**Supplementary Methods**

This paper describes the genomic location of 203 transcriptional regulators, a subset of which are examined under different environmental conditions. We previously reported the genomic binding information for 106 regulators profiled in a single growth condition[1]; we have repeated experiments for 44 of these regulators to improve the quality of the complete dataset (available at http://web.wi.mit.edu/young/regulatory_code). We have also introduced additional data analysis features to reduce noise and improve the results.

**Genetic Reagents**

The 203 transcriptional regulators were identified by searching the YPD and MIPS databases[2-4] for known and predicted transcription factors and nucleic acid binding proteins. Yeast strains were created for each of the 203 regulators in which a repeated Myc epitope coding sequence was integrated into the endogenous gene encoding the regulator. PCR constructs containing the Myc epitope coding sequence and a selectable marker flanked by regions of homology to either the 5' or 3' end of the targeted gene were transformed into the W303 yeast strain Z1256. Genomic integration and expression of the epitope-tagged protein were confirmed by PCR and Western blotting, respectively.

**Growth conditions**

Regulators were selected for profiling in a specific environment if they were essential for growth in that environment or if there was other evidence implicating them in regulation of gene expression in that environment.

A brief description of the environmental conditions used follows:

*Rich media*. Cells were grown in YPD (1% yeast extract/2% peptone/2% glucose) to an OD600 of ~0.8.

*Highly hyperoxic*. Cells were grown in YPD to an OD600 of ~0.5 followed by treatment with hydrogen peroxide (4 mM final) for 30 minutes.

*Moderately hyperoxic*. Cells were grown in YPD to an OD600 of ~0.5 followed by treatment with hydrogen peroxide (0.4 mM final) for 20 minutes.

*Amino acid starvation*. Cells were grown to an OD600 of ~0.6 in synthetic complete medium followed by treatment with the inhibitor of amino acid biosynthesis sulfometuron methyl (0.2 μg/ml final) for two hours.

*Nutrient deprived*. Cells were grown in YPD to an OD600 of ~0.8 followed by treatment with rapamycin (100 nM final) for 20 minutes.

*Filamentation inducing*. Cells were grown in YPD containing 1% butanol for either 90 minutes or 14 hours (corresponding to an OD600 of ~0.8).

*Mating inducing*. Cells were grown in YPD to an OD600 of ~0.8 followed by treatment with the alpha factor pheromone (5 μg/ml) for 30 minutes.

*Elevated temperature*. Cells were grown in YPD at 30ºC to an OD600 of ~0.5 followed by a temperature shift to 37ºC for 45 minutes.

*Galatose medium*. Cells were grown in YEP medium supplemented with galactose (2%) to an OD600 of ~0.8.

*Raffinose medium*. Cells were grown in YEP medium supplemented with raffinose (2%) to an OD600 of ~0.8.

*Acidic medium*. Cells were grown in YPD to an OD600 of ~0.5 followed by treatment for 30 minutes with succinic acid (0.05 M final) to reach a pH of 4.0.

*Phosphate deprived medium*. Cells were grown in synthetic complete medium lacking phosphate to a final OD600 of ~0.8.

*Vitamin deprived medium*. Cells were grown in synthetic complete medium lacking thiamin to a final OD600 of ~0.8.

**Genome-wide Location Analysis**

Genome-wide location analysis was performed as previously described[1,5,6]. Bound proteins were formaldehyde-crosslinked to DNA in vivo, followed by cell lysis and sonication to shear DNA. Crosslinked material was immunoprecipitated with an anti-myc antibody, followed by reversal of the crosslinks to separate DNA from protein[7,8]. Immunoprecipitated DNA and DNA from an unenriched sample were amplified and differentially fluorescently labeled by ligation-mediated PCR. Triplicate samples were hybridized to a microarray consisting of spotted PCR products representing the intergenic regions of the *S.cerevisiae* genome. Detailed protocols are available on the authors' website.

**Microarray design**

Using the Yeast Intergenic Region Primer set (Research Genetics) we PCR amplified and printed approximately 6000 DNA fragments, representing essentially all of the known intergenic regions in the yeast genome[9]. The average size of the spotted PCR products was 480 bp, and the sizes ranged from 60 bp to 1500 bp.

**Raw Data Analysis**

The microarrays were scanned using an Axon200B scanner, and the images were analyzed with Genepix 5.0. Columns corresponding to the background subtracted intensities and standard deviation of the background were extracted for further analysis. The intensities for the two channels, representing the immunoprecipitated (test) and unenriched (control) samples, were normalized by using the median of each channel to calculate a normalization factor, normalizing all datasets to a single median intensity. The log ratio of the intensity in the test channel to the control channel was calculated. To

account for biases in the immunoprecipitation reaction, these log ratios were normalized for each spot by subtracting the average log ratio of each spot across all arrays. The intensities in the test channel were then adjusted to yield this normalized ratio. Finally, an error model[10] was used to calculate significance of enrichment on each chip and to combine data for replicates to obtain a final average ratio and significance of enrichment for each intergenic region. Each intergenic region was assigned to the genes it is most likely to regulate, as described on the author's website.

We have included new refinements in our analysis relative to that used in Lee et al.[1]. Notably, we have excluded artefactual spots from analysis, selected more reliable probes for normalization and assigned quality metrics to individual arrays to identify low quality experiments.

**Error Estimates**

We previously estimated a false positive rate of 6-10% for genome-wide binding data that meets a $P \leq 0.001$ threshold. The present study is focused on DNA regions that are both bound ($P \leq 0.001$) and contain a conserved match to a binding site specificity. Of 47 sites that were used by Lee et al.[1] to determine the error rate and that met our criteria for binding sites, 45 were confirmed by independent gene-specific ChIP experiments. Thus, the frequency of false positives in this dataset is likely to be approximately 4%.

The false negative rate is more difficult to estimate, but it is likely to be approximately 24% in the present genome location dataset. This estimate was derived by determining the number of binding interactions reported in the literature for cell cycle regulators that were not identified in the genome-wide location data at $P \leq 0.001$ and associated with conserved binding sites (12/50). We selected the cell cycle literature for analysis because of the extensive study of this group of regulators and their targets.

**Motif Discovery Overview**

Binding motifs were identified in a five-step process described in detail below and summarized in Supplementary Figure 2. First, motifs were discovered by applying a suite of motif discovery programs to the intergenic sequences identified by the binding data. The resulting specificity predictions were filtered for significance using uniform metrics and then clustered to yield representative motifs. Conservation-based metrics were used to identify the highest-confidence subset of these motifs. For cases in which multiple significant binding motifs were found for a factor, we used statistical scores or information from the Transfac[11], YPD[12], and SCPD[13] databases to choose a single motif for each regulator. Sequence input files, intermediate motif discovery output, and matrix representations of the finalized motifs are available on the authors' website.

**Step 1: Initial Motif Discovery**

Motif Discovery Programs have different strengths with respect to finding specificities. To gain as comprehensive an analysis as possible, we applied five different motif-finding programs to the binding data: AlignACE[14], MEME[15], MDscan[16], the conservation-based method described in Kellis et al.[17] , and a new conservation-based method called

CONVERGE (described below). The MEME program was also used to analyze a modified input that incorporated conservation information (see "Probe Sequences").

To make the search more thorough, we ran each of these programs multiple times with different parameters. AlignACE was run using the default settings ten times with different random number seeds, in order to increase the motif space it sampled. The results from the AlignACE runs were grouped together for analysis. MEME was run using the supplied 5$^{th}$-order Markov background model, the "ZOOPS" motif model, and the "-minsites 20 -dna -revcomp" options. MEME runs were repeated using motif width ranges of 7 to 11 and 12 to 18. To run MDscan, seqeuences were ranked according the *P*-value of binding, and the program was run with the "-s 30 -r 5 -t 10" options. To compensate for the fact that MDscan searches only for motifs of fixed width, the program was run repeatedly, once with each width in the range 8 to 15 bases. The method of Kellis et al. was applied to the data as described[17]. CONVERGE was run twice using motif widths of 8 and 15.

*MEME_c*

We tested whether we could improve the performance of AlignACE, MEME and MDscan by modifying the input sequences to convey the conservation of each base in the *sensu stricto Saccharomyces* species. Using ClustalW[18] alignments for the *sensu stricto* species[17], we replaced a base in the *Saccharomyces* genome with the letter "N" if it was not conserved in 2/3 or 3/4 of the other genomes. Of the programs we tested, only MEME was able to use the modified sequences.

*CONVERGE*

We designed CONVERGE to identify motifs that are both over-represented in a set of input sequences and conserved across multiple genomes. CONVERGE input sequences consists of an ungapped DNA sequence corresponding to the primary genome, as well as one or more optional aligned sequences, which may contain gaps. The algorithm is based on the ZOOPS model of MEME and uses a 5$^{th}$-order Markov background model. However, whereas MEME searches for matches to a motif model across a set of input sequences, CONVERGE searches across the multiple-sequence alignments for each sequence. Specifically, CONVERGE treats the probability of a motif occurring at a site in the alignment as the product of the probabilities of the motif occurring at the same site in each of the aligned sequences. Thus, CONVERGE defines a site as conserved in a flexible manner that depends on the motif being discovered. Full details will be presented elsewhere.

*Probe Sequences*

Motif discovery programs were applied to the sequences of probes bound with a *P*-value $\leq 0.001$. We found that some intergenic regions were highly homologous over their entire length, and consequently skew the results of motif discovery since all subsequences are overrepresented. To remove this bias, we used BLAST[19] to identify pairs of probes with high sequence similarity over 50% of their lengths. For each pair, the shorter intergenic region was omitted from motif discovery computations. This process removed up to nine regions for some experiments, but less than one on average.

To determine the sequences present on the microarrays, we computed the expected products of the PCR used to construct the arrays. Research Genetics primer sequences were obtained from http://www.resgen.com/products/YeIRP.php3 and the March 2002 revision of the yeast genome was obtained from SGD[20]. Probes that were predicted to amplify more than two different genomic sequences were omitted from the calculations. Twenty five probe sequences neighboring repetitive, non-transcribed features (e.g. telomeric repeats, X elements and Y' elements) were also omitted.

*PSSM Representation*

Motifs from all programs were converted to a standard position-specific scoring matrix (PSSM) for subsequent analysis. AlignACE and MDscan produce alignments of binding sites, and these were first converted into matrices representing the frequency of each base (A, C, G, T) at each position of the alignments. The method of Kellis et al. represents motifs as text strings containing ambiguity codes, which were also converted to matrices of frequencies. (For example, if a motif contained the letter "S" at a particular position, a value of 0.5 would be assigned to both "C" and "G.") The matrices of base frequencies were converted to probabilities and then were adjusted with 0.001 pseudo-counts in proportion to the $0^{th}$-order background probabilities ($3.1 \times 10^{-4}$ pseudocounts for A and T, $1.9 \times 10^{-4}$ pseudocounts for G and C). Log-likelihood scores were computed by dividing the estimated probabilities by the background probability for each letter and computing the base-2 logarithm. CONVERGE and MEME both provide probability matrices, which were used directly.

**Step 2: Motif Scoring and Significance Testing**

We tested the significance of each motif by comparing how often it was found in the bound and unbound probes. To encapsulate different approaches to measuring motif over-representation, we employed three different metrics: Enrichment, ROC AUC, and for motifs discovered by the method described in Kellis et al., the "CC4" score. The enrichment score is a direct measure of the occurrence of a motif among bound probes compared to all possible gene targets, but does not distinguish between the number of motifs occurrences within each intergenic region. The ROC AUC metric is more sensitive to cases in which the number of motif occurrences is a distinguishing factor. Finally, the CC4 metric provides a way to account for the importance of the conservation of the motif among bound probes. These scores were compared to significance thresholds obtained from calculations on randomized selections of intergenic regions as described below in "Significance Thresholds"

*Enrichment score*

To obtain the enrichment score, the hypergeometric distribution was used to compare the frequency of the motif in the bound probes to that which would be expected if the intergenic regions were selected at random from the genome. A sequence was considered to contain a motif if it contained at least one or more sites scoring at least 70% of the maximum possible score of the matrix.

A *P*-value for the enrichment was computed according to the formula:

$$p = \sum_{i=b}^{\min(B,g)} \frac{\binom{B}{i}\binom{G-B}{g-i}}{\binom{G}{g}} \qquad (5)$$

where $B$ is the number of bound intergenic regions and $G$ is the total number of intergenic regions represented on the microarray (or the genome). The quantities $b$ and $g$ represent the number of intergenic regions of $B$ and $G$ matching the motif. The quantity $-log_{10}(p)$ is referred to as the enrichment score.

*ROC AUC (*Receiver Operating Characteristic Area Under Curve)
The ROC AUC refers to the area under a receiver operating characteristic curve which is assembled by ranking the sets of bound and unbound probes according to the number of motif matches they contain, and plotting the fractional rankings against each other. We used the method and code described by Clarke and Granek[21].

*Conservation CC4*
Motifs discovered using the method of Kellis et al.[17] were judged according to the CC4 metric, in which the occurrence of a conserved motif among the bound probes is compared to the expected ratio observed among all 3-gap-3 motifs in among the same set of bound probes. The binomial probability of the observed ratio was computed, and is reported in terms of the equivalent z-score.

*Significance Thresholds*
We observed that motif discovery programs produce motifs with high over-representation metrics (such as "Enrichment" and "ROC AUC") even when applied to random selections of intergenic regions. To identify the true motifs, we converted the scores from each metric into the empirical probability that a motif with a similar score could be found by the same program in randomly selected sequences. We accepted only those motifs with a $P$-value $\leq 0.001$. We selected this stringent threshold to minimize false positives, and because we observed empirically that it identified the correct motifs for many regulators with known specificity. To estimate these thresholds, we ran each program 50 times on randomly selected sequences on sets of 10, 20, 30, 40, 50, 60, 70, 80, 100, 120, 140, and 160 probes.

The observed scores from these random runs were parameterized by a normal distribution. The critical values equivalent to a $P$-value of 0.001 are provided in Supplementary Table 8 for each program and each metric. If the empirical distribution was not normal (by the Shapiro-Wilk test), the corresponding metric was not used to evaluate motifs generated by the relevant program for regulators with a similar number of bound probes.

For a particular experiment, we employed the threshold derived from the randomization set that had the size closest to the number of bound probe sequences. For example, suppose a motif found by performing ten runs of AlignACE on 32 intergenic sequences had an enrichment score of 25. The relevant score distribution has been obtained by

performing ten runs of AlignACE on each of 50 randomly selected sets of 30 intergenic sequences. The resulting distribution of enrichment scores has a mean of 14.1 and standard deviation of 2.1, and the enrichment that corresponds to significance of $P \leq 0.001$ is thus 20.43. Since the score of the candidate motif is higher, it is considered significant.

**Step 3: Motif Clustering and Averaging**

*K-medoids Clustering*

The set of significant motifs for each experiment was then clustered via k-medoids clustering[22] using the distance metric described below. The k-medoids algorithm was performed 500 times to find a clustering with a minimal sum of inter-cluster distances. To find the optimal number of clusters, this process was first performed with 10 clusters, and then repeated with incrementally fewer clusters until all average distances between members of a cluster and medoids of other clusters were sufficiently large (greater or equal to 0.18).

*Inter-Motif Distance*

We constructed a distance metric to aid in the comparison of motifs. The distance D between two aligned motifs "a" and "b" is defined as,

$$D(a,b) = \frac{1}{w}\sum_{i=1}^{w}\frac{1}{\sqrt{2}}\sum_{L\in\{ACGT\}}(a_{i,L} - b_{i,L})^2 \qquad (1)$$

where $w$ is the motif width, and $a_{i,L}$ and $b_{i,L}$ are the estimated probabilities of observing base $L$ at position $i$ of motifs $a$ and $b$, respectively. The normalizations by $w$ and $\sqrt{2}$ facilitate the interpretation as a fractional distance. For example, a distance of 0.20 indicates that the two motifs differ by about 20%.

In practice, the optimal alignment of motifs is not known. We therefore use the minimum distance between motifs among all alignments in which the motifs overlap by at least seven bases, or when the motifs are shorter, by 2 bases fewer than the shortest motif length. Alignments to the reverse complements of the motifs are included.

*Motif Averaging*

A single motif representing each cluster was computed by averaging the probabilities at each matrix position of the aligned motifs comprising the cluster. Low-information positions on the flanks of the averaged motifs were removed.

**Step 4: Conservation Testing for Averaged Motifs**

We tested the conservation of averaged motifs, and focused subsequent analysis on the motifs that met two conservation criteria: First, we required that the frequency of conserved instances of the motif compared to all instances of the motif be at least as high within bound intergenic regions as among all intergenic regions. Second, we required that discovered motifs have at least three conserved instances that are bound.

We considered a sequence a match to a motif if it had a score of at least 60% of the motif maximum. We defined a "conserved instance" to mean that the aligned sequence of at least two other *sensu stricto* species also matched the motif. In cases where fewer than two aligned sequences were available, a site was treated as "not conserved."

**Step 5: Assignment a Single Motif to Each Regulator**

Often, the motif discovery process produced several significant, distinct averaged motifs (3 on average.). These motifs could represent the desired binding specificity of the protein, or they might arise from the specificity of binding partners or have other biological significance. To identify those motifs representing the binding specificity of the profiled transcription factor, we compared the specificities to binding data in the Transfac[11], YPD[12], and SCPD[13] databases, when available, using the same inter-motif distance metric used for clustering (see above.) There were 21 regulators for which no such data were available. In these cases we chose the motif with the best enrichment score.

Specificity data from these databases is sometimes available in the forms of raw sequences, ambiguity codes, and matrices. For regulators without matrices, we assembled a single consensus sequence to represent the body of experimentally determined specificity information and converted it to a PSSM as described above. Since there is no way to independently assess the quality of the motifs assembled from the databases, we used a permissive threshold to detect similarity between the discovered motifs and the database motifs. Motifs scoring below 0.24 were accepted as matches, while motifs with scores less than 0.35 were examined manually. The scores for the motifs that were used in the Regulatory Code Map are provided in Supplementary Table 2.

**Motifs Derived from the Literature**

We used a motif derived from the databases for the remaining regulators for which either: (1) Too few intergenic regions (<10) were bound for effective motif discovery, (2) discovered motifs similar to the literature were eliminated by the conservation in Step 4, or (3) none of the discovered motifs matched the literature in Step 5. These motifs were only included if they had at least one conserved instance that was bound. The resulting compendium of 102 motifs (Supplementary Table 3) was used in all subsequent analysis.

**Regulatory Code Map**

Binding motifs for 102 regulators (Supplementary Table 3) were fused with location analysis data and conservation data to produce a map of active binding sites in intergenic regions. The entire map is available at http://web.wi.mit.edu/fraenkel/regulatory_map/. The map was constructed by finding all conserved occurrences of each motif within intergenic regions bound by the corresponding factor.

We used a binding *P*-value threshold of $P \leq 0.001$ and the definition of conservation as described in the "Conservation Test" section above. Variants of the map constructed with different binding and conservation thresholds are also available online.

Distributions of distances from the start codon (ATG) of open reading frames to binding sites in the adjacent upstream region were derived from the above data. These were compared to a distribution calculated on ten thousand "randomized" genomes in which the binding sites in each intergenic region were redistributed randomly and independently between the adjacent genes. The region from –100 to –500 (grey area in Figure 2c) contains many more binding sites than expected.

**Promoter Classification**

Promoters were classified based on the aggregate binding data from all experiments. A promoter was defined as having multiple regulator architecture if more than one regulator bound in the aggregate data, regardless of the number of regulators that bound in any particular condition. Similarly, a promoter was assigned to the single regulator architecture if it was bound by exactly one regulator in the aggregate data.

Regulators that had a tendency to use the repetitive motif architecture were identified by chi-square analysis. For each regulator, we calculated the number of promoters containing a single site and the number containing multiple sites. These values were then compared to the expected values based on the average for all factors.

Co-occurring regulatory motifs were determined based on *P* values representing the probability, based on the hypergeometric distribution, of finding the observed number of intergenic regions (or more) bound by both regulators under the null hypothesis that binding for the two regulators is independent.

**Regulator Behaviour Classification**

The binding of each regulator was compared in pair-wise fashion for every environmental condition in which that regulator was studied. Only regions bound at $P \leq 0.001$ and containing conserved matches to the corresponding motif were included in this analysis. Some regulators fall into multiple categories depending on exactly which conditions are compared.

For the "condition invariant" category the ratio of the overlap of bound probes for a regulator was greater than 0.66, and the ratio of the number of bound probes was between 0.66 and 1.5.

For the "condition enabled" category the regulator bound to no probes in one environment.

For the "condition expanded" category the ratio of the overlap of bound probes for a regulator was greater than 0.66, and the ratio of the number of bound probes was less than 0.66 or greater than 1.5.

For the "condition altered" category the regulator bound at least one probe in both environments and the ratio of the overlap of bound probes was less than 0.66.

**Experimental Confirmation of Predicted Specificity**

We compared the discovered motifs to those in the literature using an automated method, and selected the regulator for which the discrepancy was the greatest, Cin5 (Supplementary Table 2). The discovered motif, TTAcrTAA, contains a one base insertion compared to the previously reported site[23], TTACTAA. The previously known site is poorly enriched in the probes bound by Cin5 ($P \leq 0.02$), while the discovered motif is very strongly enriched ($P \leq 10^{-38.4}$).

We used a gel-shift assay to test whether the specificity for Cin5 that we inferred from our in vivo data also represented the in vitro properties for this regulator (Supplementary Figure 3). The DNA-binding domain of Cin5 was cloned into a derivative of the pET-32 vector (Novagen) fused to thioredoxin and a poly-histidine peptide, expressed in *E. coli*, and purified by affinity chromatography. Protein was incubated with a Cy5-labeled oligonucleotide containing the sequence gcgacaTTACCTAAgggc and challenged with unlabeled competitor containing either the same sequence or the previously published binding site (gcgacaTTACTAAagggc[23]). The reactions were analyzed on 10% acrylamide gels run in 0.5x TBE. Similar results were obtained for a probe containing the core sequence of TTACGTAA.

Bibliography


1. Lee, T. I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804. (2002).
2. Mewes, H. W., Albermann, K., Heumann, K., Liebl, S. & Pfeiffer, F. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res* **25**, 28-30 (1997).
3. Hodges, P. E., McKee, A. H., Davis, B. P., Payne, W. E. & Garrels, J. I. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res* **27**, 69-73 (1999).
4. Costanzo, M. C. et al. YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res* **29**, 75-9 (2001).
5. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-9. (2000).
6. Simon, I. et al. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697-708. (2001).
7. Aparicio, O. M. *Current Protocols in Molecular Biology* (ed. al., F. M. A. e.) (John Wiley and Sons, New York, 1999).
8. Orlando, V. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* **25**, 99-104 (2000).
9. Tessier, D. et al. *A DNA Microarrays Fabrication Strategy for Research Laboratories.* (eds. Rehm, H. & Reed, G.) (Wiley-VCH, Weinheim, Germany, 2002).
10. Hughes, T. R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-26 (2000).
11. Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374-8 (2003).
12. Csank, C. et al. Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol* **350**, 347-73 (2002).
13. Zhu, J. & Zhang, M. Q. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* **15**, 607-11 (1999).
14. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**, 939-45 (1998).
15. Bailey, T. L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* **3**, 21-9 (1995).

16. Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* **20**, 835-9 (2002).
17. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-54 (2003).
18. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
20. Dwight, S. S. et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res* **30**, 69-72 (2002).
21. Clarke, N. D. & Granek, J. A. Rank order metrics for quantifying the association of sequence features with gene regulation. *Bioinformatics* **19**, 212-8 (2003).
22. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of Statistical Learning; Data mining, inference and prediction* (Springer-Verlag, New York, 2001).
23. Fernandes, L., Rodrigues-Pousada, C. & Struhl, K. Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. *Mol Cell Biol* **17**, 6982-93 (1997).