A Library of Yeast Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting Nucleosome Exclusion at Promoters

Gwenael Badis,1 Esther T. Chan,2 Harm van Bakel,1 Lourdes Pena-Castillo,1 Desiree Tillo,2 Kyle Tsui,3 Clayton D. Carlson,4 Andrea J. Gossett,6 Michael J. Hasinoff,4 Christopher L. Warren,4 Marinella Gebbia,1 Shaheynoor Talukder,1 Ally Yang,1 Sanie Mnaimneh,1 Dimitri Terterov,1 David Coburn,1 Ai Li Yeo,7 Zhen Xuan Yeo,7 Neil D. Clarke,7 Jason D. Lieb,6 Aseem Z. Ansari,4,5 Corey Nislow,1,2 and Timothy R. Hughes1,2,*

1Banting and Best Department of Medical Research
2Department of Molecular Genetics
3Department of Pharmaceutical Sciences
University of Toronto, Toronto, ON M5S 3E1, Canada
4Department of Biochemistry
5The Genome Center
University of Wisconsin-Madison, Madison, WI 53706, USA
6Department of Biology, Carolina Center for Genome Sciences and Lineberger Comprehensive Cancer Center, CB# 3280, 408 Fordham Hall, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3280, USA
7Computational and Systems Biology, Genome Institute of Singapore, 60 Biopolis Street, Singapore, 138672, Republic of Singapore
*Correspondence: t.hughes@utoronto.ca
DOI 10.1016/j.molcel.2008.11.020

SUMMARY

The sequence specificity of DNA-binding proteins is the primary mechanism by which the cell recognizes genomic features. Here, we describe systematic determination of yeast transcription factor DNA-binding specificities. We obtained binding specificities for 112 DNA-binding proteins representing 19 distinct structural classes. One-third of the binding specificities have not been previously reported. Several binding sequences have striking genomic distributions relative to transcription start sites, supporting their biological relevance and suggesting a role in promoter architecture. Among these are Rsc3 binding sequences, containing the core CGCG, which are found preferentially ~100 bp upstream of transcription start sites. Mutation of RSC3 results in a dramatic increase in nucleosome occupancy in hundreds of proximal promoters containing a Rsc3 binding element, but has little impact on promoters lacking Rsc3 binding sequences, indicating that Rsc3 plays a broad role in targeting nucleosome exclusion at yeast promoters.

INTRODUCTION

The targeting of a transcription factor (TF) to specific genomic loci is determined by its DNA-binding activity, which is typically encoded by a conserved DNA-binding domain (DBD), together with cofactor interactions and the chromatin state of potential targets (Barrera and Ren, 2006). A foundation of any complete and accurate model of transcriptional regulation will be knowledge of the sequence specificities of DNA-binding proteins (Beer and Tavazoie, 2004; Segal et al., 2008). Despite intense study, there is currently no organism for which a complete encyclopedia of such TF sequence specificities exists. Even in the well-studied yeast S. cerevisiae, prior to this study, binding sequences were understood with confidence for only about half of its ~200 TFs. The majority of yeast TFs have been analyzed by ChIP-chip, but even when assayed under several different growth conditions (Harbison et al., 2004), these experiments often fail to identify either significant binding events or associated motifs, presumably because the TF is not binding DNA under the assay conditions. Further complicating de novo motif identification is the possibility that ChIP-chip and related techniques (e.g., ChIP-seq) may identify binding sequences for cofactors rather than the intended TF (Carroll et al., 2005). In some cases, it may be possible to infer TF sequence preferences on the basis of similarity among DBDs or identities of DNA-contacting residues (Berger et al., 2008; Wolfe et al., 2000), but for no DBD class is there a complete and accurate combinatorial code that dictates sequence specificity.

Incomplete knowledge of TF binding specificities hinders our understanding of basic mechanisms of transcription and nuclear organization. For example, RSC (remodel the structure of chromatin) is an abundant nuclear protein complex with a role in nucleosome organization at many yeast promoters (Cairns et al., 1996; Ng et al., 2002; Parnell et al., 2008). RSC contains two Gal4-class transcription factor-like proteins (Rsc3 and Rsc30) with very similar amino acid (AA) sequences but apparently different cellular functions (Angus-Hill et al., 2001; Wilson et al., 2006). Neither Rsc3 nor Rsc30 has known sequence specificity, and the mechanisms that target RSC to individual loci remain poorly defined.
More generally, the mechanisms responsible for nucleosome-free regions (NFRs) in yeast promoters are incompletely understood. Current models of intrinsic nucleosome-DNA preference do not explain all of the observed nucleosome positioning and occupancy (Lee et al., 2007; Segal et al., 2006; Yuan and Liu, 2008). TF binding sequences are often enriched in NFRs (Lee et al., 2007; Liu et al., 2006), and in at least some cases, TFs make strong contributions to the local chromatin landscape. For example, Abf1, Reb1, and Rap1 are found frequently in yeast promoters and are able to define chromatin domains and enable activation or repression by other TFs in diverse pathways (Chashman et al., 1990; Elemento and Tavazoie, 2005; Fourel et al., 2002; Planta et al., 1995). Abf1, Reb1, or Rap1 binding sites are known to bind to specific DNA sequences and regulate transcription (Table S1 available online). We were able to clone 207 of the 218 DBDs (or full-length proteins in the event that the DBD is unknown) as GST and/or MBP fusion proteins and, upon expression, obtained a protein for 195. We analyzed the sequence specificities of these 195 using at least one of three methods: (1) Protein Binding Microarrays (PBMs), in which the proteins are applied to an Agilent microarray consisting of 40,330 double-stranded 60-mers, each containing a unique 35-mer, such that all 10-mers are represented once and only once (Berger et al., 2006; Mintseris and Eisen, 2006); (2) Cognate Site Identifier (CSI) (Warren et al., 2006), in which proteins are applied to a NimbleGen array of stranded 60-mers, each containing a unique 35-mer, such that all 10-mers are represented once and only once (Berger et al., 2006; Mintseris and Eisen, 2006); (3) DNA immunoprecipitation-chip (DIP-chip) (Liu et al., 2005), in which purified transcription factor, bound to yeast genomic DNA, is immunoprecipitated in vitro and analyzed using microarrays.

Table S1 and our project website contain a summary of which proteins were analyzed by each method and details on motif derivation. The majority of data produced resulted from PBMs (Berger et al., 2006). To discover the motifs preferentially bound by each protein in the PBM experiments, we first took the median signal intensity across the array from the 32 spots containing each protein. We then sought DNA sequence motifs (Position Weight Matrices (PWMs), following Granek and Clarke (2005)). Z scores for each factor (see Experimental Procedures for details). The 112 resulting motifs identified are shown in Figure 1.

Figure 1. Motifs Identified in Our Study

Motifs represent Position Weight Matrices (PWMs), following Granek and Clarke (2005).

Creation of a Library of Sequence Specificities for 112 Yeast TFs

We began by creating a list of 218 yeast proteins that either contain a TF DBD or are known to bind to specific DNA sequences and regulate transcription (Table S1 available online). More than half of the proteins in our library bind to specific DNA sequences and regulate transcription. We included 176 TFs and 42 non-TFs that bind to DNA. More than half of the non-TFs bind to DNA and are regulatory proteins. We also included 218 yeast promoters as well as transcript abundance data for each gene. The promoters were obtained from the Saccharomyces Genome Database (SGD). The promoters were divided into two sets: (1) promoters with known TF binding sites and (2) promoters without known TF binding sites. We then measured the sequence preferences of the 112 motifs identified in our study.

RESULTS

Creation of a Library of Sequence Specificities for 112 Yeast TFs

We began by creating a list of 218 yeast proteins that either contain a TF DBD or are known to bind to specific DNA sequences and regulate transcription (Table S1 available online). More than half of the proteins in our library bind to specific DNA sequences and regulate transcription. We included 176 TFs and 42 non-TFs that bind to DNA. More than half of the non-TFs bind to DNA and are regulatory proteins. We also included 218 yeast promoters as well as transcript abundance data for each gene. The promoters were obtained from the Saccharomyces Genome Database (SGD). The promoters were divided into two sets: (1) promoters with known TF binding sites and (2) promoters without known TF binding sites. We then measured the sequence preferences of the 112 motifs identified in our study.

Table S1 and our project website contain a summary of which proteins were analyzed by each method and details on motif derivation. The majority of data produced resulted from PBMs (Berger et al., 2006). To discover the motifs preferentially bound by each protein in the PBM experiments, we first took the median signal intensity across the array from the 32 spots containing each protein. We then sought DNA sequence motifs (Position Weight Matrices (PWMs), following Granek and Clarke (2005)). Z scores for each factor (see Experimental Procedures for details). The 112 resulting motifs identified are shown in Figure 1.

Figure 2A illustrates how the PWM-derived scores correlate with the 8-mer-based Z scores obtained for Gzf3 using either PBM or CSI, demonstrating that the imperfect correlation cannot be attributed primarily to measurement noise in the assay or the array platform, because the 8-mer profile is consistent between these two different experiment types, even among less-preferred 8-mers. This observation may reflect shortcomings in the derivation. The majority of data produced resulted from PBMs (Berger et al., 2006). To discover the motifs preferentially bound by each protein in the PBM experiments, we first took the median signal intensity across the array from the 32 spots containing each 8-mer and expressed this as a Z score (Berger et al., 2006). We then sought DNA sequence motifs (Position Weight Matrices (PWMs), following Granek and Clarke (2005)). Z scores for each factor (see Experimental Procedures for details). The 112 resulting motifs identified are shown in Figure 1.

Figure 2A illustrates how the PWM-derived scores correlate with the 8-mer-based Z scores obtained for Gzf3 using either PBM or CSI, demonstrating that the imperfect correlation cannot be attributed primarily to measurement noise in the assay or the array platform, because the 8-mer profile is consistent between these two different experiment types, even among less-preferred 8-mers. This observation may reflect shortcomings in the derivation. The majority of data produced resulted from PBMs (Berger et al., 2006). To discover the motifs preferentially bound by each protein in the PBM experiments, we first took the median signal intensity across the array from the 32 spots containing each 8-mer and expressed this as a Z score (Berger et al., 2006). We then sought DNA sequence motifs (Position Weight Matrices (PWMs), following Granek and Clarke (2005)). Z scores for each factor (see Experimental Procedures for details). The 112 resulting motifs identified are shown in Figure 1.

Figure 2A illustrates how the PWM-derived scores correlate with the 8-mer-based Z scores obtained for Gzf3 using either PBM or CSI, demonstrating that the imperfect correlation cannot be attributed primarily to measurement noise in the assay or the array platform, because the 8-mer profile is consistent between these two different experiment types, even among less-preferred 8-mers. This observation may reflect shortcomings in

Figure 2A illustrates how the PWM-derived scores correlate with the 8-mer-based Z scores obtained for Gzf3 using either PBM or CSI, demonstrating that the imperfect correlation cannot be attributed primarily to measurement noise in the assay or the array platform, because the 8-mer profile is consistent between these two different experiment types, even among less-preferred 8-mers. This observation may reflect shortcomings in
We verified by Electrophoretic Mobility Shift Assay (EMSA) that Stp3 and Yml081w bind to DNA sequences matching 2004). We verified by Electrophoretic Mobility Shift Assay from an ancient whole genome duplication (WGD) (Kellis et al., respectively, their corresponding yeast paralogs that arose are very similar to those we obtained from Stp3 and Zms1, 880 Molecular Cell are very similar to those we obtained from Stp3 and Zms1, PWM and consensus models (Benos et al., 2002). PWMs do, however, identify the best binding sequences in all of our experiments, and since they are compact, intuitive, and compatible with existing analysis techniques, we used PWMs for the remainder of our analyses.

63 of the 112 Motifs in our Library Correspond to Known Motifs
We next asked if the 112 motifs we obtained agree with those previously identified for the same proteins, from either global ChIP-chip analysis (Harbison et al., 2004; Maclsaac et al., 2006), or individual studies in the literature (Nash et al. [2007] and others), by manual comparison of logos, consensus sequences, and individual binding sites (Table S1). Sixty-three of our motifs bear an obvious correspondence to previous information (although not always all previous information), while 11 are inconsistent. For the remaining 38, we did not find any previously known motifs, although most of these motifs we obtained are consistent with expectation in some way (see below).

In Cases of Discrepancies with Existing Data, Evidence Supports the Newly Discovered Motifs
For some of the 11 discrepancies, additional evidence suggests that our measurements are likely to represent at least a correct in vitro monomeric binding sequence (Table S2). For example, our FoxN1 motif is a close match to that of its human homolog, FoxN1 (Schlake et al., 1997). Our motifs for Stp4 and Yml081w are very similar to those we obtained from Stp3 and Zms1, respectively, their corresponding yeast paralogs that arose from an ancient whole genome duplication (WGD) (Kellis et al., 2004). We verified by Electrophoretic Mobility Shift Assay (EMSA) that Stp3 and Yml081w bind to DNA sequences matching our motifs and not those previously described (Figure S1).

A few other discrepancies can be explained by the methodology we employed. For example, the A/T-rich motif we obtained for Sum1 is different from the published motif because when cloning DBDs, we selected the N-terminal AT hook domain rather than the C-terminal fragment that binds the established Sum1 motif but does not contain a known conserved domain (Pierce et al., 2003). Despite this discrepancy, promoter scans with our Sum1 motif do have a high correspondence to ChIP-chip results, suggesting that this additional DNA-binding activity of Sum1 may contribute to targeting in vivo (Spearman correlation p < 10⁻²⁰, Wilcoxon Rank Sum p < 0.000011 with 61 targets defined by Harbison et al. [2004] at p < 0.001).

Other variations from the literature are likely reproducible in vitro phenomena that are characteristic of members of a structural class. Four of the eight GATA-class proteins we analyzed (Ecm23, Srd1, Gat3, and Gat4) bound unexpectedly to sequences resembling the palindrome AGATCT. No binding sequences have been described for three of these four proteins, Ecm23, Srd1, and Gat4, and we know of no other in vitro or in vivo data that confirms or refutes our observations. A noncanonical motif different from AGATCT was derived for the fourth protein, Gat3, on the basis of ChIP-chip and sequence conservation of putative target sites (Maclsaac et al., 2006), and has not been experimentally pursued to our knowledge. Our motif does not correlate with the ChIP-chip data, which is highly enriched for subtelomeric loci. However, we confirmed by EMSA that Gat3 binds the sequence we identified more strongly than the sequence identified by ChIP-chip and that Ecm23 binds to the newly-identified motif (Figure S1).

Three of the discrepancies (Ecm22, Pu3, and Um66) are for Gal4-class proteins, which also have characteristic behavior in our analyses. It appears that our data largely capture monomeric specificities rather than the dimeric motifs typically associated with proteins in this class (MacPherson et al., 2006) (for all DBD classes, we counted correct monomeric specificities as consistent with previous information for dimeric proteins). Still, all but two of the motifs we obtained for Gal4-class proteins do contain the expected CGG core sequence (MacPherson et al., 2006), which is not always the case for the motifs derived from other studies. The capture of monomeric specificities could be a consequence of the domain definitions used for expression or the epitope-tagging strategy. In order to include dimerization contacts, our Gal4-class contacts included 50 AAs of flanking sequence beyond the boundaries of the DBD (or to the end of the protein if within 50 AAs). The choice of flanking sequence length was based on inspection of crystal structures of Gal4-class dimers binding to the DNA. However, the family is structurally diverse in the way the DBD dimerizes, and it may be that for some members of the family, the flanking sequence that was included was insufficient to mediate dimerization. In addition, our constructs are N-terminal GST fusions; Gal4-class DBDs are typically found at the N-terminus of yeast proteins and either dimerization or DNA binding by dimers may be influenced by N-terminal GST tags. The array designs we used may also fail to detect long motifs because the arrays are designed primarily to detect sequences up to ~10 bases (for PBM and CSI). Nonetheless, Gal4-class proteins do sometimes function in vivo as monomers (Kim et al., 2003; Larochelle et al., 2006; Vik and Rine, 2001) and several of our monomeric motifs are enriched in the promoters of functionally-related genes and at specific promoter positions (see below).

Figure 2. Comparison of Motif Representation and Reproducibility of 8-mer Profiles across Platforms
(A) PWM scores (Granek and Clarke, 2005) for all possible 8-mers for the single motif with highest Pearson correlation to the PBM 8-mers, plotted against the Z scores from the PBM. Data are plotted as asinh values, which are similar to natural log, but return real values for negative numbers (by definition, half of all Z scores are negative).
(B) CSI Z scores (combined from up to four array spots containing the 8-mer) versus Z scores from PBM.
Correspondence between Amino Acid Sequence Similarity and DNA-Binding Specificities Supports New Motifs

Based on our examination of literature and databases (described above and in Table S1), we classified 38 of the proteins for which we obtained a motif as having no previously established binding sequences. However, most of the 38 are members of structural classes that have characteristic binding site properties, and many are members of gene families that might be expected to share related sequence specificities. Indeed, most of our new motifs conform to expectation. The C2H2 zinc finger family provides several such examples (Figure 3). All three Mig proteins share virtually identical DNA-binding activities, as expected (Lutfiyya et al., 1998), as do Stp3 and Stp4, as described above. In contrast, C2H2 zinc finger proteins with unique motifs (Azf1, Crz1, Fzf1, Rpn4, Rei1, and Rim101) all have less than 60% identity to any other yeast protein in the DBD. ClustalW-derived phylograms similar to Figure 3 are given for all other structural classes in Figure S2. Three major observations include (1) Two Gal4-class proteins with related DBD sequences, Rsc3 and Rsc30, prefer sites that contain CGCG rather than the CGG typical of this class of proteins. Not coincidentally, perhaps, these two proteins are also unusual in having glycine at a position that is almost always lysine or arginine (corresponding to K20 in the Gal4 DBD). The lysine or arginine normally found at this position is in close proximity to the phosphate backbone in crystal structures of protein-DNA complexes (Figure S3). It is also just two positions C terminal to the residue that makes base-specific contacts to the usual CGG half-site. Thus, the unusual glycine at this position in Rsc3 and Rsc30 may affect the orientation of the domain with respect to DNA, resulting in the unusual DNA-binding specificity discovered here. (2) Dot6 and Ybl054w, a pair of related SANT domain proteins originating from the WGD (Kellis et al., 2004), both bound to sequences containing the core CGATG, which resembles the PAC (Polymerase A and C) motif (Deguard-Chablat et al., 1991). However, we found no evidence indicating that they bind to the promoters of genes containing these motifs (Harbison et al., 2004). (3) We obtained similar motifs containing the core TGTCA for Tos8 and Cup9, a pair of homeodomain proteins originating from the WGD. Neither protein has previously established binding specificity.

Many Motifs Are Enriched Upstream of Functionally Related Genes

We next scanned the yeast genome with the motifs and asked if the potential binding sites for each TF are associated with genes that share functional classes. Twenty-seven of the 112 motifs had a hypergeometric p value of < 5 × 10−3 (corresponding to a Bonferroni-corrected p value of 0.01) for enrichment of at least one GO Biological Process category among the top 100 promoter/motif hits. Expected enrichments include Ste12 (Sterile 12), with “cell-cell fusion” (p < 2.2 × 10−14), and Pdr1 (Pleiotropic Drug Resistance), with “response to drug” (p < 1 × 10−6). Our analysis is consistent with the function of Rgt1 (Restores Glucose Transport 1) as a Gal4-class TF that binds DNA as a monomer in vivo (Kim et al., 2003), since our monomeric motif is associated with “hexose transport” (p < 6.1 × 10−10). Ypr196w and Ydr520c binding sequences were also enriched in the promoters of hexose transporters (p < 2.4 × 10−6; 6.35 × 10−7); the motifs for these proteins are related to that of Rgt1 and the top promoter/motif matches are found in an overlapping, but not identical, set of transporters, suggesting a more complex regulatory network of sugar utilization than is currently known. We were also intrigued to find that the monomeric motif we obtained for Lys14 has the same enrichment in promoters of lysine biosynthesis genes as the established dimeric motif (p < 3.8 × 10−6.
for both), suggesting that both binding modes may be used in vivo.

**Many New Motifs Are Preferentially Found in the NFR**

We next examined how the occurrences of the motifs we discovered were distributed within promoters. Figure 4 (top) shows that most of our 21 monomeric Gal4 motifs occur preferentially in the position of the NFR (approximately –130 to –50, relative to TSS), providing support for their widespread in vivo relevance. Figure 4 (middle) shows 14 motifs we classified as new and unexpected; several of these are also located preferentially in the NFR. The most striking examples are Rsc3 and Rsc30, which share very similar binding preferences to sequences containing CGCG. At a stringent motif score threshold, these sequences are 16-fold more likely to occur in the position of the NFR than they are within genes. Only a handful of other TFs have this extreme bias (Lee et al., 2007), most notably Abf1 and Reb1, which are capable of remodeling chromatin in the vicinity of their binding sites. Among these are 416 of 667 RSC targets defined in Ng et al., using a combined p value cutoff of < 0.01 (the p value of this overlap among 4,947 genes is p < 4.36 \times 10^{-19}). The correspondence to Rsc3 ChIP-chip occupancy (defined in Ng et al. [2002] using a p value cutoff < 0.01) is lower, although still significant (162 out of 293 targets; p < 0.0011). We note, however, along with others (Parnell et al., 2008), that ChIP-chip experiments with RSC subunits, particularly Rsc3, tend to have very low enrichment ratios. One possible explanation, consistent with the activity of RSC as an enzyme that displaces nucleosomes, may be that the association of RSC with target promoters is transient, as may be the case for the DNA-binding TFIIIC module, which also has relatively low ChIP-chip enrichments (Roberts et al., 2003; Soragni and Kassavetis, 2008). We therefore sought an alternative functional assay to ask if Rsc3 binding sites in promoters influence nucleosome occupancy.

**RSC3 Is Required for the Formation of Nucleosome-free Regions at Promoters Containing Rsc3 Binding Sites**

We assayed nucleosome occupancy in the rsc3-1 mutant (Angus-Hill et al., 2001) using MNase digestion mapping and full-genome tiling arrays with 4 nt resolution (Lee et al., 2007). The biochemical defect of rsc3-1 is unknown, but the mutations (M709I and L828S) are outside the DBD (AA1-37). We compared nucleosomal DNA enrichment (i.e., ratio of nucleosomal DNA
versus total genomic DNA) in the rsc3-1 mutant to that in an isogenic wild-type control grown at the same temperature (37 degrees for 6 hr). Figure 5A shows an example locus in which nucleosome depletion over a Rsc3 binding sequence in a promoter region is dependent on RSC3. Figure 5B shows that this phenomenon occurs at many yeast promoters, with a clear preference for the affected region to be located near +100 from TSS. Moreover, the location of the increase in nucleosome occupancy (and the position of the NFR itself) tracks with the Rsc3 binding sequence across hundreds of promoters. Such changes are not observed at promoters that do not contain Rsc3 binding sequences (Figure 5C); in fact, nucleosome occupancy appears to decrease in these promoters, perhaps as a consequence of microarray signal normalization or redistribution of nucleosomes in vivo. This observation illustrates specificity of this phenomenon for Rsc3 binding sequences and not just NFRs in general. Unlike a previous study that used a greater tiling interval on selected promoters to examine the effects of mutating another RSC

Figure 5. Rsc3 Influences Nucleosome Occupancy at Proximal Promoters Containing Rsc3 Binding Sequences
(A) A segment of Chromosome XIII with a Rsc3 binding sequence (gray vertical line) that is depleted in wild-type but occupied in the rsc3-1 mutant. (B and C) Changes in promoter nucleosome occupancy profiles between rsc3-1 and a wild-type control for promoters containing Rsc3 binding sequences (B) or containing Reb1 binding sequences, but not Rsc3 binding sequences (C). Promoters are sorted by the position of the highest scoring Rsc3 or Reb1 binding sequence location in the promoter, which is shown at left in (B) and (C). Additional sites of equivalent PWM score are also indicated.
subunit (Parnell et al., 2008), we saw little or no effect on nucleosome positioning or occupancy at tRNA genes (Figure S4), indicating that the effects we observed are distinct from a general loss of RSC activity. We also surveyed RNA abundance in the rsc3-1 strain using the same arrays and observed a clear trend in which the Pol II promoters with an increase in nucleosome occupancy tend to exhibit lower RNA abundance (Figure 6). Overall, our results are consistent with a function for Rsc3 in nucleosome removal and promoting transcription from Pol II promoters that contain Rsc3 binding sequences in the NFR region.

In order to ask whether the effect of Rsc3 is mediated by Rsc, we compared the relative occupancy of Rsc8 in wild-type and rsc3-1 strains using ChIP-chip. In previous studies (Damelin et al., 2002; Ng et al., 2002; Parnell et al., 2008), Rsc8 has the highest occupancy ratios of any RSC subunit with up to 6-fold enrichment at tRNAs. In our wild-type strain, Rsc8 occupancy ratios are also highest at tRNAs (maximum enrichment 8.5-fold in our analysis, Figure S4), and at Pol II promoters, there is a significant correspondence between Rsc8 occupancy and the Rsc3 motif score (Spearman rank correlation \( p < 1.3 \times 10^{-5} \)). We found that occupancy at tRNAs is not affected by rsc3-1 (Figure S4), suggesting that Rsc is targeted to Pol III transcripts by a RSC3-independent mechanism. Surprisingly, in rsc3-1, we saw a global (albeit modest) increase in occupancy of Rsc8 at Pol II promoters (Figure 6), which could be an indirect effect of the fitness defects seen in rsc3-1 mutant cells (Angus-Hill et al., 2001), and/or the dramatic alterations we observed in chromatin organization and transcript profiles. Nonetheless, the increase is clearly smaller for promoters in which nucleosome occupancy increases in response to rsc3-1 (Figure 6), and it is also smaller for those promoters carrying a Rsc3 sequence (Wilcoxon rank sum test \( p < 2.7 \times 10^{-5} \) among Rsc8-bound promoters, with Rsc3 positives defined as genes with a Rsc3 site in the NFR [−150 to −70]). Together, these observations suggest that Rsc3 may function by targeting RSC but do not rule out the possibility that Rsc3 acts by other mechanisms.

**Other TFs Contribute to Nucleosome Occupancy at Promoters Containing Their Cognate Binding Sequences**

Finally, we asked whether other TFs have an impact on nucleosome occupancy and transcription similar to that observed for Rsc3. Indeed, the correspondence between Rsc3 binding sequences and the impact of the rsc3-1 mutant on nucleosome occupancy in promoters and transcript levels from the corresponding gene is similar to that seen with Abf1 and Reb1 (Figure 6 and Figure S5). Binding sequences for these TFs are found in the proximal promoter of hundreds of yeast genes, and as predicted from their known roles as chromatin modifiers, mutation of each TF results in a specific increase in the occupancy of nucleosomes over the potential binding site (Figure 6), with the most affected NFRs in the mutants typically containing the TF binding sequence. We also analyzed nucleosome occupancy in mutants in the essential DNA-binding proteins Tbf1, Rap1, and Mcm1; all three appear to influence nucleosome occupancy at promoters containing their cognate binding sequences, although the number of promoters affected is smaller than for Rsc3, Abf1, and Reb1 (Figures S5 and S6). By way of comparison, there is no relationship between binding sequences for Cep3, a centromere-binding protein, and nucleosome occupancy at Pol II promoters in a cep3 mutant (Figure 6 and Figure S6). There is, however, a match to the Cep3 motif in all sixteen yeast centromeres, and the array signal in our nucleosome preparations at each centromere is depleted in the cep3 mutant (Figure S7).

**DISCUSSION**

Our in vitro survey of yeast TF-DBD sequence specificities raises the number of yeast TFs with known sequence preference to 174, or ~80% (Table S1). This expanded index of sequence preferences provides a resource for exploration of the function and evolution of gene regulatory networks. Our comparison of predicted promoter preferences to GO categories represents only one possible exploratory approach; by examining correlations between theoretical promoter affinity for TFs (Granek and Clarke, 2005) and relative induction or repression in individual microarray experiments, we have identified many statistically significant associations (A.L.Y., Z.X.Y., N.D.C., and T.R.H., unpublished data). In addition, because motif representations almost certainly do not fully describe in vitro TF binding preferences (e.g., see Figure 2) and because previous studies have concluded that weak and/or noncanonical binding sites are likely to be functional in some instances (Blackwell et al., 1993; Buck and Lieb, 2006; Tanay, 2006), in the future it may be useful to scan the genome with indices of relative affinity to individual sequences rather than positional models of specificity.

One aspect of global gene expression and regulation that has been difficult to model is precisely how factors within cells assemble at promoters rather than other genomic locations with similar sequence characteristics. In our study, Rsc3 emerged as a major player in NFR formation/maintenance and promoter function for hundreds of yeast genes. Our data are consistent with prior conjecture that Rsc3 uses its sequence-specific binding activity to target RSC to promoters and creating the NFR (Angus-Hill et al., 2001; Parnell et al., 2008; Wilson et al., 2006). Our data are also consistent with previous ChIP-chip analyses of RSC because promoters containing Rsc3 binding site are enriched in RSC immunoprecipitates. Rsc3 itself is frustratingly refractory to study by ChIP-chip (Parnell et al., 2008); although there is a significant enrichment of Rsc3 binding sites among ChIP-chip targets, the enrichment ratios, the overlap with Rsc3 binding sequences, and the resolution of published ChIP-chip data (Damelin et al., 2002; Ng et al., 2002; Parnell et al., 2008) are all too low to specify exact target interactions. Therefore, we cannot rule out that the effects of Rsc3 on occupancy of many promoters are indirect, although we have no other explanation for the extremely strong association between Rsc3 binding sequences and the promoter nucleosome occupancy changes in the rsc3-1 mutant (Figures 5 and 6). Several other TFs bind to sequences containing CGCG (e.g., Mbp1, Swi6, Dal82, and Rsc30), but no other known TF binding site (Harbison et al., 2004) or binding sequence (Maclsaac et al., 2006 and this study) correlates as powerfully with the rsc3-1 data as does that of our Rsc3 PWM (Spearman rank correlation \( p < 4.4 \times 10^{-43} \) between the Rsc3 PWM score and the relative change in the NFR in rsc3-1 shown in Figure 6). Moreover, motif searches in
Figure 6. Comparison of the Effects of Mutations in Essential DNA-Binding Proteins on Nucleosome Profiles at all Promoters

Within each panel, promoters are sorted by change in occupancy in the NFR. Locations of binding sequences for the mutated factor are illustrated at left in tiling intervals matching those of the array and shown as heat-maps. The change in nucleosome occupancy in the mutant is shown in the middle. Relative transcript levels are illustrated at right. The rsc3-1 panel (upper left) also shows the change in relative enrichment in Rsc8-TAP ChIP-chip between the rsc3-1 and wild-type strains.
the promoters most affected in rsc3-1 yield CGCG-containing motifs (data not shown).

Promoters in diverse organisms are enriched for both characteristic DNA structural features and binding sites for specific proteins (Lee et al., 2007). Our analyses extend these observations and, furthermore, demonstrate that many TFs contribute globally to either establishment or maintenance of the NFR (Figures 5 and 6 and Figures S3 and S4). Our data also link NFR formation to promoter function, since in all of the TF mutants we analyzed, an increase in nucleosome occupancy in the NFR generally corresponds to a decrease in transcript levels (Figure 6 and Figure S4). However, it is also true that correlation between binding sequences and effect of mutation is imperfect in all of the TF mutants we analyzed, supporting the notion that NFRs, and promoters, are created by a combination of factors, likely including both DNA structural features and specific TF recognition sites. It is curious and somewhat unexpected that the TFs that play key roles in NFR formation in yeast are not highly F conserved proteins: obvious orthologs of Reb1, Abf1, and Rsc3 are not found outside of fungi (Wilson et al., 2006). Possibly, TFs involved in promoter establishment evolve with gene architecture, chromosome structure, and nuclear organization. If this is the case, then large-scale study of TF binding specificities in other organisms may be needed as much to understand how the cell identifies genomic landmarks as to map regulatory pathways.

EXPERIMENTAL PROCEDURES

Additional details and data are found in the Supplemental Experimental Procedures (see below) and on our project web site (http://hugheslab.ccbri.utoronto.ca/supplementary-data/yeastDBD/).

Cloning and Protein Expression

We cloned PCR amplicons (pfam-defined DBDs plus 50 flanking residues) into pMAGIC (Li and Elledge, 2005). Resulting inserts were transferred into pTH1137, a T7-GST-tagged variant of pML280 (Berger et al., 2008). We obtained proteins by either purification from E. coli C41 DE3 cells (Lucigen), or in vitro transcription/translation reactions (Ambion ActivePro Kit) without purification, as indicated on our project web site.

Microarray Analysis of TF Binding Specificities

The Supplemental Experimental Procedures contain a detailed description of microarray analyses and motif derivation methods. PBM arrays and assays were as described (Berger et al., 2006). CSI methods essentially followed (Warren et al., 2006). DIP-chip was carried out as described previously (Liu et al., 2005), and the resulting DNA was hybridized to NimbleGen microarrays covering the yeast genome at 32 bp resolution.

Nucleosome and Expression Analyses Using Tiling Arrays

Extraction of nucleosomal DNA from the samples and hybridization onto the yeast tiling array was performed according to Lee et al. (2007). Isolation of total RNA and hybridization onto the tiling arrays followed (Juneau et al., 2007), except that Actinomycin D was added in a final concentration of 6 μg/ml during cDNA synthesis to prevent antisense artifacts (Perocchi et al., 2007).

ChIP-chip

We grew isogenic wild-type and rsc3-1 strains, each carrying Rsc8-TAP, in parallel under rsc3-1-restrictive growth conditions. After formaldehyde cross-linking and chromatin extraction, we performed a single pull-down with IgG sepharose. Following decrosslinking, we analyzed these samples on Nimblegen tiling arrays using a two-color procedure, comparing the pulled-down DNA to genomic DNA. We then compared relative enrichment between wild-type and rsc3-1.

Scoring Promoter Sequences and GO Enrichment

The probability of a transcription factor binding somewhere within a promoter was estimated using PWMs obtained in this study and the program GOMER (Graneck and Clarke, 2005), run with default parameters, with promoters defined as the 600 bp region 5’ to the ORF. The top 100 hits were input into FunSpec (Robinson et al., 2002).

Additional Information

Additional information including clone sequences and 8-mer scores and motifs for all TFs can be found in Tables S4–S7.

ACCESSION NUMBERS

Affymetrix tiling array data are available at ArrayExpress (record E-MEXP-1754); all other microarray data are available at GEO (record GSE12349).

SUPPLEMENTAL DATA

The Supplemental Data include eight tables, seven figures, Supplemental Experimental Procedures, and Supplemental References and can be found with this article online at http://www.cell.com/molecular-cell/supplemental/S1097-2765(08)00842-3.

ACKNOWLEDGMENTS

This work was supported by Genome Canada through the Ontario Genomics Institute, the Ontario Research Fund, a grant from the CIHR to C.N. and T.R.H. (MOP 86705), and grants from NIH (GM089420) and USDA/Hatch to A.Z.A. G.B.B. was supported by a CIHR postdoctoral fellowship, H.v.B. by the Netherlands Organization for Scientific Research (825.06.033), C.D.C. by American Heart Association Predoctoral Fellowship No. 0615615Z, and C.L.W. by Computation and Informatics in Biology and Medicine Training Grant T15LM007359. A.Z.A. is a Shaw Scholar. J.D.L. and A.J.G. are supported by NIH R01-GM072518. We thank Brenda Andrews, Charlie Boone, Li Zhijiang, Zhaolei Zhang,QUAD Morris, Larry Hiesler, Martha Bulyk, Mike Berger, Cong Zhu, and Andrew Gehrke for assistance and helpful discussions.

Received: August 13, 2008

Revised: November 5, 2008

Accepted: November 26, 2008

Published: December 24, 2008

REFERENCES


