Published in final edited form as:
Mol Cell. 2008 December 26; 32(6): 878–887. doi:10.1016/j.molcel.2008.11.020.

A new 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. Hughes**[1,2,8]

[1] Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M5S 3E1

[2] Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 3E1

[3] Department of Pharmaceutical Sciences, University of Toronto, Toronto, ON M5S 3E1

[4] Department of Biochemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706

[5] The Genome Center, University of Wisconsin-Madison, Madison, Wisconsin 53706

[6] Department of Biology and Carolina Center for Genome Sciences, CB# 3280, 408 Fordham Hall, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3280

[7] Computational and Systems Biology, Genome Institute of Singapore, 60 Biopolis St, Singapore, 138672

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 which have not been previously reported. Several newly discovered 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.

8Corresponding author: t.hughes@utoronto.ca; Tel. 416-946-8260; FAX 416-978-8258.

Supplementary material and URLs. Supplementary data files including clone sequences and 8-mer scores and motifs for all TFs are posted at http://hugheslab.ccbr.utoronto.ca/supplementary-data/yeastpbm/. Affymetrix tiling array data is available at ArrayExpress (record E-MEXP-1754); all other microarray data is available at GEO (record GSE12349).

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 encyclopaedia 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 (Chasman et al., 1990; Elemento and Tavazoie, 2005; Fourel et al., 2002; Planta et al., 1995). Abf1, Reb1, or Rap1 binding sites are found in only a minority of promoters, however (Harbison et al., 2004), highlighting the probability that additional nucleosome-displacing factors, or combinations of factors, remain to be identified.

Here, we have measured the sequence preferences of the majority of yeast TF DBDs, using a combination of systematic microarray-based approaches. These data provide a resource for genomic analyses, and for the study of the evolution of both the genome and the TFs themselves. Our data include binding preferences for 36 proteins for which there was previously no reported binding specificity information, and provide independent support for many more that were previously inferred from ChIP-chip or identified on the basis of one or a few binding sequences. Among the proteins for which we have defined specificities for the first time are Rsc3 and Rsc30. Binding sequences for these proteins occur preferentially between −125 and −75 upstream of TSS, and Rsc3 is essential for the maintenance of a nucleosome-free region in hundreds of yeast promoters as well as transcript abundance from these promoters.

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 (Supplementary Table 1). 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: (i) 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); (ii) Cognate Site Identifier (CSI) (Warren et al., 2006), in which proteins are applied to a Nimblegen array of 262,148 DNA hairpins each containing an 11bp randomized region permitting display of all possible 10-mers; and/or (iii) DNA immunoprecipitation chip (Dip-chip) (Liu et al., 2005), in which a purified transcription factor, bound to yeast genomic DNA, is immunoprecipitated *in vitro* and analyzed using microarrays.

Supplementary Table 1 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 8-mer, and expressed this as a Z-score (Berger et al., 2006). We then sought DNA sequence motifs (Position Weight Matrices or PWMs) that produced predicted binding scores (Granek and Clarke, 2005) that correlated with the 8-mer based Z-scores for each factor (see Experimental Procedures for details). The 112 resulting motifs identified are shown in Fig 1. Fig 2A illustrates how the PWM-derived scores correlate with the 8-mer Z-score data for Gzf3. Fig 2B, which shows a comparison of 8-mer Z-scores obtained for Gzf3 using either PBM or CSI, demonstrates 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 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; MacIsaac 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 (Supplementary Table 1). Sixty-three of our motifs bear an obvious correspondence to previous information (although not always all previous information), while 11 are inconsistent. The remaining 38 represent newly discovered specificities, although most of these motifs are consistent with expectations 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 (Supplementary Table 2). For example, our Fhl1 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 (Supplementary Fig 1).

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, however, 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 $P < 10^{-92}$; 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, or 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 (MacIsaac 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 (~Supplementary Fig 1).

Three of the discrepancies (Ecm22, Put3, and Ume6) are for Gal4-class proteins, which also have characteristic behaviour 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 a number of Gal4-class protein-DNA complexes, all are of dimers in the crystal. 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 intolerant of or otherwise 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).

Correspondence between amino acid sequence similarity and DNA binding specificities supports new motifs

Most of the 36 proteins we classified as having no previously established binding sequences 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 (Fig 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, Rim101) all have less than 60% identity to any other yeast protein in the DBD. ClustalW-derived phylograms similar to Fig 3 are given for all other structural classes in Supplementary Fig 2. Three major observations include: (i) 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 (Supplementary Fig 3). 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. (ii) 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 (Dequard-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). (iii) 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 in shared functional classes. Twenty-seven of the 112 motifs had a hypergeometric P-value of < 0.000005 (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” ($P < 2.2 \times 10^{-14}$) and Pdr1 (Pleiotropic Drug Resistance), with “response to drug” ($P < 1 \times 10^{-6}$). 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” ($P < 6.1 \times 10^{-10}$). Ypr196w and Ydr520c binding sequences were also enriched in the promoters of hexose transporters ($P < 2.4 \times 10^{-8}$; 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 that 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 ($P < 3.8 \times 10^{-6}$ 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. Fig 4A 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. Fig 4B shows 14 motifs we classified as new and

unexpected; several of these are also located preferentially in the NFR. The most striking instances 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 remodelling chromatin in the vicinity of their binding sites. At a more liberal PWM score threshold, 708 yeast genes contain a potential Rsc3 binding sequence in the NFR region (−130 to −75), compared to only 146 found in an identical amount of ORF sequence. These 708 genes represent a broad spectrum of functional classes, including 169 (of 1101) that are essential for cell viability (hypergeometric $P < 2.2 \times 10^{-6}$). Given that RSC is an abundant protein complex that repositions nucleosomes (Angus-Hill et al., 2001;Cairns et al., 1996;Parnell et al., 2008), we reasoned that Rsc3 and Rsc30 may play a broad role in directing the establishment or maintenance of nucleosome-free regions in promoters. We focused on Rsc3 because it is essential, and therefore its activity is required under typical laboratory growth conditions.

Promoters containing Rsc3 binding sequences are likely to be bound by RSC

Three previous studies have analyzed RSC binding sites in the yeast genome using ChIP-chip (Damelin et al., 2002; Ng et al., 2002; Parnell et al., 2008), two involving Rsc3. Promoters containing the Rsc3 motif displayed a statistically significant correspondence to overall RSC occupancy in these previous studies: among 5,015 (4,947 with ChIP-chip data) yeast genes with well-defined TSS (Lee et al., 2007), 2,325 (2,296 with ChIP-chip data) have a match to our Rsc3 motif (using our liberal cutoff). 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 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 vs. total genomic DNA) in the *rsc3-1* mutant to that in an isogenic wildtype control grown at the same temperature (37 degrees, for 6 hours). Fig 5A shows an example locus in which nucleosome depletion over a Rsc3 binding sequence in a promoter region is dependent on *RSC3*. Fig 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 (Fig 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 subunit (Parnell et al., 2008), we saw little or no effect on nucleosome positioning or occupancy at tRNA genes (Supplementary Fig 4), 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 (Fig 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 wildtype 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 wildtype strain, Rsc8 occupancy ratios are also highest at tRNAs (maximum enrichment 8.5-fold in our analysis, Supplementary Fig 4), and at Pol II promoters there is a significant correspondence between Rsc8 occupancy and the Rsc3 motif score (Spearman rank correlation $P < 1.3 \times 10^{-9}$). Furthermore, occupancy at tRNAs is not affected by *rsc3-1* (Supplementary Fig 4), 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 (Fig 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* (Fig 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 (Fig 6 and Supplementary Fig 5). 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 (Fig 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 (Supplementary Figs 5 and 6). By way of comparison, there is no relationship between binding sequences for Cep3, a centromere-binding protein, and nucleosome occupancy at Pol II promoters (Fig 6 and Supplementary Fig 5). There is, however, a perfect 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 (Supplementary Fig 7; signal from centromere probes could reflect occupancy by centrosomes).

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% (Supplementary Table 1). This expanded index of sequence preferences provides a new 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 hundreds of statistically significant associations (unpublished data). In addition, because motif representations almost certainly do not fully describe *in vitro* TF binding preferences (e.g. see Fig 2), and because previous studies have concluded that weak and/or non-canonical 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 (Figs 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 ((MacIsaac 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 Fig 6). Moreover, motif searches in the promoters most affected in *rsc3-1* yield CGCG-containing sequences (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 TFs contribute to either establishment or maintenance of the NFR (Figs 5, 6, and Supplementary Figs 3 and 4). 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 (Fig 6 and Supplementary Fig 4). Correlation between binding sequence and effect of the mutation is, however, imperfect in all cases, 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-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 Supplementary Methods and on our project web site (see below).

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 Supplementary methods 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 32bp 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 artefacts.

ChIP-chip

We grew isogenic wildtype and *rsc3-1* strains, each carrying Rsc8-TAP, in parallel under *rsc3-1* restrictive growth conditions. After formaldehyde crosslinking and chromatin extraction we performed a single pulldown 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 wildtype 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 (Granek and Clarke, 2005), run with default parameters, with promoters defined as the 600bp region 5′ to the ORF. The top 100 hits were input into FunSpec (Robinson et al., 2002).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by Genome Canada through the Ontario Genomics Institute, the Ontario Research Fund, a grant from the CIHR to CN and TRH (MOP 86705) and grants from NIH (GM069420) and USDA/Hatch to AZA. GBB was supported by a CIHR postdoctoral fellowship, HvB by the Netherlands Organization for Scientific Research (825.06.033), CDC by American Heart Association Predoctoral Fellowship No. 0615615Z, CLW by Computation

and Informatics in Biology and Medicine Training Grant T15LM007359. AZA is a Shaw Scholar. JDL and AJG are supported by NIH R01-GM072518. We thank Brenda Andrews, Charlie Boone, Li Zhijang, Zhaolei Zhang, Quaid Morris, Larry Hiesler, Martha Bulyk, Mike Berger, and Andrew Gehrke for assistance and helpful discussions.

References

Angus-Hill ML, Schlichter A, Roberts D, Erdjument-Bromage H, Tempst P, Cairns BR. A Rsc3/Rsc30 zinc cluster dimer reveals novel roles for the chromatin remodeler RSC in gene expression and cell cycle control. Mol Cell 2001;7:741–751. [PubMed: 11336698]

Barrera LO, Ren B. The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding. Curr Opin Cell Biol 2006;18:291–298. [PubMed: 16647254]

Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell 2004;117:185–198. [PubMed: 15084257]

Benos PV, Bulyk ML, Stormo GD. Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res 2002;30:4442–4451. [PubMed: 12384591]

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 2008;133:1266–1276. [PubMed: 18585359]

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 2006;24:1429–1435. [PubMed: 16998473]

Blackwell TK, Huang J, Ma A, Kretzner L, Alt FW, Eisenman RN, Weintraub H. Binding of myc proteins to canonical and noncanonical DNA sequences. Mol Cell Biol 1993;13:5216–5224. [PubMed: 8395000]

Buck MJ, Lieb JD. A chromatin-mediated mechanism for specification of conditional transcription factor targets. Nat Genet 2006;38:1446–1451. [PubMed: 17099712]

Cairns BR, Lorch Y, Li Y, Zhang M, Lacomis L, Erdjument-Bromage H, Tempst P, Du J, Laurent B, Kornberg RD. RSC, an essential, abundant chromatin-remodeling complex. Cell 1996;87:1249–1260. [PubMed: 8980231]

Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR, et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell 2005;122:33–43. [PubMed: 16009131]

Chasman DI, Lue NF, Buchman AR, LaPointe JW, Lorch Y, Kornberg RD. A yeast protein that influences the chromatin structure of UASG and functions as a powerful auxiliary gene activator. Genes Dev 1990;4:503–514. [PubMed: 2361590]

Damelin M, Simon I, Moy TI, Wilson B, Komili S, Tempst P, Roth FP, Young RA, Cairns BR, Silver PA. The genome-wide localization of Rsc9, a component of the RSC chromatin-remodeling complex, changes in response to stress. Mol Cell 2002;9:563–573. [PubMed: 11931764]

Dequard-Chablat M, Riva M, Carles C, Sentenac A. RPC19, the gene for a subunit common to yeast RNA polymerases A (I) and C (III). J Biol Chem 1991;266:15300–15307. [PubMed: 1869554]

Elemento O, Tavazoie S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol 2005;6:R18. [PubMed: 15693947]

Fourel G, Miyake T, Defossez PA, Li R, Gilson E. General regulatory factors (GRFs) as genome partitioners. J Biol Chem 2002;277:41736–41743. [PubMed: 12200417]

Granek JA, Clarke ND. Explicit equilibrium modeling of transcription-factor binding and gene regulation. Genome Biol 2005;6:R87. [PubMed: 16207358]

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. Transcriptional regulatory code of a eukaryotic genome. Nature 2004;431:99–104. [PubMed: 15343339]

Juneau K, Palm C, Miranda M, Davis RW. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. Proc Natl Acad Sci U S A 2007;104:1522–1527. [PubMed: 17244705]

Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 2004;428:617–624. [PubMed: 15004568]

Kim JH, Polish J, Johnston M. Specificity and regulation of DNA binding by the yeast glucose transporter gene repressor Rgt1. Mol Cell Biol 2003;23:5208–5216. [PubMed: 12861007]

Larochelle M, Drouin S, Robert F, Turcotte B. Oxidative stress-activated zinc cluster protein Stb5 has dual activator/repressor functions required for pentose phosphate pathway regulation and NADPH production. Mol Cell Biol 2006;26:6690–6701. [PubMed: 16914749]

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet 2007;39:1235–1244. [PubMed: 17873876]

Li MZ, Elledge SJ. MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules. Nat Genet 2005;37:311–319. [PubMed: 15731760]

Liu X, Lee CK, Granek JA, Clarke ND, Lieb JD. Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. Genome Res 2006;16:1517–1528. [PubMed: 17053089]

Liu X, Noll DM, Lieb JD, Clarke ND. DIP-chip: rapid and accurate determination of DNA-binding specificity. Genome Res 2005;15:421–427. [PubMed: 15710749]

Lutfiyya LL, Iyer VR, DeRisi J, DeVit MJ, Brown PO, Johnston M. Characterization of three related glucose repressors and genes they regulate in Saccharomyces cerevisiae. Genetics 1998;150:1377–1391. [PubMed: 9832517]

MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 2006;7:113. [PubMed: 16522208]

MacPherson S, Larochelle M, Turcotte B. A fungal family of transcriptional regulators: the zinc cluster proteins. Microbiol Mol Biol Rev 2006;70:583–604. [PubMed: 16959962]

Mintseris J, Eisen MB. Design of a combinatorial DNA microarray for protein-DNA interaction studies. BMC Bioinformatics 2006;7:429. [PubMed: 17018151]

Nash R, Weng S, Hitz B, Balakrishnan R, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, et al. Expanded protein information at SGD: new pages and proteome browser. Nucleic Acids Res 2007;35:D468–471. [PubMed: 17142221]

Ng HH, Robert F, Young RA, Struhl K. Genome-wide location and regulated recruitment of the RSC nucleosome-remodeling complex. Genes Dev 2002;16:806–819. [PubMed: 11937489]

Parnell TJ, Huff JT, Cairns BR. RSC regulates nucleosome positioning at Pol II genes and density at Pol III genes. Embo J 2008;27:100–110. [PubMed: 18059476]

Pierce M, Benjamin KR, Montano SP, Georgiadis MM, Winter E, Vershon AK. Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression. Mol Cell Biol 2003;23:4814–4825. [PubMed: 12832469]

Planta RJ, Goncalves PM, Mager WH. Global regulators of ribosome biosynthesis in yeast. Biochem Cell Biol 1995;73:825–834. [PubMed: 8721998]

Roberts DN, Stewart AJ, Huff JT, Cairns BR. The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships. Proc Natl Acad Sci U S A 2003;100:14695–14700. [PubMed: 14634212]

Robinson MD, Grigull J, Mohammad N, Hughes TR. FunSpec: a web-based cluster interpreter for yeast. BMC Bioinformatics 2002;3:35. [PubMed: 12431279]

Schlake T, Schorpp M, Nehls M, Boehm T. The nude gene encodes a sequence-specific DNA binding protein with homologs in organisms that lack an anticipatory immune system. Proc Natl Acad Sci U S A 1997;94:3842–3847. [PubMed: 9108066]

Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J. A genomic code for nucleosome positioning. Nature 2006;442:772–778. [PubMed: 16862119]

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 2008;451:535–540. [PubMed: 18172436]

Soragni E, Kassavetis GA. Absolute gene occupancies by RNA polymerase III, TFIIIB and TFIIIC in Saccharomyces cerevisiae. J Biol Chem. 2008

Tanay A. Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res 2006;16:962–972. [PubMed: 16809671]

Vik A, Rine J. Upc2p and Ecm22p, dual regulators of sterol biosynthesis in Saccharomyces cerevisiae. Mol Cell Biol 2001;21:6395–6405. [PubMed: 11533229]

Warren CL, Kratochvil NC, Hauschild KE, Foister S, Brezinski ML, Dervan PB, Phillips GN Jr, Ansari AZ. Defining the sequence-recognition profile of DNA-binding molecules. Proc Natl Acad Sci U S A 2006;103:867–872. [PubMed: 16418267]

Wilson B, Erdjument-Bromage H, Tempst P, Cairns BR. The RSC chromatin remodeling complex bears an essential fungal-specific protein module with broad functional roles. Genetics 2006;172:795–809. [PubMed: 16204215]

Wolfe SA, Nekludova L, Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct 2000;29:183–212. [PubMed: 10940247]

Yuan GC, Liu JS. Genomic sequence is highly predictive of local nucleosome depletion. PLoS Comput Biol 2008;4:e13. [PubMed: 18225943]

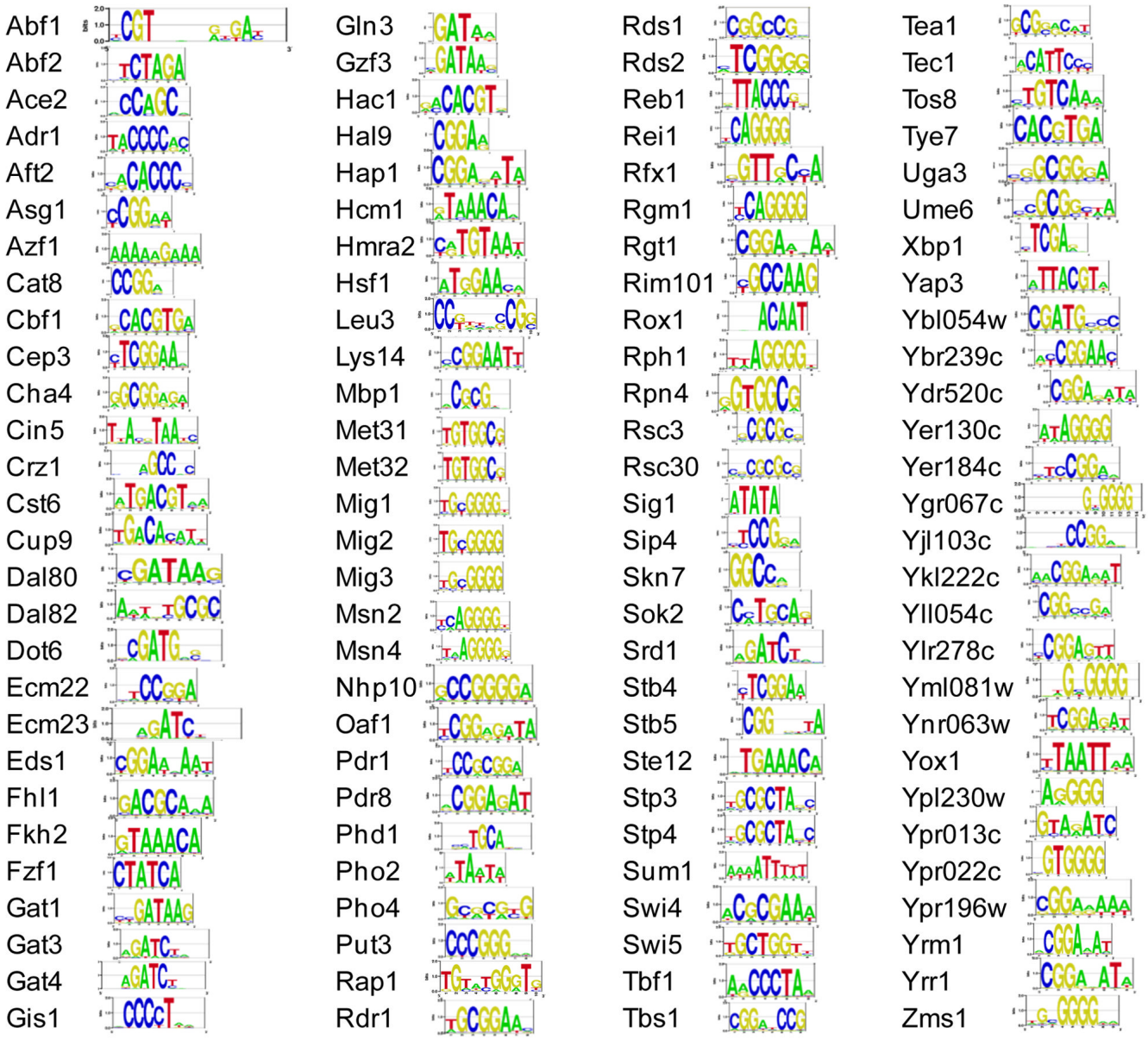

Fig 1.
Motifs identified in our study.

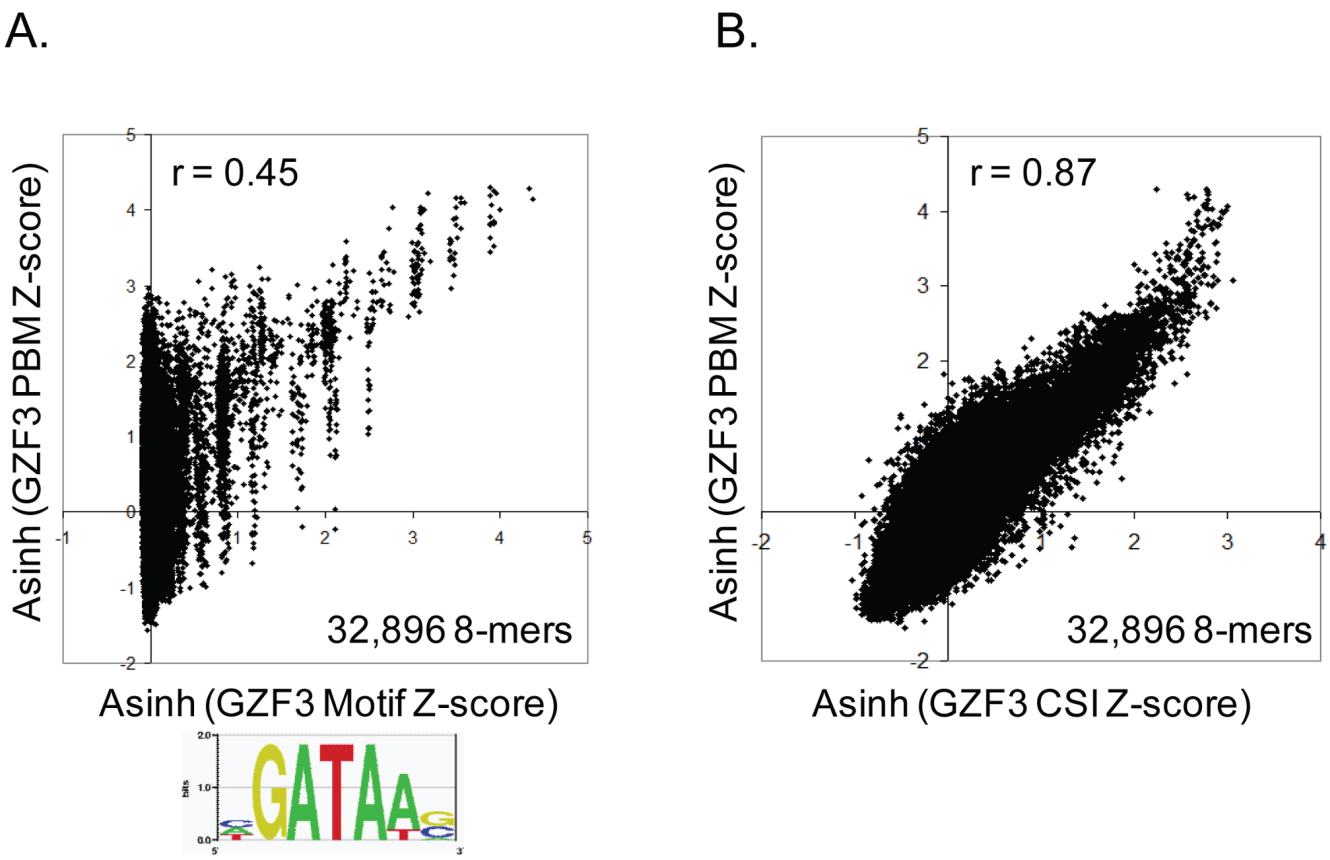

Fig 2. Comparison of motif representation and reproducibility of 8-mer profiles across platforms

(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. (B) CSI Z-scores (combined from up to four array spots containing the 8-mer) vs. Z-scores from 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).

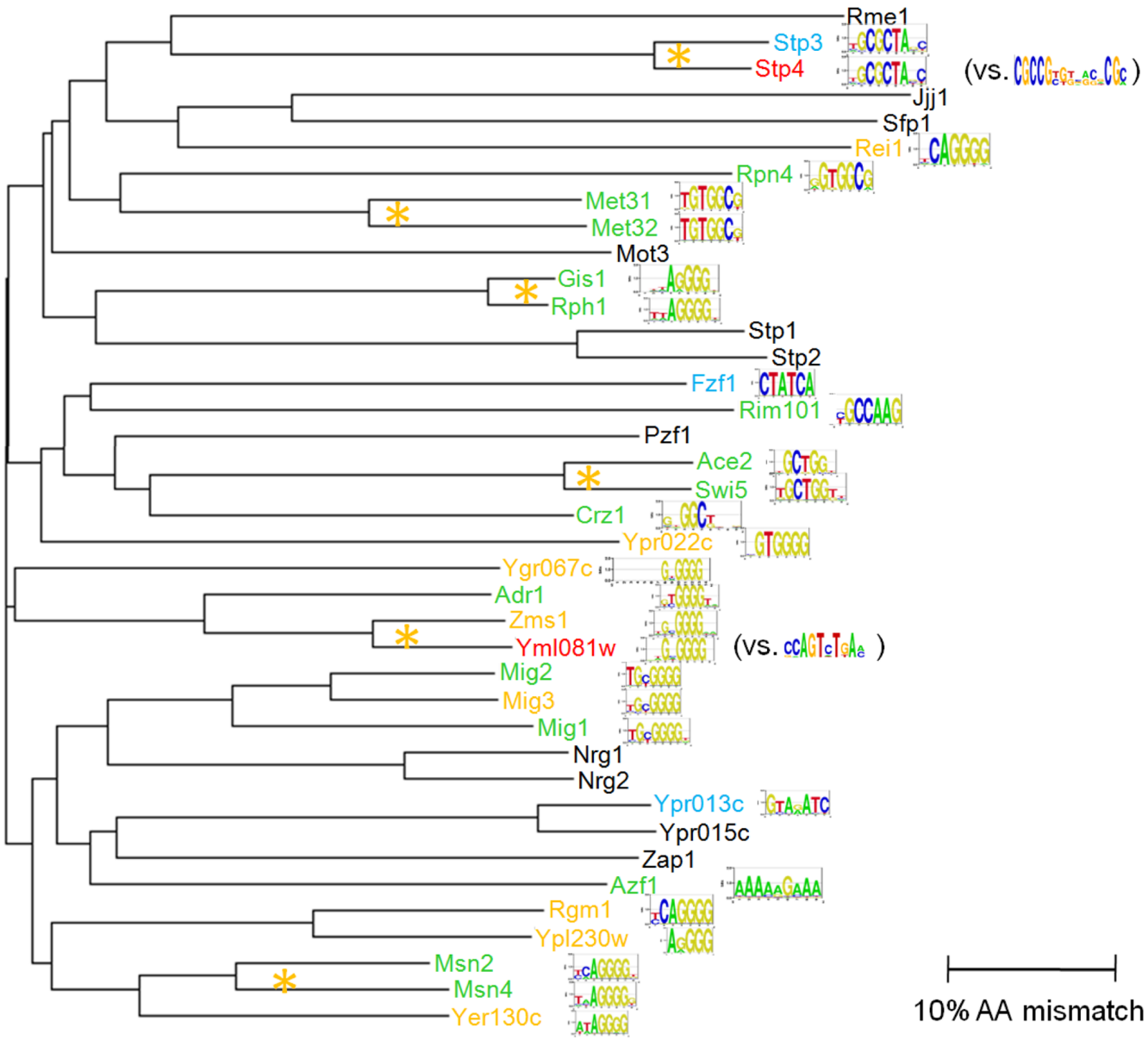

Fig 3. Similarity among C2H2 zinc finger motifs reflects DNA-binding domain sequence similarity

The phylogram tree was created using online EBI ClustalW with default settings. Our motifs are shown next to the gene names; inconsistent motifs from (MacIsaac et al., 2006) are shown for Stp4 and Yml081w. Yellow asterisks represent pairs arising from the WGD. Colors of protein names reflect our classifications of consistency with prior data: Green; known motif obtained; Red; discrepancy between our motif and that previously reported; Yellow, new motif but consistent with expectation based on homology; Blue, new unexpected motif.

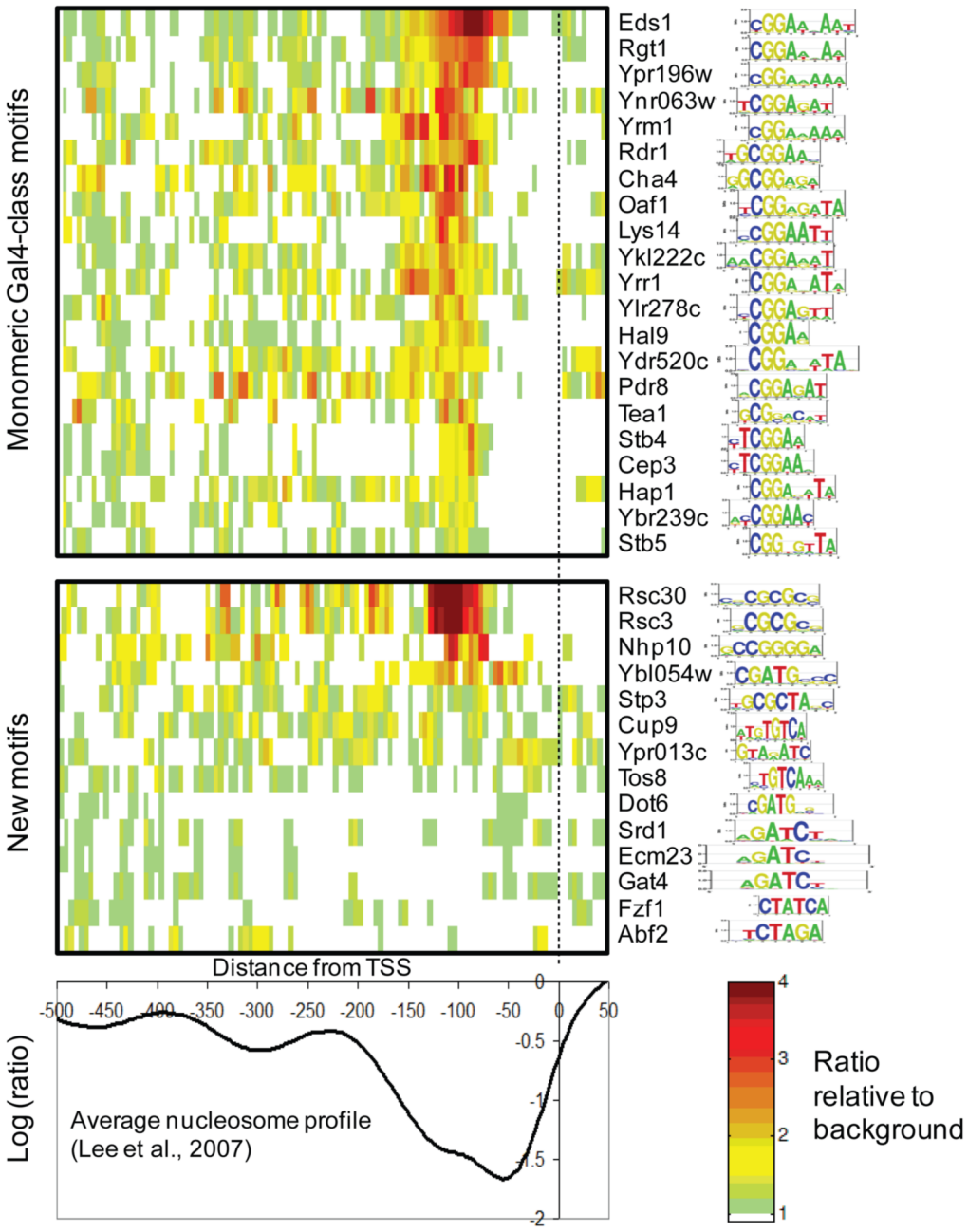

Fig 4. Bias in the position of TF binding sequences in 5,015 promoters with well-defined TSS (Lee et al., 2007)

Motif scores (Granek and Clarke, 2005) were calculated for 8-bp windows, and high-scoring 8-mers were tallied along equivalent positions of all of the yeast promoter sequences using a cutoff selected to capture only the linear range of Z vs. PWM score in PBM experiments (cutoff values are given in Supplementary data). Background was calculated from the first 100 bases of yeast ORFs. TFs are sorted by relative enrichment between −125 and −75.

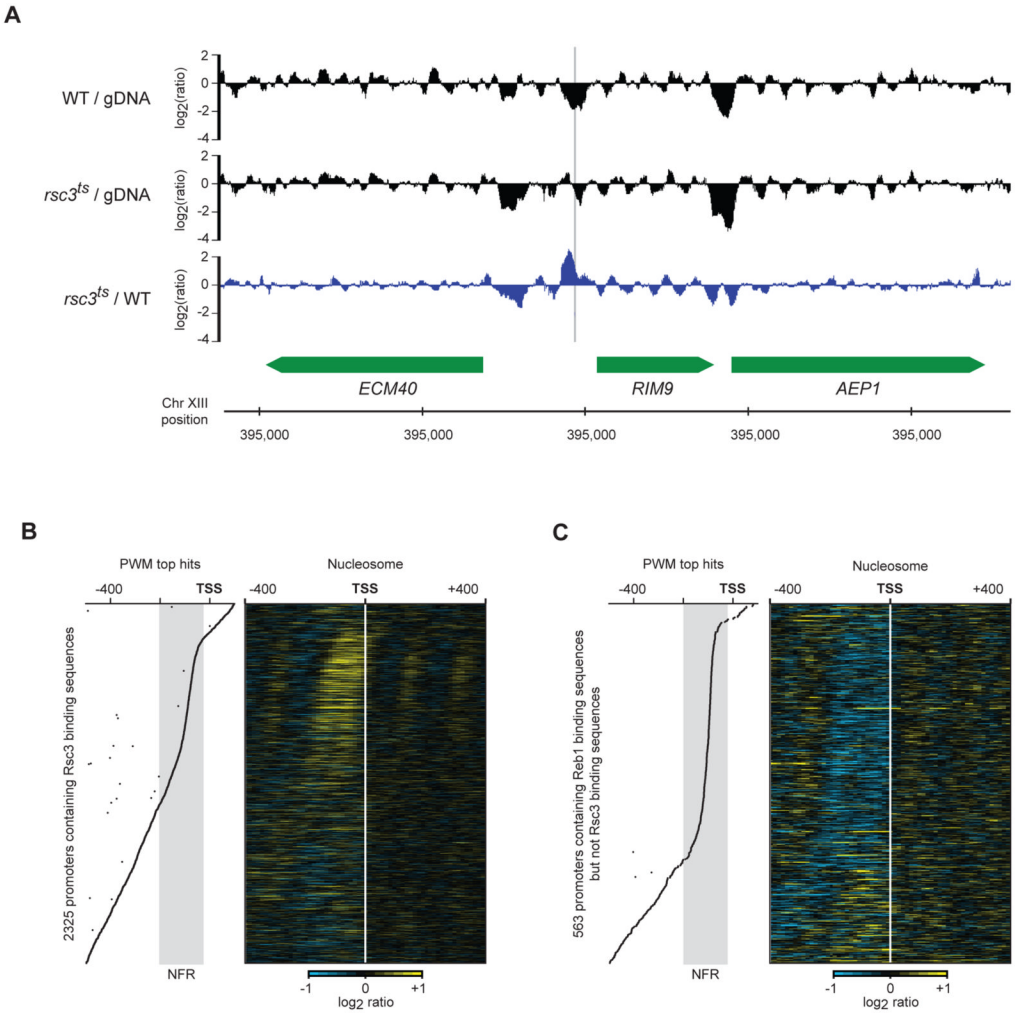

Fig 5. Rsc3 influences nucleosome occupancy at proximal promoters containing Rsc3 binding sites

(A) A segment of Chromosome XIII with a Rsc3 binding sequence (grey vertical line) that is depleted in wildtype but occupied in the *rsc3-1* mutant. (B, C) Changes in promoter nucleosome occupancy profiles between *rsc3-1* and a wildtype control for promoters containing Rsc3 binding sequences (C), or containing Reb1 binding sequences but not Rsc3 binding sequences (D). 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 panels B and C. Additional sites of equivalent PWM score are also indicated.

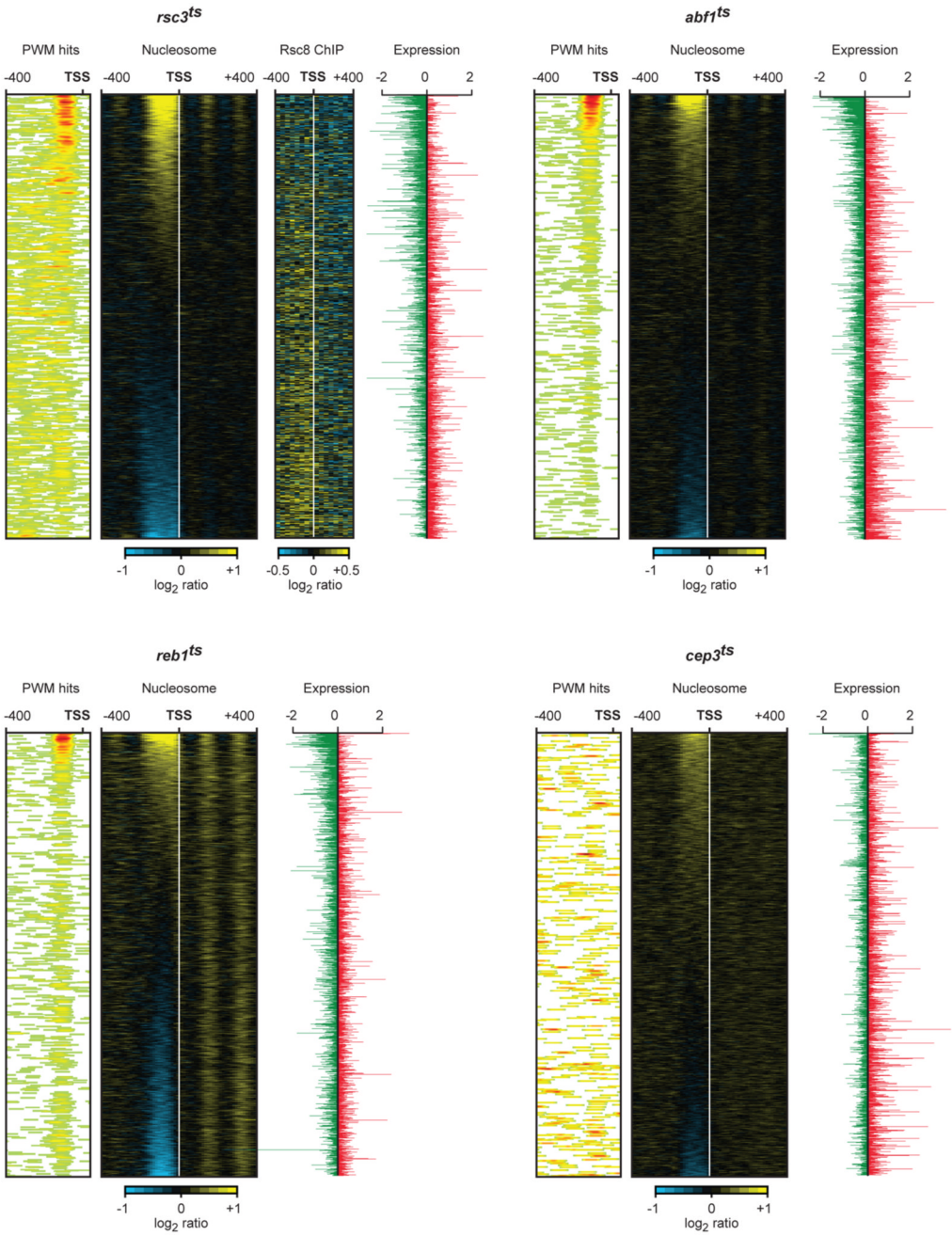

Fig 6. Comparison of the effects of mutations in essential DNA-binding proteins on nucleosome profiles at all promoters

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. 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 wildtype strains.