ABSTRACT

Motivation: There is accumulating evidence that the chromatin environment of transcription factor (TF) binding sites in promoter regions has a critical influence on their regulatory potential. Recent studies have mapped TF binding sites and nucleosome positions throughout the yeast genome; however, there is a lack of computation tools to integrate these data types.

Results: We have developed the Ceres software to facilitate the integrated analysis of TF binding sites and nucleosome positions in the model eukaryote S.cerevisiae. Ceres enables users to dynamically display the spatial organization of TF binding sites and nucleosome positions of individual genes, or the average profiles for large gene sets. Ceres provides novel statistical tools to test for the enrichment of TF binding sites and chromatin environments for user-selected gene sets. Ceres also enables users to search the genome for combinations of TF binding sites that are associated with specific chromatin environments. Preliminary analysis using the Ceres software indicates that functional and conserved TF binding sites are often associated with specific chromatin environments. Availability: http://bioinformatics1.smb.wsu.edu/Ceres

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Gene transcription is regulated by short DNA sequences in promoter regions that typically encode binding sites for sequence-specific transcription factor (TF) proteins. Experimental and computational methods have been developed and employed to map transcription factor (TF) binding sites in many eukaryotic genomes. TF binding sites have been most thoroughly mapped in the yeast (Saccharomyces cerevisiae) genome. For example, a groundbreaking study used chromatin immunoprecipitation microarray (ChIP-chip) experiments and computational approaches to map bound and evolutionarily conserved TF binding sites in the yeast genome for 105 distinct yeast TFs (Harbison et al., 2004). A number of databases and web servers have been developed to facilitate the analysis of yeast TF binding sites, including SCPD (Zhu and Zhang, 1999), YEASTRACT (Monteiro et al., 2008; Teixeira et al., 2006) and MYBS (Tsai et al., 2007).

The nucleosome environment of TF binding sites can strongly influence how said binding sites regulate gene transcription (Morse, 2007). In simple cases, the positioning of TF binding sequences relative to nucleosomes has qualitative effects on gene transcription. For example, the location of a TF binding sequence in an accessible nucleosome-free region (NFR) presumably would enhance TF binding and subsequent gene activation. The chromatin environment of TF binding sites can also affect quantitative features of gene transcription. For example, a recent study by O’Shea and colleagues (Lam et al., 2008) investigated the effects of chromatin environment on the function of the Pho4 TF in regulating the expression dynamics of phosphate response (PHO) genes. They discovered that Pho4 binding sites located in different chromatin environments had quantitatively different effects on the activation threshold and maximal induction level of PHO gene transcription (Lam et al., 2008). Indeed, quantitative modeling of PHO gene expression is facilitated by the integration of information of both TF binding sites and nucleosome positions (Kim and O’Shea, 2008).

These and other studies have demonstrated that the nucleosome environment has a critical influence on the regulatory potential of TF binding sites. A number of studies have mapped nucleosome positioning in yeast throughout the genome under normal conditions (Lee et al., 2007; Mavrich et al., 2008; Shivaswamy et al., 2008; Yuan et al., 2005), and in response to environmental perturbation (i.e. heat shock) (Shivaswamy et al., 2008). Bioinformatics tools are available to analyze nucleosome positioning and other chromatin data (Albert et al., 2008; Bock et al., 2009; Kuhn et al., 2009; O’Connor and Wyrick, 2007). However, there is a critical lack of tools to facilitate the integrated analysis of TF binding site data in conjunction with the respective chromatin environment of the binding sites.

To meet this need, we have developed the Ceres database and software tools. The Ceres database contains predicted, conserved and experimentally identified binding sites throughout the yeast genome for 105 distinct yeast TFs (Harbison et al., 2004). Importantly, each binding site has been classified according to its chromatin environment (nucleosome interior, linker DNA, or NFR) in each of three genome-wide nucleosome-mapping data sets (Lee et al., 2007; Shivaswamy et al., 2008). This classification scheme

*To whom correspondence should be addressed.
enables the user to perform integrated analyses of TF binding sites and chromatin environments. Here we describe the novel analysis, visualization and mining tools provided by the Ceres software platform. Using these tools, we show that the chromatin environment of TF binding sites is strongly correlated with the conservation, binding state and functionality of these sites.

2 METHODS

2.1 Data sources and integration

Ceres contains a comprehensive and high quality set of yeast TF binding sites identified by Young and colleagues (Harbison et al., 2004). The TF binding sites were identified by experimental ChIP-chip studies in yeast, and by sequence conservation among related yeast species (Harbison et al., 2004). Altogether, binding site data was available for 105 yeast TFs. Each TF binding site was mapped to the promoter region(s) in which it was located, and classified according to its sequence conservation (‘High’, ‘Medium’, or ‘None’) and ChIP-chip binding affinity (‘High’, ‘Medium’, or ‘None’), following previously described methods (see Supplementary Methods section for details).

Ceres currently contains three genome-wide nucleosome positioning data sets. The first data set is from MNase-microarray analysis (Lee et al., 2007) of nucleosome positions from yeast in normal (YPD 30°C) growth conditions. Two other data sets are from MNase-seq analysis of nucleosome positions from yeast in normal (YPD 30°C) and heat-shocked (YPD 39°C for 15 min) growth conditions (Shivaswamy et al., 2008).

We developed a novel classification scheme for connecting DNA sequence elements with nucleosome positioning data. In this scheme, TF binding sites were classified according to their chromatin environment in each of the nucleosome positioning data sets (Supplementary Fig. 1). Not only was the chromatin environment recorded, but the unique ID of the relevant nucleosome was annotated to that TF binding site in the ‘Chromatin Environment’ database table (Supplementary Fig. 2), thus hard-coding direct links between TF binding sites and chromatin data. TF binding sites whose midpoint was located inside a nucleosome were classified as having a nucleosome ‘interior’ chromatin environment (Supplementary Fig. 1A). TF binding sites located between closely spaced nucleosomes were classified as having a ‘linker’ chromatin environment (Supplementary Fig. 1B). TF binding sites located entirely inside a NFR (defined as a continuous segment of more than 147 bp of nucleosome free DNA sequence; see Supplementary Methods section) were classified as having a ‘nucleosome free’ chromatin environment (Supplementary Fig. 1C). As we describe below, this classification scheme enables Ceres users to perform complex queries of promoter sequence elements and chromatin features. Additional details about software design, data filtering and validation, and statistical analysis tools and methods can be found in the Supplementary Materials and Methods section.

3 RESULTS

3.1 Analysis of TF binding site chromatin environments

We tested whether TF binding sites in the yeast genome showed a bias toward residing in particular chromatin environments. We used Ceres to calculate the chromatin environment for each TF binding site in the yeast genome, and compared the observed distribution of chromatin environments to a background distribution of randomly permuted TF binding sites. If we consider the TF binding sites in aggregate, we observed a slight, but statistically significant, enrichment of TF binding sites occurring in NFRs compared to the background distribution. This trend was observed in the normal MNase-microarray nucleosome data set (P < 0.002), and to a lesser extent in the heat shock and normal MNase-seq nucleosome data sets (P < 0.002 in both cases). These observations confirm the results of previous studies, which have shown that TF binding sites tend to be highly enriched in NFRs of yeast promoters (Lee et al., 2007; Narlikar et al., 2007; Yuan et al., 2005).

Intriguingly, the bias toward the nucleosome free chromatin environment was more pronounced for TF binding sites that were conserved in other yeast species (Fig. 1A) or whose promoters were bound by the respective TF (Fig. 1B). A corresponding depletion of binding sites present in the nucleosome interior chromatin environment was observed in the bound and conserved data sets (Fig. 1). Similar results were obtained regardless of which nucleosome positioning data set was used (data not shown). This analysis suggests that sites whose promoter is bound by the cognate TF or that are evolutionarily conserved are associated with specific chromatin environments.

As expected from the aggregate analysis, the binding sites of many individual TFs were biased toward NFRs (data not shown). The binding site distribution of the Rap1 TF, which regulates the expression of ribosomal protein genes, is shown as an example (Supplementary Fig. 3). Rap1 binding sites, particularly sites that are conserved or bound by Rap1, were strongly biased towards NFRs, as anticipated by previous studies (Bernstein et al., 2004; De Santis et al., 2002; Yu and Morse, 1999). However, the binding sites of some TFs show enrichment in other chromatin environments. For example, sites for the Ume6 transcriptional repressor show a dual preference for linker regions and NFRs (Supplementary Fig. 4A).

3.2 Functional TF binding sites are associated with specific chromatin environments

We hypothesized that the chromatin environment might distinguish functional TF binding sites from non-functional binding sites. As a simple test of this hypothesis, we analyzed microarray data from strains in which yeast TFs were mutated or down-regulated (Hu et al., 2007). An example of this analysis for the Rap1 TF is shown in Figure 2A.

Rap1 target genes were classified into nucleosome interior, linker DNA, and nucleosome free categories based on the chromatin environment of the Rap1 binding site present in the promoter of the target gene. A Wilcoxon rank sum test was used to test whether any class of Rap1 target genes showed altered expression in a Rap1 depletion strain [data derived from reference (Hu et al., 2007)]. We found that genes with Rap1 binding sites located in NFRs were significantly down-regulated when the Rap1 TF was depleted (P < 10^{-5}). This result is illustrated in Figure 2A, which plots, for the Rap1 depletion microarray data set, the relative change of expression of each set of target genes (calculated as the deviation of the rank sum of the target genes from the mean; see Supplementary Methods section). A negative change indicates that depletion of the Rap1 TF causes a corresponding decrease in target gene expression. Such a result indicates that Rap1 is required for expression of at least a subset of the target genes, and thus that the target gene set contains a significant fraction of functional Rap1 binding sites.

In contrast, the expression of genes with Rap1 binding sites in other chromatin environments (i.e. linker or nucleosome interior) was not significantly affected in the Rap1 depletion strain (Fig. 2A; P > 0.05). Similar results were obtained whether we analyzed all predicted Rap1 binding sites, or just those binding sites whose
Fig. 1. Conserved or bound TF binding sites are enriched in the nucleosome free chromatin environment. (A) Frequency of conserved TF binding sites among different chromatin environments. Frequencies were calculated from the aggregate binding sites of all 105 TFs in Ceres. The background frequencies were calculated from randomly permuted TF binding site positions (‘Supplementary Materials’ section). Chromatin environments were assigned using the normal MNase-microarray nucleosome data set (Lee et al., 2007). Medium conservation indicates the binding site sequence was conserved in at least one other related yeast species. High conservation indicates the binding site was conserved in two or more additional yeast species. (B) Aggregate distribution of ‘bound’ TF binding sites, analyzed as in part A. Medium binding corresponds to a ChIP-chip P-value < 0.005. High binding to a ChIP-chip P-value < 0.001.

Fig. 2. The chromatin environment of binding sites demarcates functional from non-functional binding sites. (A) Analysis of gene expression changes in a rap1 deletion strain. Rap1 target genes were classified according to the chromatin environment of the corresponding binding site (in the normal MNase-microarray nucleosome data set). The change in expression of each category of target genes in the rap1 deletion strain was displayed as a percent change in the rank sum statistic (ΔT; ‘Supplementary Materials’ section). P-values were calculated using the Wilcoxon rank sum test (‘Supplementary Materials’ section). (B) Summary of TF binding sites showing functional associations with target gene expression when located in specific chromatin environments. A Wilcoxon rank sum test was used to identify significant associations between binding site chromatin environment categories and changes in gene expression in the corresponding mutant TF data set (Hu et al., 2007). Significant associations were identified using a Bonferroni-corrected P-value threshold of 0.05. Names with asterisks indicate TFs that negatively regulated target gene expression; names without asterisks indicate positive regulation of target gene expression.

promoter was bound by Rap1 or whose sequences were conserved in other yeast species (Fig. 2A). This analysis suggests that functional Rap1 binding sites are associated with NFRs.

Similar analysis was performed on 99 yeast TFs in which mutant expression data were available. The results, which are summarized in Figure 2B, indicated that the binding sites for 19 of the 99 TFs were significantly associated with a change in expression in the TF mutant or deletion strain in one or more chromatin environments. While this percentage of TFs (~19%) seems relatively low, it should be noted that many of the TFs may not actively regulate transcription in YPD media, in which most of the TF mutant experiments were conducted (Hu et al., 2007); and that a relatively stringent P-value
The positions of the Pho4 and Fkh1 TF binding sites are indicated. The user can input up to 100 gene names or accessions and select the TF binding site criteria (i.e. ‘Medium’ binding and conservation stringency) and nucleosome data set to visualize. An example of the compact visualization output, in which Ceres was used to visualize the PHO5 gene promoter, is shown in Supplementary Figure 5A. The positions of the Pho4 and Fkh1 TF binding sites are indicated by color-coded vertical lines. A TF table (Supplementary Fig. 5B) provides a key to the color-coded TF binding sites and lists the number of binding sites, and the significance (P-value) of their association with the selected promoter(s). The nucleosome positions are indicated by the gray-scale ovals. For this data set (Lee et al., 2007), the nucleosomes are represented as either well positioned (dark gray) or weakly positioned (light gray).

3.3 Visualization tools

The Ceres database contains tools to visualize TF binding sites and nucleosome positions in multiple user-selected promoter sequences. The user can input up to 100 gene names or accessions and select the TF binding site criteria (i.e. ‘Medium’ binding and conservation stringency) and nucleosome data set to visualize. An example of the compact visualization output, in which Ceres was used to visualize the PHO5 gene promoter, is shown in Supplementary Figure 5A. The positions of the Pho4 and Fkh1 TF binding sites are indicated by color-coded vertical lines. A TF table (Supplementary Fig. 5B) provides a key to the color-coded TF binding sites and lists the number of binding sites, and the significance (P-value) of their association with the selected promoter(s). The nucleosome positions are indicated by the gray-scale ovals. For this data set (Lee et al., 2007), the nucleosomes are represented as either well positioned (dark gray) or weakly positioned (light gray).

The Promoter Display tool was used to visualize the nucleosome positions among multiple data sets. The Promoter Display tool was used to visualize the pattern of nucleosome positioning at the HSP10 gene (Fig. 3A), which encodes a small heat-shock protein, in normal and heat-shocked cells [15 min at 39°C; (Shivaswamy et al., 2008)]. The nucleosome positions in normal and heat-shocked cells are depicted as color-coded ovals, based on a quantitative measure of how strongly the nucleosomes are positioned (see Shivaswamy et al. (2008) for more details). Bound TF binding sites, including the Heat-Shock Factor 1 (Hsf1) binding sites, are depicted as color-coded vertical lines in the HSP10 promoter sequence (Fig. 3A). Inspection of Figure 3A indicates that nucleosome depletion in the HSP10 promoter is considerably greater upon heat shock, indicating that significant chromatin remodeling may be occurring when HSP10 transcription is significantly induced upon heat shock (>4-fold; (Gasch et al., 2000)). In contrast, there is relatively little remodeling of nucleosome positions at the promoter or coding region of GAL10 (Fig. 3B), which encodes a galactose metabolic enzyme.

The Chromatin Trend tool can also be used to display the average nucleosome profile for a set of user-selected genes. This profile can be based on the average nucleosome positioning score or simply the average nucleosome occupancy (positioning scores are ignored). The Chromatin Trend tool was used to visualize the average nucleosome occupancy of ribosomal protein genes that contained a Rap1 binding site within a 500bp promoter region (medium binding stringency; 92 genes total). In the output, the average nucleosome profile is plotted as a solid black line; the gray shading indicates the standard deviation from this average among the set of 92 selected genes (Fig. 4A). Statistical analysis indicates that the nucleosome profiles of these genes were significantly correlated (r = 0.575, P = 2 × 10^{-5}). Inspection of the data indicates that a large nucleosome depleted region is localized in the proximal promoter between −50 and −300bp. For comparison, we used Ceres to analyze the average nucleosome profile of yeast ribosomal protein genes that did not contain a bound Rap1 binding site within the 500bp proximal promoter region (127 genes). This display indicated that ribosomal protein genes not bound by Rap1 contain a much smaller nucleosome depleted region, centered at approximately −130 bp (Fig. 4B). Hence, Rap1 binding is correlated with a significant expansion of the promoter proximal nucleosome depleted region, as anticipated by previous studies (Bernstein et al., 2004; De Sanctis et al., 2002; Yu and Morse, 1999).

Intriguingly, the distribution of Rap1 binding sites in the bound ribosomal protein genes is centered at −300 bp (Supplementary Fig. 6), correlating with the expansion of the nucleosome depleted region in these promoters. Functional analysis using the Ceres gene ontology mining tool indicates that the Rap1-bound ribosomal protein genes primarily encode cytoplasmic ribosome subunits, while a majority of the unbound ribosomal protein genes code for mitochondrial ribosome subunits (Supplementary Fig. 7). These results are in accordance with a previous study, which showed
that the expression of the cytoplasmic translational machinery is regulated differently than that of the mitochondrial translational machinery (Jimenez et al., 2003). Our analysis indicates that not only do the cytoplasmic and mitochondrial ribosomal genes differ in their regulation by Rap1, but also differ in the nucleosome organization of their respective promoter regions.

3.4 Statistical analysis of enriched TF binding sites and chromatin features

An important feature of the Ceres software is the ability to mine for promoter elements or chromatin features associated with selected gene sets. This functionality is provided through the Gene Set tool. The Gene Set tool reports enriched or depleted TF binding sites present in the promoter regions among a set of user selected genes. The tool also reports the statistical significance of the observed enrichment or depletion. In addition, the Gene Set tool reports if particular chromatin environments or gene ontology terms are associated with the selected gene set.

As an example, we used the Gene Set tool to analyze a set of 41 genes that were identified in a proteomics study as being highly induced during heat-shock (Boy-Marcotte et al., 1999). This gene set is enriched for the ‘cellular response to heat’, ‘response to stress’ and ‘unfolded protein binding’ gene ontology categories ($P < 10^{-10}$), among others, as expected. We tested for bound TF binding sites that were enriched among the promoter sequences of these genes. The output identified the Hsf1 ($P < 10^{-10}$), Msn2 ($P < 10^{-10}$) and Msn4 ($P < 10^{-10}$) TF binding sites (medium binding criteria) as being significantly enriched (Supplementary Fig. 8A). These results are in accordance with previous studies, which have shown that these TFs...
regulate heat-shock response (Martínez-Pastor et al., 1996; Schmitt and McIntee, 1996; Wu, 1995).

The Site Environments tool can be used to test whether TF binding sites in specific chromatin environments are enriched in the selected gene set. We applied this analysis to the heat-shock gene set. The results indicated that in normal growth conditions the Hsf1, Msn2 and Msn4 binding sites were enriched in both the nucleosome interior ($P < 10^{-5}$) and nucleosome free ($P < 10^{-8}$) chromatin environments (Supplementary Fig. 8B). In contrast, in the heat-shock conditions, the Hsf1, Msn2 and Msn4 binding sites were enriched predominately in the nucleosome free chromatin environment ($P < 10^{-10}$; Supplementary Fig. 8C).

Ceres can also test for statistically significant enrichment or depletion in the density of positioned nucleosomes in a gene set. Analysis of the 41 heat-shock induced genes indicated that the positioned nucleosome density was significantly depleted in heat-shocked cells ($P < 10^{-6}$; Supplementary Fig. 8D), but not in cells grown in normal conditions (Supplementary Fig. 8D). This depletion could explain the change in Hsf1, Msn2 and Msn4 binding site chromatin environments (see above).

3.5 Data search tools

The Ceres software enables users to search for genes containing user-specified combinations of TF binding sites. Importantly, users can also search for TF binding sites that are associated with specific chromatin environments. To demonstrate the utility of this tool, we searched for genes containing a bound Bas1 TF binding site located in a nucleosome-free chromatin environment (Fig. 5A). This query identified 31 genes, which included many known Bas1 target genes, including genes involved in histidine and nucleotide biosynthesis (e.g. HIS4, ADE2, ADE4, ADE5,7, ADE6, ADE8, ADE12, ADE13, ADE17) (Daiguan-Formier and Fink, 1992; Tice-Baldwin et al., 1989). Similar queries identified bound Bas1 TF binding sites associated with nucleosome interior chromatin environments (15 genes) and linker environments (eight genes). Unlike the genes containing nucleosome-free associated Bas1 TF binding sites, the other queries were not enriched for known Bas1 target genes. For example, only two of the known Bas1 target genes were found in the nucleosome interior query, and both of these genes (ADE8 and ADE17) also contained at least one additional Bas1 TF binding site associated with a NFR (data not shown).

Further analysis revealed that genes with Bas1 TF binding sites in a nucleosome-free chromatin environment were significantly downregulated in expression in a bas1 mutant strain (Fig. 5B). In contrast, genes with Bas1 TF binding sites in other chromatin environments did not show altered expression (nucleosome interior) or showed slightly increased expression (linker DNA) in a bas1 mutant strain (Fig. 5B). Again, these results highlight the importance of the chromatin environment for demarcating functional TF binding sites.

4 DISCUSSION

We have described a novel software resource for the integrated analysis of nucleosome positioning and TF binding sites in the promoter regions of yeast genes. While previous tools have been developed to analyze promoter sequences in yeast (Monteiro et al., 2008; Teixeira et al., 2006; Tsai et al., 2007; Zhu and Zhang, 1999), the Ceres software described here is the first analysis tool to tightly integrate the analysis of promoter sequence elements with chromatin data. Importantly, we demonstrate that the chromatin environment of TF binding sites is strongly correlated with the conservation, binding state and functionality of these sites. For this reason, we anticipate that the integrated analysis, visualization and
data mining tools provided in Ceres will greatly facilitate the study of transcriptional regulation in the model eukaryote \textit{S. cerevisiae}.

Our analysis suggests that many TF binding sites functionally regulate transcription only when located in specific chromatin environments (e.g. nucleosome free). In the case of Rap1, and perhaps other TFs, it is likely that the functional TF binding sites themselves stimulate the formation of the preferred chromatin environment [e.g. nucleosome depleted region; (Bernstein et al., 2004; Buck and Lieb, 2006; De Sanctis et al., 2002; Yu and Morse, 1999)]. In such cases, the chromatin environment of the TF binding sites serves as a potential marker for distinguishing functional from non-functional binding sites. In other cases, however, the nucleosome environment may directly regulate the potential of the binding site to be bound by the TF, and thus activate (or repress) transcription (Morse, 2007).

The current version of the Ceres software tool focuses on the nucleosome environment of TF binding sites. However, there is evidence that the local pattern of histone modifications may also influence promoter elements and TF binding sites. For example, a recent study suggests that histone H3 K4 mono-methylation is associated with active enhancer sequences in the human genome (Buck and Lieb, 2006). It is possible that histone modifications and other chromatin features (e.g. histone variant incorporation) in the future.

**ACKNOWLEDGEMENTS**

The authors thank McKenna Kyress for helpful comments on the manuscript, and Dr Vishwanath Iyer for providing additional information about nucleosome data sets.

**Funding:** National Science Foundation (grant DBI-0605016) to R.T.M. and J.J.W.

**Conflict of Interest:** None declared.

**REFERENCES**


Tice-Baldwin,K. et al. (1989) BAS1 has a Myb motif and activates HIS4 transcription only in combination with BAS2. Science, 246, 931–935.


