Supplementary Methods for

"High-Resolution DNA Binding Specificity Analysis of Yeast Transcription Factors"

Cong Zhu, Kelsey Byers, Rachel McCord, Zhenwei Shi, Michael Berger, Daniel Newburger, Katrina Saulrieta, Zachary Smith, Mita Shah, Mathangi Radhakrishnan, Anthony Philippakis, Yanhui Hu, Federico De Masi, Marcin Pacek, Andreas Rolfs, TVS Murthy, Joshua LaBaer, and Martha L Bulyk

Table of contents:

A. Cloning of S. cerevisiae TFs as full-length and DNA-binding domain constructs

We compiled a list of 303 known and predicted yeast TFs. We considered only those genes present in the sequenced strain S288C (Goffeau et al. 1996), which is a MATalpha, not MATa strain. These 303 ORFs were identified as candidate TFs based on at least one of the following three criteria: (1) annotated in Gene Ontology (Ashburner et al. 2000; Consortium 2001; Harris et al. 2004) as "DNA-binding" and one of "Transcription Factor", "Transcriptional Activator", or "Transcriptional Repressor" in the Yeast Proteome Database (YPD) (Costanzo et al. 2000); (2) annotated as "Transcription Factor" in the Munich Information for Protein Sequences (MIPS) Database (Mewes et al. 2002); (3) selected for ChIP-chip in a prior global study (Harbison et al. 2004). TFs classified as being of "unknown function" include 21 "uncharacterized ORFs" plus an additional 8 that are "verified ORFs" yet are annotated in SGD as "protein of unknown function". This list is accurate as of January 2008 annotations in YPD and MIPS. All ORFs not classified as "unknown function" were classified as "characterized". These include all "verified ORFs" that had some described function in SGD other than "protein of unknown function", and so here "characterized" is a quite loose definition. Note that 19 of our proteins were 'long shots', in that their annotated domains had no evidence for sequence-specific DNA binding; the domains of such proteins include: PHD (9), Zf_CCCH (2), bromodomain (4), SIR2 (2) and SNF2_N (2).

In addition, we also added in the following ORFs for the following reasons: (1) *SPT15*, the TATA-binding protein (component of TFIID) in *S. cerevisiae*; (2) *TBF1* (TTAGGG repeat-Binding Factor) (Brigati et al. 1993; Liu and Tye 1991), which is a known sequence-specific DNA binding protein; and (3) two homologous genes, RSC3 and RSC30, which are paralogous genes that contain Zn_2Cys_6 DNA binding domains and encode components of the RSC chromatin remodeling complex. Rsc3 and RSC30 were of interest to us because they have been proposed to bind DNA and recruit RSC to specific loci (Angus-Hill et al. 2001).

We cloned these 245 full-length ORFs and 99 DBDs into Gateway-compatible Entry and Destination vectors, pDONR201 or pDONR221, and pDEST-GST (Braun et al. 2002), as described previously (Hu et al. 2007). A separate, partially redundant set of 118 DBDs were a

generous gift from Tim Hughes of U. Toronto and Jason Lieb of U. North Carolina. Thus, our final clone collection includes 245 ORFs as full-length constructs, and 208 as DBDs. We were not able to get either FL or DBD clones for 22 of the ORFs in our list. The missing full-length ORFs have been for the most part lengthy or otherwise difficult to clone. With regard to the DBD collection, a number of the ORFs do not have identifiable DNA binding domains. We did not pursue cloning of DBDs if the Pfam-annotated DBDs spanned >80% of the full-length protein, or in some cases, if careful manual inspection indicated that these were not likely DBDs based on their descriptions (e.g. bromodomain).

B. High-Throughput Expression and Purification of GST-Tagged Yeast TFs

For all 245 full-length TF and 208 DBD clones, we performed high-throughput over-expression in *E. coli* cultures and subsequent affinity purification using glutathione resin in 96-well plates essentially as described previously (Hu et al. 2007). In this study, the high-throughput over-expression in 1.2-ml 96-well plates was done robotically, while the high-throughput purification was done in 96-well plates using a multichannel pipettor.

C. PBMs on TFs Purified in High-Throughput Manner from E. coli

Our PBMs on yeast TFs overexpressed in and purified from *E. coli* in high-throughput indicated that these proteins can yield DNA binding site motifs (Hu et al. 2007). In this present study, all the proteins were overexpressed robotically but manually purified in 96-well plates.

D. Western blots

We expressed and purified in high-throughput 245 TFs as full-length TFs fused to GST. For each purified protein we performed Western blots to assess quality and to approximate its

concentration, as described previously (Hu et al. 2007). Briefly, proteins were analyzed on precast 4%–12% XT Criterion gradient gels (BioRad) according to the manufacturer's protocols. Immunoblots were probed with rabbit anti-GST antibody (Sigma) at 20 ng/ml final concentration and developed using SuperSignal West Femto Maximum Sensitivity Substrate (Pierce) according to the manufacturer's protocols. Western blot analysis indicated that 57% (140/245) of these expression cultures yielded at least 10 pmol of the desired protein, which is sufficient for one PBM experiment in 8x15K array format plus duplicate PBM experiments in 4x44K array format at a minimum concentration of 25 nM TF in the PBM experiment, from single 1.2-ml wells of high-throughput (96-well) *E. coli* cultures. 208 DBD GST-fusion clones underwent high-throughput expression and purification at HIP, of which 185 DBDs (89%) were acceptable by Western blots. Examples of Westerns of purified full-length TFs are shown in Fig. 3 of our previous *Genome Research* paper (Hu et al. 2007). Overall this resulted in 246 TFs for which we have acceptable protein as either full-length or DBD or both. We were not able to get sufficient amounts of acceptable protein, as judged by both size and concentration, as either the full-length or DBD version, for the remaining 43 constructs, corresponding to 35 nonredundant TFs.

E. Protein Binding Microarrays (PBMs)

Briefly, our maximally compact, synthetic DNA sequence design (Berger et al. 2006; Philippakis et al. 2008) for PBMs represents all possible DNA sequence variants of a given length *k* (i.e., all "*k*-mers") on a single microarray. We constructed microarrays covering all 10 bp binding sites by converting high-density single-stranded oligonucleotide arrays to double-stranded DNA arrays (Berger et al. 2006). Importantly, our universal arrays are designed to evenly sample the space of higher order *k*-mers (Berger et al. 2006; Philippakis et al. 2008), which allows us to generate longer motifs (Berger et al. 2006). Moreover, our universal array designs employed in this present study were designed to cover various *k*-mers consisting of half-sites with long spacers (i.e., "Gal4-type" motifs).

To design the sequence for our 'all 10-mer' universal protein binding microarrays (PBMs), we utilized de Bruijn sequences of order 10 (Berger et al. 2008; Berger et al. 2006; Philippakis et al.

2008). A de Bruijn sequence of order k is a circular string of length 4^k that contains every k-mer exactly once when overlaps are considered. We created two separate designs for replicate experiments, which we optimized to achieve maximal coverage of gapped k-mers (described below). To generate de Bruijn sequences of order 10 for our universal PBMs, we used a linear-feedback shift register corresponding to the primitive polynomial:

$$3x^{10}+3x^9+2x^8+1x^7+2x^6+2x^5+3x^4+3x^3+1x^2+2x$$

For longer motifs containing nine or more positions with an information content (Schneider et al. 1986) of 0.3 or greater in the PWM constructed from the PBM data using Seed-and-Wobble (Berger and Bulyk; Berger et al. 2006), influences of flanking sequence, position, and orientation were minimized by performing a replicate PBM experiment on a separate microarray containing a second de Bruijn sequence (Berger et al. 2006; Philippakis et al. 2008). The two de Bruijn sequences for our two PBMs differ by cyclic permutations of A, C, G, and T. We empirically selected these particular de Bruijn sequences because they uniformly cover all contiguous 10-mers and all gapped 10-mers spanning 11 total positions. Further, they exhibit optimal coverage of contiguous and gapped 8-mers. Any 8-mer is guaranteed to occur 16 times in a de Bruijn sequence of order 10 (32 times for non-palindromes). Our de Bruijn sequences exhibit this redundancy for all gapped 8-mers spanning up to 12 total positions (except for sequence variants of the single pattern 1111-1-1--11), as well as all gapped 8-mers of the pattern 1111-gap-1111 with a gap of up to 20 positions. Thus, all 4^8 sequence variants for each of these 357 patterns (nearly 23.4 million 8-mers) occur at least 16 times each.

After generating these de Bruijn sequences *in silico*, we partitioned them into subsequences of length 36 nucleotides (nt) and overlapping by 11 nt, resulting in 41,944 36-mers for each microarray. Any 36-mer with a run of five or more consecutive guanines was replaced by its reverse complement to avoid problems in primer extension for double-stranding. We appended a common 24-nt sequence to each 3' end complementary to our primer in order to create 60-mer sequences that would become the probes on our custom-designed microarrays. These microarrays were synthesized by Agilent technologies in their "4x44K" format. Each slide contains all possible 10-mers in approximately 44,000 probes, repeated in four identical subgrids that can be physically separated into four chambers for four separate experiments. The

additional probes beyond this set of 41,944 were designated as control sequences for a variety of purposes.

In order to efficiently test all proteins in our collection, we first performed an initial survey that utilized eight-chambered arrays in Agilent's "8x15K" format. We created a novel microarray design for these experiments consisting of one de Bruijn sequence of order 9 and one de Bruijn sequence of order 8 in approximately 15,000 total spots. Consequently, each subgrid contains every 9-mer at least twice (when reverse complements are considered), every non-palindromic 8-mer at least 10 times, and every 7-mer at least 40 times. As before, we specifically chose de Bruijn sequences that enabled us to efficiently capture gapped *k*-mers, including the "Gal4-type" motifs consisting of half-sites with long spacers.

To generate the de Bruijn sequence of order 9, we used a linear-feedback shift register corresponding to the primitive polynomial:

$$3x^9+1x^8+1x^7+1x^6+3x^5+1x^4+2x^3+3x^2+2x$$

We selected this de Bruijn sequence because it uniformly covers all contiguous 9-mers and a large fraction of all 9-mers spanning up to 12 total positions. Additionally, it provides equal coverage of all 7-mers spanning up to 12 total positions with a single exception (corresponding to the pattern 1-111---111). We partitioned this de Bruijn sequence into 36-mer subsequences overlapping by 12 nucleotides to form the spots on the microarray.

To generate the de Bruijn sequence of order 8, we used a linear-feedback shift register corresponding to the primitive polynomial:

$$3x^8+1x^7+1x^6+1x^5+1x^4+2x^3+3x^2+2x$$

This de Bruijn sequence uniformly covers all contiguous 8-mers, all 8-mers spanning 9 total positions, and all gapped 8-mers of the pattern 1111-gap-1111 with gaps up to 10 positions. It also provides equal coverage of all 7-mers spanning up to 11 total positions with four exceptions (all separate from the pattern missed above). We partitioned this de Bruijn sequence into 36-mer subsequences overlapping by 15 nucleotides to form the spots on the microarray.

All GST-TF fusion proteins that were deemed acceptable by Western blot (i.e., sufficient amount

of protein of the appropriate size) were first tested on new or stripped and re-used 8x15K arrays, allowing us to test TFs both rapidly and at lower cost. If a TF was represented by both a full-length and also a DBD construct, we tested both in separate PBMs. If a protein resulted in a candidate motif whose seed *k*-mer either had an enrichment score of at least 0.45 or was a non-G-run *k*-mer with an enrichment score of at least 0.35 from the 8x15K test array PBM data, we subsequently examined it in the 4x44K array format. For 7 proteins, the observed PBM binding profiles were clearly due to contamination from an adjacent well on the 96-well protein prep plate, and so those data were eliminated from further analysis.

F. Identification of DNA binding site motifs using Seed-and-Wobble

PBM k-mer scores:

As described previously (Berger and Bulyk 2006), every non-palindromic 8-mer occurs on at least 32 spots in each chamber of our universal PBM. Because of this redundancy, we are able to provide a robust estimate of the relative preference of a transcription factor for every contiguous and gapped 8-mer that is covered on our array. Here, for each 8-mer, we provide the median normalized signal intensity and a rank-based statistical enrichment score. Median normalized signal intensity refers to the median normalized signal intensity for the set of probes containing a match to each 8-mer (usually ~32 probes, but some might be flagged occasionally because of dust flecks, etc., and therefore removed from further consideration). We have shown previously that higher PBM median signal intensity corresponds to stronger protein-DNA binding affinity (Berger and Bulyk 2006); however, experimental variability and differences in absolute signal intensities and nonspecific binding can make this measure difficult to compare for different TFs.

Our enrichment score is a rank-based, non-parametric statistical measure that is invariant to protein concentration and readily allows different experiments to be compared on the same scale. This enrichment score has been described previously in detail (Berger and Bulyk 2006). Briefly, for each 8-mer (contiguous or gapped) we consider the collection of all probes containing a match as the "foreground" feature set and the remaining probes as the "background" feature set.

We compare the ranks of the top half of the foreground with the ranks of the top half of the background by computing a modified form of the Wilcoxon-Mann-Whitney (WMW) statistic scaled to be invariant of foreground and background sample sizes. The enrichment score ranges from +0.5 (most favored) to -0.5 (most disfavored). As described previously (Berger and Bulyk 2006), in order to combine the data from duplicate experiments for a TF using microarrays created with independent sequence designs, we computed enrichment scores for all 8-mers for each separate experiment and then calculated the mean enrichment score for each 8-mer directly.

Motif construction using Seed-and-Wobble:

In addition to reporting scores for each individual 8-mer for each TF, we compactly represent these binding data as position weight matrices (PWMs). Our "Seed-and-Wobble" algorithm for deriving motifs from universal array PBM data has been described previously (Berger and Bulyk; Berger et al. 2006). Briefly, the algorithm works in two stages. In the first stage (the "Seed" stage), we identify the single 8-mer (ungapped or gapped) with the greatest enrichment score. For this study, we considered all 8-mers spanning up to 10 total positions, as well as all 8-mers of the pattern 1111-gap-1111 with gaps up to 12 positions, as candidate seeds. In the second stage (the "Wobble" stage), we systematically test the relative preference of each nucleotide variant at each position, both within and outside the seed. This is accomplished by examining each of the four nucleotides at each position within the 8-mer seed (keeping the other 7 positions fixed) and computing the modified WMW statistic using the entire set of probes containing one of the four variants. For positions outside the 8-mer seed, we first identify the single position within the seed with the lowest information content, treat it as a gapped position, and query every other position for which the resulting 8-mer is covered in our de Bruijn sequence (i.e., all 4^8 sequence variants of that pattern exhibit 32-fold redundancy). Finally, we transform the motif derived from this method into a PWM using a Boltzmann distribution (Workman et al. 2005). In order to derive a single motif combined from separate experiments, we choose the 8-mer with the greatest average enrichment score as a seed, use it to build a PWM on each separate array, and average the matrix elements, as described previously (Berger et al. 2006).

G. Comparison of PBM motifs

We used CompareACE to compare our 89 PBM-derived motifs against a list of 4,282 PWMs for previously published motifs, including both known TF binding site motifs and purely *in silico* derived candidate regulatory motifs, that we compiled from the literature (Beer and Tavazoie 2004; Cliften et al. 2003; Harbison et al. 2004; Hughes et al. 2000; Kellis et al. 2003; Lee et al. 2002; MacIsaac et al. 2006; Morozov and Siggia 2007; Mukherjee et al. 2004; Narlikar et al. 2006; Sudarsanam et al. 2002; Tavazoie et al. 1999; Workman et al. 2006). We required a minimum CompareACE motif similarity score of 0.7 to consider motifs as matching.

H. Scoring of potential target genes

The PBM *k*-mer data were used to score potential target genes in yeast. A predicted total occupancy score for a given TF was calculated for the upstream promoter region of each gene by summing the background-subtracted median PBM signal intensities for each overlapping 8-mer, considering all those 8-mers with a PBM enrichment score of at least 0.35, over the sequence up to 600 bp upstream of translation start. (If a verified ORF overlapped with this 600 bp upstream sequence, the sequence to score was truncated at the beginning of this nearby ORF) In this summation, we used the median intensities calculated for all sequence variants of the 8-mer pattern that produced the highest overall enrichment score, considering both contiguous 8-mers as well as 8-mers spanning up to 10 total positions, or 8-mers of the pattern 1111-gap-1111 with gaps up to 12 positions. The median value of the median intensities over all 8-mers was used as a measure of the background signal and was subtracted from each individual 8-mer's intensity before summation. For TFs with duplicate PBM data, the ranks of the genes resulting from median intensity summing from each individual array were averaged to produce the final gene rankings for subsequent analysis.

I. Analysis of ChIP-chip data

To determine how well our *in vitro* PBM data agree with *in vivo* binding data for the TFs, we analyzed the PBM-derived scores for ChIP-chip derived target intergenic regions (IGRs) of each TF. All yeast IGRs were scored using PBM 8-mer data as described above for each TF, except that in this case the entire IGR length was scored, rather than only 600 bp upstream of translation start. This enables a direct comparison between scores derived from PBM data and those measured from ChIP-chip experiments, in which the entire IGR (~100-1500 bp in length) was included on the array. Target IGR sets for each TF were defined as IGRs bound by the TF in a ChIP-chip experiment at $p<0.001$ in any experimental condition as reported by the authors of that study (Harbison et al. 2004). In cases where the results from multiple experimental conditions were combined, the ChIP-chip p-values were Bonferroni-corrected for multiple hypothesis testing before the target gene set was chosen. TFs were excluded from the analysis if fewer than 10 IGRs were bound at $p<0.001$ in the ChIP-chip data. An area under the receiver operating characteristic (ROC) curve (AUC statistic) was then calculated by comparing the PBM-derived ranks of IGRs within the ChIP-chip 'bound' IGRs (foreground set, or "class 1") to the ranks of the rest of the yeast IGRs (background set, or "class 0"). For comparison, ChIP-chip-derived motifs, if available, were used to rank the IGRs as well. ScanACE (Hughes et al. 2000; Roth et al. 1998) was used to score ChIP-chip motif matches in all yeast IGRs at a threshold of 2 SD below the mean motif score. If multiple matches occurred within an IGR, these scores were summed to obtain a final score for each IGR. The resulting ChIP-chip IGR ranking was then used to calculate an AUC statistic comparing the ChIP-chip derived ranks for ChIP-chip target IGRs versus background IGRs. P-values measuring the significance of each AUC statistic were calculated using the mean and standard deviation of a distribution of AUC scores derived using 1,000 random permutations of the foreground/background assignments for each set of IGR scores. This p-value is sensitive to sample size (*i.e.*, to the size of the foreground and background sets), whereas the AUC is much more robust to sample size (measures what can be thought of as 'effect size' while p-value measures 'statistical significance'). Therefore, we use the AUC to compare datasets and the p-value to draw a cutoff on statistical significance. We consider AUC scores with $p<0.05$ as significant enrichment of likely PBM-predicted target genes within the ChIP-chip binding data. We noted any cases in which PBM-derived ranks better

explained the ChIP-chip target IGRs (higher AUC) than did the ChIP-chip motif ranks. For three TFs (Aro80, Stp4, and Yox1), fewer than 100 IGRs contained any ScanACE matches to the ChIP-chip derived motif and thus there were not enough uniquely ranked values to allow a reliable p-value estimation by the permutation-based method. All other p-values are reported in Fig. S6.

Disagreements between the ChIP-chip p-values and PBM scores may have a variety of explanations. In addition to the potential differences discussed in the main body of the paper, the ChIP-chip data may suffer from false positive signals due to cross-hybridization. The long intergenic region fragments present on the array increase the probability that the short sheared DNA fragments immunoprecipitated by the ChIP antibody may recognize regions within array fragments other than their appropriate target sequence.

To determine which ChIP-chip bound regions would be classified as "direct" versus "indirect" binding sites according to our PBM *k*-mer data, we required a scoring method that would provide discrete groups of target intergenic regions instead of the continuous ranking of regions accomplished by the methods described above. Thus, we searched for high scoring *k*-mers within IGRs, using E score > 0.45 as a consistent threshold between TF datasets. We counted any IGR bound by a certain TF in ChIP-chip that contained at least one *k*-mer for that TF with an E score > 0.45 as a likely "direct" target of that TF, while ChIP-chip bound IGRs without *k*-mers surpassing this threshold were considered likely "indirect" targets. In Fig. S7, we also report the numbers of direct and indirect binding targets according to permissive versus conservative thresholds. We did not require cross-species conservation for either our *k*-mer matches or for the sites identified by MacIsaac *et al.*, since a recent ChIP-chip study provided evidence that many functionally orthologous TF binding sites may not be aligned in genome sequence alignments (Borneman et al. 2007).

J. Analysis of functional category enrichment among a TF's predicted target genes

We calculated the predicted total occupancy score at the promoter of a yeast gene for a given TF

using the median intensity 8-mer summation method described above. We then considered the top 200 scoring genes for analysis of functional category enrichment among a TF's predicted target genes using the tool FunSpec (Robinson et al. 2002). We report results obtained at 2 different p-value significance thresholds as calculated by FunSpec, including $p<0.005$ without Bonferroni correction and $p<0.05$ with Bonferroni correction. In general, the top 200 genes gave much more informative results than 100 in the positive control TFs of known function, and so we used the top 200 genes for all the TFs.

K. Prediction of condition-specificity using CRACR

We utilized our CRACR (Combination Rank-order Analysis of Condition-specific Regulation; pronounced "cracker") algorithm for prediction of the condition-specific functions of *S. cerevisiae* TFs (TFs) by integrating many (currently 1,693) microarray gene expression data sets and TF binding data from PBMs essentially as described previously (McCord et al. 2007).

Briefly, CRACR searches for conditions in which genes downstream of intergenic regions (IGRs) exhibiting significant TF binding in PBMs are enriched among differentially expressed genes. Each gene in the yeast genome is first ranked by the predicted total occupancy of the sequence up to 600 bp upstream of its translational start site by a TF as calculated from PBM data using 8-mer median intensity summations as described above. We then order all yeast genes according to their expression in a single condition and use a mean-centered area under a receiver operating characteristic (ROC) curve (AUC) statistical test to compare the PBM-defined ranks of similarly expressed genes within a sliding foreground window to the ranks of a background set of genes outside this window. This "area statistic" (McCord et al. 2007) indicates whether the similarly expressed genes within each window are enriched (positive area) or depleted (negative area) for likely TF targets. The statistical significance of the maximum enrichment in an expression condition is determined by permutation testing. In addition to discovering individual conditions in which a TF is likely to be regulating its target genes, each individual expression dataset is annotated with terms describing the biological functions affected by the given experimental condition (i.e., "heat shock response" or "sporulation") to facilitate a search for the

general biological functions of a particular TF (Marion et al. 2004), FuncAssociate (Berriz et al. 2003) is used to calculate the enrichment of these "condition annotation terms" within a set of conditions significant for a TF using a file describing the associations between annotation terms and conditions analogous to an association file between genes and GO terms, allowing the calculation of an annotation enrichment p-value adjusted for multiple hypothesis testing and accounting for the degree of significance of each condition as measured by its maximum area statistic.

L. Prediction of transcriptional co-regulation by TFs

Using the TFs Pbf1, Pbf2, and Stb3, which clustered together in the CRACR clustergram, consistent with their PAC and RRPE binding site motifs co-regulating rRNA processing and transcription target genes (Hughes et al. 2000), as a guide in setting a conservative CRACR similarity threshold, we considered all TF clusters with equal or greater CRACR similarity as that of the Pbf1/Pbf2/Stb3 cluster, as groups of putative co-regulatory TFs.

M. EMSAs

EMSA DNA Probe Design:

Based on DNA binding specificities derived from PBMs, a 60-nt EMSA probe was designed such that the 5' 40 nt sequence corresponds to a putative target intergenic region in the yeast genome and contains the predicted DNA binding site, and the next 20 nt corresponds to a common priming sequence at the 3' end that can anneal to a universal biotinylated primer. The negative probes were designed such that their sequences either don't contain the putative target site or contain a mutated version of the target site. Specific probe sequences used in EMSAs were as follows:

Probes for Yer130c EMSAs:

Positive probe iYOR259C:

5' – CACCGTTGATAGCTCCCCCCCCCTATTGTCGTAGTCATGTC*GAAAGGATGGGTGCGACGCG* - 3'

Positive probe iYGL158W:

5' – GATTTTCCGCGGTTCCGACCCCCTATCCTAGAAACACGGAA*GAAAGGATGGGTGCGACGCG* - 3'

Positive probe itT(UGU)G2:

5' – CCACGGAGGTTATCTTCACCCCCTATTTTCATTTTAAATTT*GAAAGGATGGGTGCGACGCG* - 3'

Negative probe iYGR145W:

5' – AGGTTTTTTTTTTTTAGGATGACTAGAAAAGGAAATCGAA*GAAAGGATGGGTGCGACGCG* - 3'

Universal biotinylated primer:

5' – Biotin-CGCGTCGCACCCATCCTTTC - 3'

Probes for Pbf1 and Pbf2 EMSAs:

Positive probe iYKL144C:

5' – CAAATAAAAATTTTAAGCGATGAGCATCGCCTGAATATTA*CTGATTGCGGCGACCCATGG* - 3'

Negative probe iYKL144CmP:

5' – CAAATAAAAATTTTAAGCCATCACCATCGCCTGAATATTA*CTGATTGCGGCGACCCATGG* - 3'

Universal biotinylated primer:

5' – Biotin-CCATGGGTCGCCGCAATCAG - 3'

In the above probe sequences, the predicted binding sites are indicated by underlining and the common 3' primer sequence is indicated by italics.

Primer extensions to create double-stranded DNA EMSA Probes:

Lyophilized oligonucleotides (Integrated DNA Technologies, Inc.) were resuspended in TE pH 8.0 to a working stock of 100 µM. Primer extensions reactions were performed in 1x Thermopol

Buffer (NEB, 20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$,10 mM KCl, 2 mM $MgSO_4$, 0.1 % Triton X-100) using final concentrations of 0.8 mM dNTPs (Amersham), 4 µM 20-nt common primer, and 4 µM 60-nt template oligonucleotide in a 25 µL reaction. Primer extensions were performed in a thermocycler using the following protocol:

1) 95° C for 3 minutes
2) Ramp down to 60° C (0.1° C per second)
3) Hold at 60° C

At Step 3, 8 units of *Bst* Polymerase Large Fragment (NEB) in 1x Thermopol buffer were added to each reaction as a "hot start". After polymerase was added, the reaction was allowed to continue to Steps 4 and 5:

4) Incubate at 60° C for 90 minutes
5) Hold at 3° C

The 60-bp double-stranded EMSA probes were filtered using MinElute PCR Purification Kit (Qiagen) according to the manufacturer's instructions, and diluted to a working concentration of 10 nM after concentrations were determined using a UV spectrophotometer.

EMSA experiments:

EMSAs were performed using the LightShift© Chemiluminescent EMSA Kit (Pierce), essentially according to the manufacturer's protocols. Briefly, each 20 µl binding reaction contained 1x Binding Buffer (10 mM Tris, 50 mM KCl, 1 mM DTT), 2.5% glycerol, 0.5 µg Salmon Testes DNA (Sigma), 10 mM KCl, 4 µg BSA (NEB), 0.05 % NP-40, and 50 µM zinc acetate (for use in Yer130c EMSA, since Yer130c is a ZnF-C2H2 protein, but not in Pbf1 and Pbf2 EMSAs, since Pbf1 and Pbf2 are not zinc finger proteins). Approximately 5 nM DNA probe and approximately 0.2 µM protein were used in each reaction. The binding reactions were allowed to incubate at room temperature for 1 hour. A precast 6% polyacrylamide DNA

retardation gel (Invitrogen) was pre-run for 30 minutes at 100 V, and then 5 µl of 5x loading buffer was added to the binding reaction, and subsequently 20 µl of the reaction was run on the gel at 100 V for 45 minutes. The gel was then transferred to a charge-modified 0.45 µm nylon membrane (Sigma) for 1.5 hours at 100 V, and subsequently UV-crosslinked to the membrane at 120 µJ/cm^2. The membrane was then treated with developing buffers (Lightshift Blocking Buffer with stabilized Streptavidin-Horse Radish Peroxidase conjugate, Wash Buffer, Substrate Equilibration Buffer, Luminol/Enhancer Solution and Peroxide Solution) according to manufacturer's protocol, then exposed to X-ray film and developed.

N. Yeast strains and growth conditions

All yeast strains were isogenic with *BY4741*. *BY4741*, *ΔPBF1* and *ΔPBF*2 were purchased from Open Biosystems. The *ΔPBF1ΔPBF2* double deletion mutant was generated by replacing *PBF2* with *URA3* using standard homologous recombination protocol in the *ΔPBF1* background (Baudin et al. 1993). PCR primer sequences used for PBF2 deletions are 5'-CCGTGCACGTTCCAGTCTTCCCTCCCTTCTCTGCTCCGTGAGATTGTACTGAGAGTGCAC-3' and 5'-TTTTTTTATTTTTATTTTTTTTTCATTTTAAGTTTTCCCCCTGTGCGGTATTTCACACCG-3'.

We used the PCR epitope-tagging procedure (Schneider et al. 1995) to generate yeast strains with a 3xHA (hemagglutinin) N-terminal epitope tag using plasmid pMPY-3xHA (Schneider et al. 1995). PCR primer sequences for Pbf1-3xHA are: (Forward: 5' -AAGAATATATCACTGTTCTTATTGAAGTTCCCTCGCGATGACTAGGGAACAAAAGCTGGA- 3'
and Reverse: 5' -TAACATGTCCTGATGAAACAGAAACGCTACTGAGTTTCGGCAACTGTAGGGCGAATTGGG- 3') and for Pbf2-3xHA are: (Forward: 5' -TGCACGTTCCAGTCTTCCCTCCCTTCTCTGCTCCGTGATGTCCAGGGAACAAAAGCTGGA- 3'

and Reverse: 5' -
TGCTGCTCAAATGAATGGAAGCTGAGTTCAAACTGGTTGAAATCTGTAGGGCGAATT
GGG- 3'). The *URA3* marker was further looped out using 5-fluoro-orotate selective media. All yeast were grown in standard yeast YPD medium as described (Sherman 2002) if not otherwise specified.

O. Chromatin immunoprecipitation (ChIP) and quantitative real-time PCR (ChIP-qPCR)

We carried out chromatin immunoprecipitation as described previously (Aparicio et al. 2005) with minor modifications. An overnight culture from a single colony of Pbf1 or Pbf2 HA-tagged strains was diluted into fresh medium and grown to OD_{600} between 0.3 and 0.4. Three independent cultures were grown in parallel in order to carry out triplicate biological replicates for ChIP assays. Cells were then subjected to heat shock treatment, i.e., growth temperature shifted from 25°C to 37°C, for 5 min, and then they were fixed with 1% formaldehyde for 20 min, and then quenched by ~1.5 M glycine for 5 min. Cells were then washed and resuspended in FA lysis buffer / 2 mM PMSF (50 mM HEPES, pH7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS) and lysed with zirconia-silica beads (Biospec Products) by vortexing at 4°C for 90 min. Chromatin was pelleted by centrifugation and solubilized by sonication. Soluble chromatin was then immunoprecipitated using monoclonal anti(HA-tag) antibody (12CA5, Santa Cruz Biotechnology) at 4°C overnight with rotation. Protein A Sepharose beads (GE Healthcare) were added to the sample and incubated for one additional hour at 4°C. Chromatin was then washed and recovered. Peptides and remaining proteins were digested with Pronase (Roche) and crosslinking chromatin was reversed by incubating at 65°C overnight, followed by DNA purification using Qiagen PCR-purification spin column. PCRs were performed using iQ™ SYBR Green SuperMix (Bio-Rad) on an iCycler real-time PCR thermocycler. The enrichment was defined as the ratio of the PCR product amount in the "IP" sample versus "INPUT" sample using an open reading frame (ORF)-free region on chromosome V (ChrV) as a control. ChIP-qPCR primer sequences are as follows:

ChrV forward: 5'- AGCTAGGTGAGAGAAAGCAAAGGT -3'

ChrV reverse: 5'- AGTGTGTACGATCTTAGTTCCAATGG -3'

ENO2 forward: 5'- AGTCAGCATACACCTCACTAGGGT -3'
ENO2 Reverse: 5'- GCTTGGTGCCACTTGTCACATACA -3'

SAS10 forward: 5'- GCATCAGTGAGATGAGCTATGATGAG -3'
SAS10 reverse: 5'- AAATTGGACTTTGCAGGGCGGATG -3'

NOP2 forward: 5'- TGAGTAGGATCCAACGTGCCAAAG -3'
NOP2 reverse: 5'- GCGACAACTGTATTTGCAGCTC -3'

MTR4 forward: 5'- TCACTTTCTTGCGATGAGATGCAC -3'
MTR4 reverse: 5'- GCAGGGAATGTTGAGTCACCGAAA -3'

KRR1 forward: 5'- GCCAATTTGGATATTTGTGTGACCC -3'
KRR1 reverse:5'- TAGCAGGCTTGCACATCTGA -3'

ERB1 forward: 5'- CAGTACCTTTCTTCGCTAGGATCT -3'
ERB1 reverse: 5'- TGCTAATTAAGTAGGATTGAATTGTCGC -3'

P. Gene expression profiling and quantitative RT-PCR (RT-qPCR)

Three independent cultures for each of the *BY4741*, *ΔPBF1*, *ΔPBF*2 and *ΔPBF1ΔPBF2* strains were grown in parallel in order to carry out triplicate biological replicates. Cells were then subjected to heat shock treatment, i.e., growth temperature shifted from 25°C to 37°C, for 20 min, and subsequently spun down and flash-frozen at -80°C. RNA was extracted and purified using Qiagen RNeasy Mini kit with DNase I treatment. Gene expression profiling was performed using Affymetrix™ Yeast Genome 2.0 GeneChip® oligonucleotide arrays essentially following the manufacturer's protocol. Microarray data were analyzed as described previously (Choe et al. 2005). We imposed a false discovery rate (FDR) of 0.0001 as the cut-off value to identify

differentially expressed genes. GO term enrichment analysis was performed by applying FuncAssociate (Berriz et al. 2003) on lists of differentially expressed genes ordered by their expression ratio. Microarray data were deposited into the GEO database with accession number GSE13684.

qPCR reactions were performed as described above. RT-qPCR primer sequences are as follows:

ACT1 forward: 5'- ACGTTCCAGCCTTCTACGTTTCCA -3'
ACT1 reverse: 5'- ACGTGAGTAACACCATCACCGGAA -3'

ENO2 forward: 5'- TCATTGCTGCTGCTTTCGTCAAGG -3'
ENO2 Reverse: 5'- TAGCGTTAGCACCCAACTTGGACT -3'

SAS10 forward:5'- GCTGATGTGGACGCACAAGACAAA -3'
SAS10 reverse: 5'- CATCCAACAGACGTTGCTGCCTTT -3'

NOP2: forward: 5'- ACGTCGATGGGTTCTTTGTCGCTA -3'
NOP2 reverse: 5'- TTCTTCGTCCTCGAAGGTTGCGAA -3'

MTR4 forward: 5'- GCTTGCCGAACCTTTGAAGGCTAT -3'
MTR4 reverse: 5'- ACGAACCTTCGTAAACGTCGGTCA -3'

KRR1 forward: 5'- TTTCCTCCTGCCCAATTGCCTAGA -3'
KRR1 reverse: 5'- TGCTCTCTCTTCCTGCCTTTCGAT -3'

ERB1 forward: 5'- TATGGCACGACCTGGATTTGGCTA -3'
ERB1 reverse: 5'- TCATCAGCGGCAGAGCTGAATAG -3'

Supplementary References

Angus-Hill, M.L., Schlichter, A., Roberts, D., Erdjument-Bromage, H., Tempst, P., and Cairns, B.R. 2001. A Rsc3/Rsc30 zinc cluster dimer reveals novel roles for the chromatin remodeler RSC in gene expression and cell cycle control. *Mol Cell* **7:** 741-751.

Aparicio, O., Geisberg, J., Sekinger, E., Yang, A., Moqtaderi, Z., and Struhl, K. 2005. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences *in vivo*. *Current Protocols in Molecular Biology***:** 21.23.21.-21.23.33.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., and Cullin, C. 1993. A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic Acids Res* **21:** 3329-3330.

Beer, M.A. and Tavazoie, S. 2004. Predicting gene expression from sequence. *Cell* **117:** 185-198.

Berger, M. and Bulyk, M. Universal protein binding microarrays for the comprehensive characterization of the DNA binding specificities of transcription factors. *Nature Protocols***:** in press.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., Khalid, F., Zhang, W., Newburger, D., Jaeger, S.A., Morris, Q.D., Bulyk, M.L., and Hughes, T.R. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133:** 1266-1276.

Berger, M.F. and Bulyk, M.L. 2006. Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol Biol* **338:** 245-260.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24:** 1429-1435.

Berriz, G.F., King, O.D., Bryant, B., Sander, C., and Roth, F.P. 2003. Characterizing gene sets with FuncAssociate. *Bioinformatics* **19:** 2502-2504.

Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. 2007. Divergence of transcription factor binding sites across related yeast species. *Science* **317:** 815-819.

Braun, P., Hu, Y., Shen, B., Halleck, A., Koundinya, M., Harlow, E., and LaBaer, J. 2002. Proteome-scale purification of human proteins from bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **99:** 2654-2659.

Brigati, C., Kurtz, S., Balderes, D., Vidali, G., and Shore, D. 1993. An essential yeast gene encoding a TTAGGG repeat-binding protein. *Mol Cell Biol* **13:** 1306-1314.

Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M., and Halfon, M.S. 2005. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol* **6:** R16.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71-76.

Consortium, T.G.O. 2001. Creating the gene ontology resource: design and implementation. *Genome Res.* **11:** 1425-1433.

Costanzo, M.C., Hogan, J.D., Cusick, M.E., Davis, B.P., Fancher, A.M., Hodges, P.E., Kondu, P., Lengieza, C., Lew-Smith, J.E., Lingner, C., Roberg-Perez, K.J., Tillberg, M., Brooks, J.E., and Garrels, J.I. 2000. The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res* **28:** 73-76.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. 1996. Life with 6000 genes. *Science* **274:** 546, 563-547.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E., and Young, R.A. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99-104.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32:** D258-261.

Hu, Y., Rolfs, A., Bhullar, B., Murthy, T., Zhu, C., Berger, M., Camargo, A., Kelley, F., McCarron, S., Jepson, D., Richardson, A., Raphael, J., Moreira, D., Taycher, E., Zuo, D., Mohr, S., Kane, M., Williamson, J., Simpson, A., Bulyk, M., Harlow, E., Marsischky, G., Kolodner, R., and LaBaer, J. 2007. Approaching a complete repository of sequence-verified, protein-encoding clones for *Saccharomyces cerevisiae*. *Genome Res***:** Epub Feb 23., 2007.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296:** 1205-1214.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241-254.

Lee, T., Rinaldi, N., Robert, R., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, D., Ren, B., Wyrick, J., Tagne, J., Volkert, T., Fraenkel, E., Gifford, D., and Young, R. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799-804.

Liu, Z.P. and Tye, B.K. 1991. A yeast protein that binds to vertebrate telomeres and conserved yeast telomeric junctions. *Genes Dev* **5:** 49-59.

MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D., and Fraenkel, E. 2006. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7:** 113.

Marion, R.M., Regev, A., Segal, E., Barash, Y., Koller, D., Friedman, N., and O'Shea, E.K. 2004. Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc Natl Acad Sci U S A* **101:** 14315-14322.

McCord, R.P., Berger, M.F., Philippakis, A.A., and Bulyk, M.L. 2007. Inferring condition-specific transcription factor function from DNA binding and gene expression data. *Molecular Systems Biology* **3:** 100.

Mewes, H., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. 2002. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30:** 31-34.

Morozov, A.V. and Siggia, E.D. 2007. Connecting protein structure with predictions of regulatory sites. *Proc Natl Acad Sci U S A* **104:** 7068-7073.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36:** 1331-1339.

Narlikar, L., Gordan, R., Ohler, U., and Hartemink, A.J. 2006. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics* **22:** e384-392.

Philippakis, A.A., Qureshi, A.M., Berger, M.F., and Bulyk, M.L. 2008. Design of compact, universal DNA microarrays for protein binding microarray experiments. *J. Comput. Biol. (presented at RECOMB 2007 conference)* **15:** 655-665.

Robinson, M., Grigull, J., Mohammad, N., and Hughes, T. 2002. FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics* **3:** 35.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16:** 939-945.

Schneider, B.L., Seufert, W., Steiner, B., Yang, Q.H., and Futcher, A.B. 1995. Use of polymerase chain reaction epitope tagging for protein tagging in Saccharomyces cerevisiae. *Yeast* **11:** 1265-1274.

Schneider, T.D., Stormo, G.D., and Gold, L. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188:** 415-431.

Sherman, F. 2002. Getting started with yeast. *Methods in Enzymology* **350:** 3-41.

Sudarsanam, P., Pilpel, Y., and Church, G.M. 2002. Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. *Genome Res* **12:** 1723-1731.

Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* **22:** 281-285.

Workman, C.T., Mak, H.C., McCuine, S., Tagne, J.B., Agarwal, M., Ozier, O., Begley, T.J., Samson, L.D., and Ideker, T. 2006. A systems approach to mapping DNA damage response pathways. *Science* **312:** 1054-1059.

Workman, C.T., Yin, Y., Corcoran, D.L., Ideker, T., Stormo, G.D., and Benos, P.V. 2005. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res* **33:** W389-392.