
H3K4me2
H4ac
H3K9ac
H3K27ac
H4
H3K4me1
H2BK16ac
H3K14ac
H3
H3K4me3
H3K18ac
H4K16ac
H3K36me3
H2BK11ac
H4K8ac

Figure 5

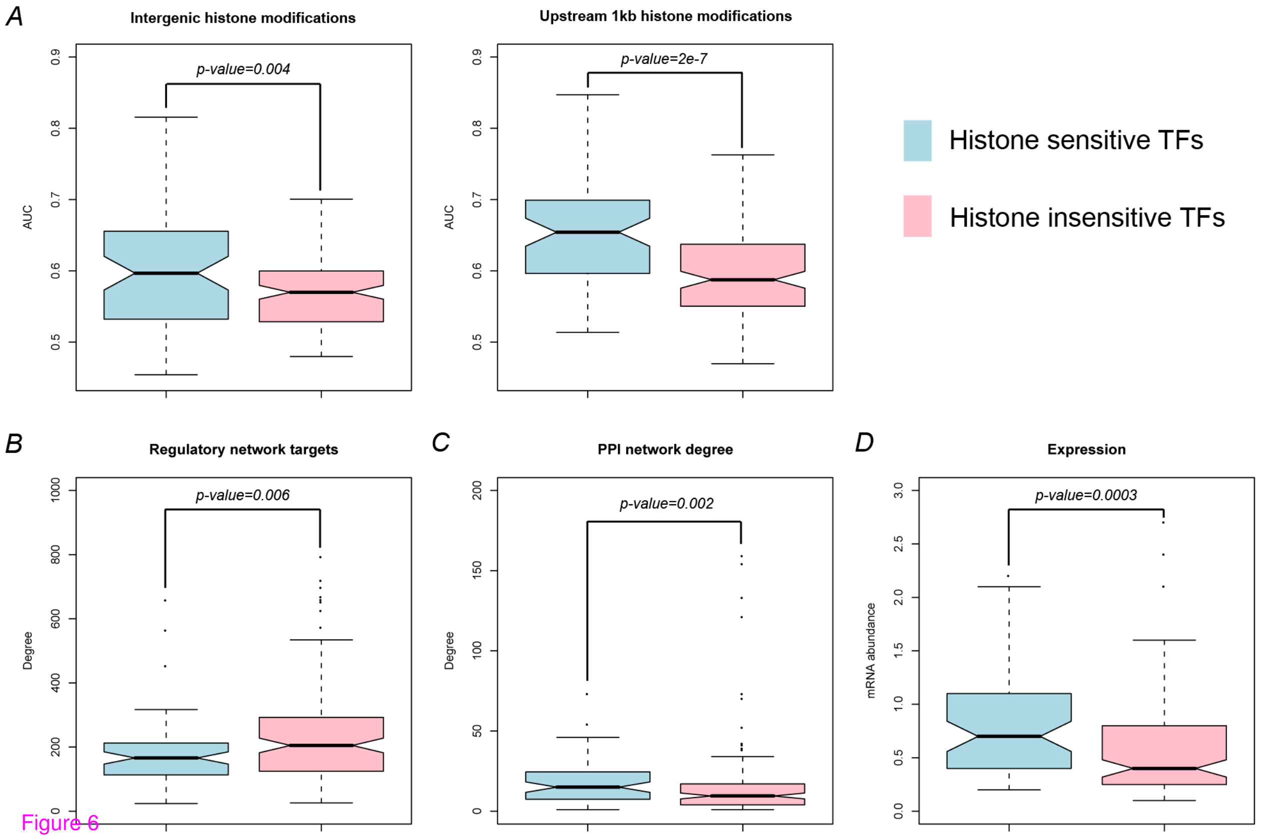
A
Intergenic histone modifications
AUC
0.5
0.6
0.7
0.8
0.9
p-value=0.004
Upstream 1kb histone modifications
AUC
0.5
0.6
0.7
0.8
0.9
p-value=2e-7
Histone sensitive TFs
Histone insensitive TFs
B
Regulatory network targets
Degree
0
200
400
600
800
1000
p-value=0.006
C
PPI network degree
Degree
0
50
100
150
200
p-value=0.002
D
Expression
mRNA abundance
0.0
0.5
1.0
1.5
2.0
2.5
3.0
p-value=0.0003

Figure 6

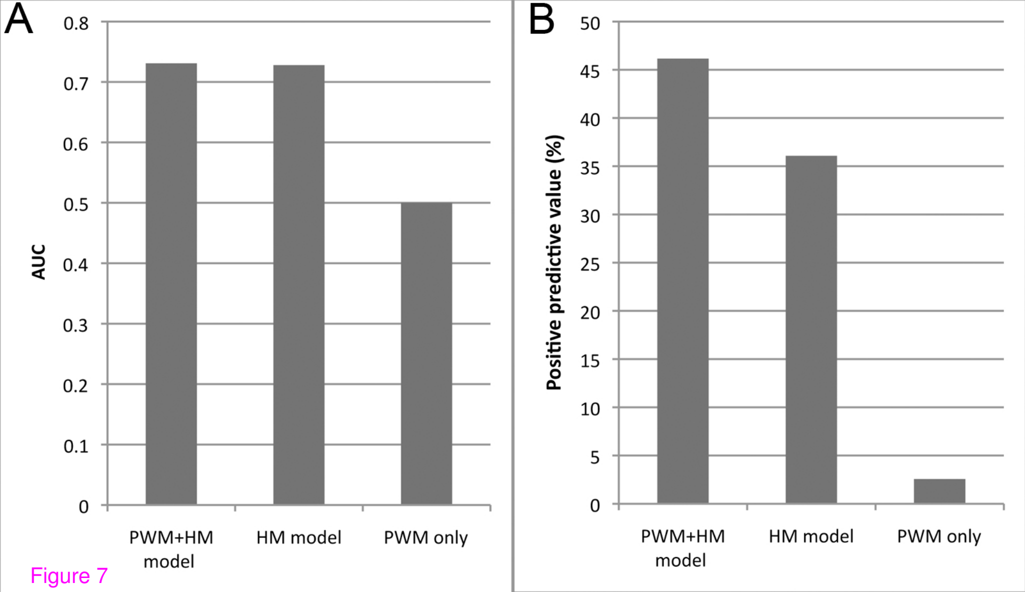
A
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
AUC
PWM+HM model
HM model
PWM only
B
50
45
40
35
30
25
20
15
10
5
0
Positive predictive value (%)
PWM+HM model
HM model
PWM only

Figure 7

Additional files provided with this submission:

Additional file 1: Table S1.doc, 44K
http://genomebiology.com/imedia/1485415472627339/supp1.doc
Additional file 2: Table S2.doc, 42K
http://genomebiology.com/imedia/1264968825627339/supp2.doc
Additional file 3: Table S3.doc, 45K
http://genomebiology.com/imedia/6566813162733932/supp3.doc
Additional file 4: Table S4.doc, 67K
http://genomebiology.com/imedia/1684927288627339/supp4.doc
Additional file 5: Table S5.doc, 127K
http://genomebiology.com/imedia/3367737256273393/supp5.doc
Additional file 6: TableS6.pdf, 49K
http://genomebiology.com/imedia/1473162938627339/supp6.pdf
Additional file 7: Figure S1.pdf, 49K
http://genomebiology.com/imedia/4038558876273393/supp7.pdf