# SUPPLEMENTAL TABLES, FIGURES AND ANALYSIS

## EXPERIMENTAL METHODS

### *S. cerevisiae* Strains

All strains were generated specifically for this study starting from an *ADE2* strain in the W303 strain background (*trp1 leu2 ura3 his3 can1 GAL+ psi+*), except for Sko1-HA (which was a gift from the Struhl lab, also in a W303 background), and are listed in Table S5. Gene deletions were introduced by transformation with PCR products including auxotrophic or antibiotic resistance markers, flanked by the 40 bp sequence found directly upstream and downstream from the gene, and selection on the appropriate medium[1]. Strains with multiple gene knockouts (KOs) were constructed by mating the strains with single deletions and dissecting the resulting tetrads. Epitope-tagged strains (GFP, HA or TAP) were constructed in a similar way[2] but the primers directed the recombination to the sites directly upstream and directly downstream from the appropriate stop codon. PCR was used to confirm the location of the marker gene or epitope tag insertion.

### Expression Microarrays

mRNA was extracted from frozen cells in two steps. First, cells were lysed in 65 ºC phenol/SDS by vortexing and the total RNA was purified by phenol/choloroform and then chloroform extraction (http://cat.ucsf.edu/pdfs/TotalRNAIsolation.pdf). The mRNA was then separated from rRNA, tRNA and other contaminants by purification using oligo-dT cellulose (http://cat.ucsf.edu/pdfs/polyARNAIsolation.pdf). 4 μg of this polyA

RNA was converted into cDNA by reverse PCR with a 2:3 ratio of amino allyl-UTP:dTTP, and purified using a DNA Clean and Concentrator-5 kit (Zymo Research). These samples were then labeled with NHS-ester Cy3 or Cy5 (GE Biosciences) by incubation in sodium bicarbonate for 2-6 hours, and purified away from the free dye using a DNA Clean and Concentrator-5 kit.

The expression difference between strains/conditions was determined by two-color microarray. 200 ng of Cy3 and 200 ng of Cy5 labeled cDNA were hybridized to an Agilent microarray with 6200 60 base probes (G4140A arrays), in Agilent hybridization buffer for 17 hrs at 65 ºC, rotating at approximately 5 rpm in SureHyb chambers (Agilent). These arrays were then washed in 6.7X SSPE buffer for at least 1 min, 0.67X SSPE buffer for 30 sec and then Agilent drying and ozone protection wash for 30 sec (1X SSPE = 0.15M NaCl, 10 mM sodium phosphate, 1 mM EDTA, pH 7.4) at room temperature. The arrays were scanned immediately using an Axon 4000B scanner. The average intensity of the Cy3 and Cy5 fluorescence at each spot was then extracted using GenePix software (version 5 or 6, Molecular Devices) and the data loaded into the Rosetta Resolver database (version 6) where a LOWESS-like normalization procedure was applied to correct for the observed non-linear dependence of the Cy5/Cy3 ratio on total signal level (we confirmed that this algorithm accurately eliminated the dye bias by performing several dye-swap experiments). Data was then exported from this database with a gene filter requiring that the signal in at least one channel is 1.5-fold above background. The background signal was set equal to the average value of a negative

control probe, spotted at 89 positions on the array, that shares no significant homology to any sequence in yeast.

In preliminary studies carried out using home printed arrays, we examined the time-course of transcriptional response to 0.375 M KCl in the wild-type, *hog1*Δ, *msn2*Δ*msn4*Δ and *sko1*Δ strains (data not shown). Here we found that the peak of the transcriptional response was at 20-40 min, in agreement with previous studies[3]. Examining the expression of the mutant strains we found an increase in the expression of Sko1-dependent genes in the *msn2*Δ*msn4*Δ strain, and an increase in the expression of Msn2/4-dependent genes in the *sko1*Δ strain, at 40 min time-points and later. This is likely due to a defect in re-establishing the osmotic balance in these strains. We therefore focused our further analysis on the 20 min time-point to avoid measuring these secondary effects.

## EXPRESSION COMPONENT ANALYSIS

### Mutant Cycle Approach

To estimate the contribution of each network component (Hog1 and TFs) to the expression level of individual genes, we developed the Mutant Cycle approach, where we co-analyze the expression data from several mutant strains. As described in Fig. 2a, to dissect the interaction between Hog1 and Msn2/4, we compared the gene expression of four strains: wild-type (wt), *hog1*Δ, *msn2/4*Δ, and *hog1*Δ*msn2/4*Δ, using DNA microarrays (measured in triplicate):

B = wt vs *hog1Δmsn2/4Δ*

C = wt vs *hog1Δ*

D = wt vs *msn2/4Δ*

E = *msn2/4Δ* vs *hog1Δmsn2/4Δ*

F = *hog1Δ* vs *hog1Δmsn2/4Δ*

For each gene, we described these measurements as the (noisy) sum of three underlying components: H (the influence of Hog1 alone on expression), M (the influence of Msn2/4 alone on expression), and Co (the effect of the interaction between Hog1 and Msn2/4), as described in the text. This allows us to rewrite the equations above as:

B = H+M+Co (as all three components exist in the wild-type strain, and are absent in the double-deletion strain)

C = H+Co (wild-type strain contains all three components, while *hog1Δ* contains only the M component)

D = M+Co

E = H

F = M

This system of equations can be formulated as the following matrix multiplication:

(Matrix 1)

$$\begin{bmatrix} \text{wt vs } hog1\Delta msn2/4\Delta \\ \text{wt vs } hog1\Delta \\ \text{wt vs } msn2/4\Delta \\ msn2/4\Delta \text{ vs } hog1\Delta msn2/4\Delta \\ hog1\Delta \text{ vs } hog1\Delta msn2/4\Delta \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \text{ x } \begin{bmatrix} \text{H} \\ \text{M} \\ \text{Co} \end{bmatrix} + \begin{bmatrix} \varepsilon_{\text{wt vs } hog1\Delta msn2/4\Delta} \\ \varepsilon_{\text{wt vs } hog1} \\ \varepsilon_{\text{wt vs } msn2/4\Delta} \\ \varepsilon_{msn2/4\Delta \text{ vs } hog1\Delta msn2/4\Delta} \\ \varepsilon_{hog1\Delta \text{ vs } hog1\Delta msn2/4\Delta} \end{bmatrix}$$

or $\boldsymbol{Y = X * \beta + \varepsilon}$, where Y are the measured values, X is the design matrix, β is the contribution of the three components, and ε is the noise. For each gene, we wish to find a β which minimizes the errors, ε.

To solve this linear model, we applied a multiple linear regression algorithm which minimizes the least squares fit of $\boldsymbol{X*\beta}$, assuming a zero-mean Normal distribution of the errors $\boldsymbol{\varepsilon}$. Specifically, the equation above $\boldsymbol{X * \beta = Y}$ is multiplied (from the left) by $X^T$, to get: $\boldsymbol{X^T * X * \beta = X^T * Y}$. In our case, the matrix $\boldsymbol{X^T * X}$ is non-singular, and so we invert

$X^T * X$ and use it to multiply the equation (from left), and obtain a unique solution for the vector of regression coefficient $\beta = (X^T * X)^{-1} * X^T * Y$.

It is assumed that all the coefficients in $\beta$ have a zero-centered normal distribution, and so we can estimate their variance and covariance values. Specifically, $Cov(\beta) = \sigma^2 * (X^T * X)^{-1}$, where $\sigma^2$ is the variance of the fit. These properties pave the way for testing hypotheses about the estimated values of regression coefficients $\beta$.

It should be noted that since $Y$ was actually measured in triplicate, we can concatenate the 3 sets of values so that $n=|Y|= 15$. We also replicate the design matrix $X$ to match. This allows for more accurate regression, by estimating the error in each array separately. Calculations were performed based on the REGRESS function of MATLAB, (version 7.0 R14), and following [4].

The same approach was applied to dissect the pair-wise interactions between Sko1, Hot1 and Msn2/4 (See Fig. 3c and Fig. S6a). Specifically, we determined the values of three components (SH for the Sko1Hot1 effect, M for the Msn2/4 effect, and SHM for the effect of their interaction) by comparing gene expression in the wt strain, *msn2/4Δ*, *sko1Δhot1Δ* and *sko1Δhot1Δmsn2/4Δ,* as shown in Fig. S6 using the matrix below.

(Matrix 2)

$$\begin{bmatrix} \text{wt vs } sko1\Delta hot1\Delta \\ \text{wt vs } msn2/4\Delta \\ msn2/4\Delta \text{ vs } sko1\Delta hot1\Delta msn2/4\Delta \\ sko1\Delta hot1\Delta \text{ vs } sko1\Delta hot1\Delta msn2/4\Delta \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} x \begin{bmatrix} \text{SH} \\ \text{M} \\ \text{SHM} \end{bmatrix}$$

In addition, we extended the Mutant Cycle approach to examine the three-way interaction between Sko1, Hot1 and Msn2/4. For this we used the following experiments:

wt vs *msn2/4Δ*

wt vs *sko1Δhot1Δ*

*sko1Δhot1Δ* vs *sko1Δhot1Δmsn2/4Δ*

*msn2/4Δ* vs *sko1Δhot1Δmsn2/4Δ*

wt vs *hot1Δ*

*hot1Δ* vs *hot1Δmsn2/4Δ*

*msn2/4Δ* vs *hot1Δmsn2/4Δ*

wt vs *sko1Δ*

*sko1Δ* vs *sko1Δmsn2/4Δ*

*msn2/4Δ* vs *sko1Δmsn2/4Δ*

Here, we decompose these measurements into the sum of ten components, reflecting the effect of each factor: Sko1, Hot1, Msn2/4, and each combination of two or three factors: Sko1Hot1, Sko1Msn24, Hot1Msn24, and Sko1Hot1Msn2/4.

As before, we formulate the measurements as a noisy matrix multiplication:

(Matrix 3)

$$\begin{bmatrix} \text{wt vs } msn2/4\Delta \\ \text{wt vs } sko1\Delta hot1\Delta \\ sko1\Delta hot1\Delta \text{ vs } sko1\Delta hot1\Delta msn2/4\Delta \\ msn2/4\Delta \text{ vs } sko1\Delta hot1\Delta msn2/4\Delta \\ \text{wt vs } hot1\Delta \\ hot1\Delta \text{ vs } hot1\Delta msn2/4\Delta \\ msn2/4\Delta \text{ vs } hot1\Delta msn2/4\Delta \\ \text{wt vs } sko1\Delta \\ sko1\Delta \text{ vs } sko1\Delta msn2/4\Delta \\ msn2/4\Delta \text{ vs } sko1\Delta msn2/4\Delta \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} x \begin{bmatrix} \text{Sko1} \\ \text{Hot1} \\ \text{Msn2/4} \\ \text{Sko1Hot1} \\ \text{Sko1Msn2/4} \\ \text{Hot1Msn2/4} \\ \text{Sko1Hot1Msn2/4} \end{bmatrix}$$

and find a β which minimizes the system errors.

**Statistical significance**

Here we test the hypothesis $\boldsymbol{H_0}$ that one of the regression coefficients of $\boldsymbol{\beta}$, denoted $\boldsymbol{\beta_j}$, equals zero. Under $\boldsymbol{H_0}$, $\boldsymbol{\beta_j / std(\beta_j)}$ should have a ***t***-distribution with ***(n-p)*** degrees of freedom, where ***n*** is the number of experiments in the cycle (15 in the case of Hog1Msn2/4), and ***p*** is the size of $\boldsymbol{\beta}$ (3 in the case of Hog1Msn2/4). By computing the cumulative distribution function of the ***t*** distribution, we can estimate the likelihood of $\boldsymbol{\beta_i}$ under $\boldsymbol{H_0}$. This estimation approximates of probability of seeing such a value (or larger) at random, and we will treat it as a "zero p-value" for $\boldsymbol{\beta_j}$ (ref 4).

The same rationale can be applied to testing additional critical values. For example, we defined a threshold ***thr*** for a non-marginal contribution of a factor (above 1.5-fold), and used a similar approach to test if the contribution of some factor is significantly above the threshold (for ***H*** and ***M***, which are assumed to be positive) or non-marginal (two-tailed version, for ***Co***, which could be either positive or negative). In this case, we assume that under the null assumption $\boldsymbol{H_0}$, the mean value of $\boldsymbol{\beta_j}$ is smaller than ***thr***, and so $\boldsymbol{(\beta_j - thr) / std(\beta_j)}$ should have a t-distribution with ***(n-p)*** degrees of freedom.

**Accuracy of the Approach**

To ensure that the expression components fully and accurately account for the raw microarray data we compared the data predicted from the fitted component values back to the raw data used to calculate these values (Figs. S1-S3). These plots demonstrate that the expression components determined in the global fit to the array data accurately and completely describe the expression changes found in the individual mutant strains.

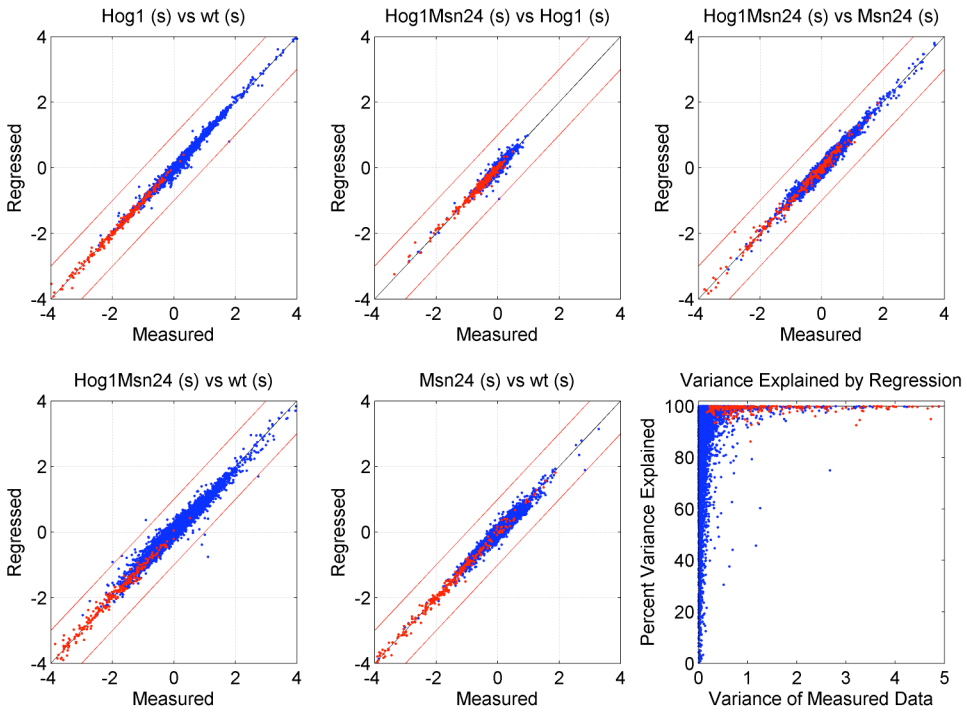



**Fig. S1** Analysis of the fit to the data in Fig. 2. The data from each microarray experiment B-F in Fig. 2a is plotted against the data expected based on the component values extracted from the global fit to these arrays (regressed values). Each point shows the measured and expected $\log_2$ fold-change for a single gene, colored red if it is in the 273 gene Hog1 network and blue if it is outside the network. The last panel shows the percent of total variance explained by the fit (color coded as above). For each gene, we computed the variance (V) of expression for the expression measurements (B-F). The percent of variance explained by the regression is given by 100*(V-R)/V, where R is the variance of the residual data (measured minus regressed data).

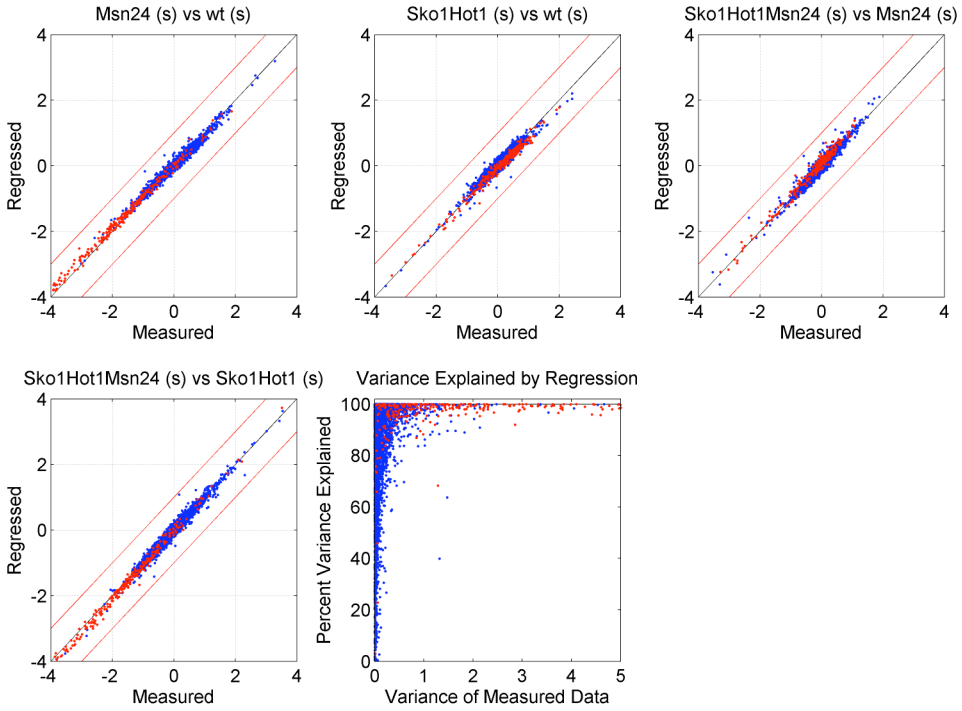



**Fig. S2** Analysis of the fit to data from the cycle in Fig. 3c. Comparison of measured and expected data, as described in Fig. S2, except that the arrays and fitted components are for Fig. 3c.

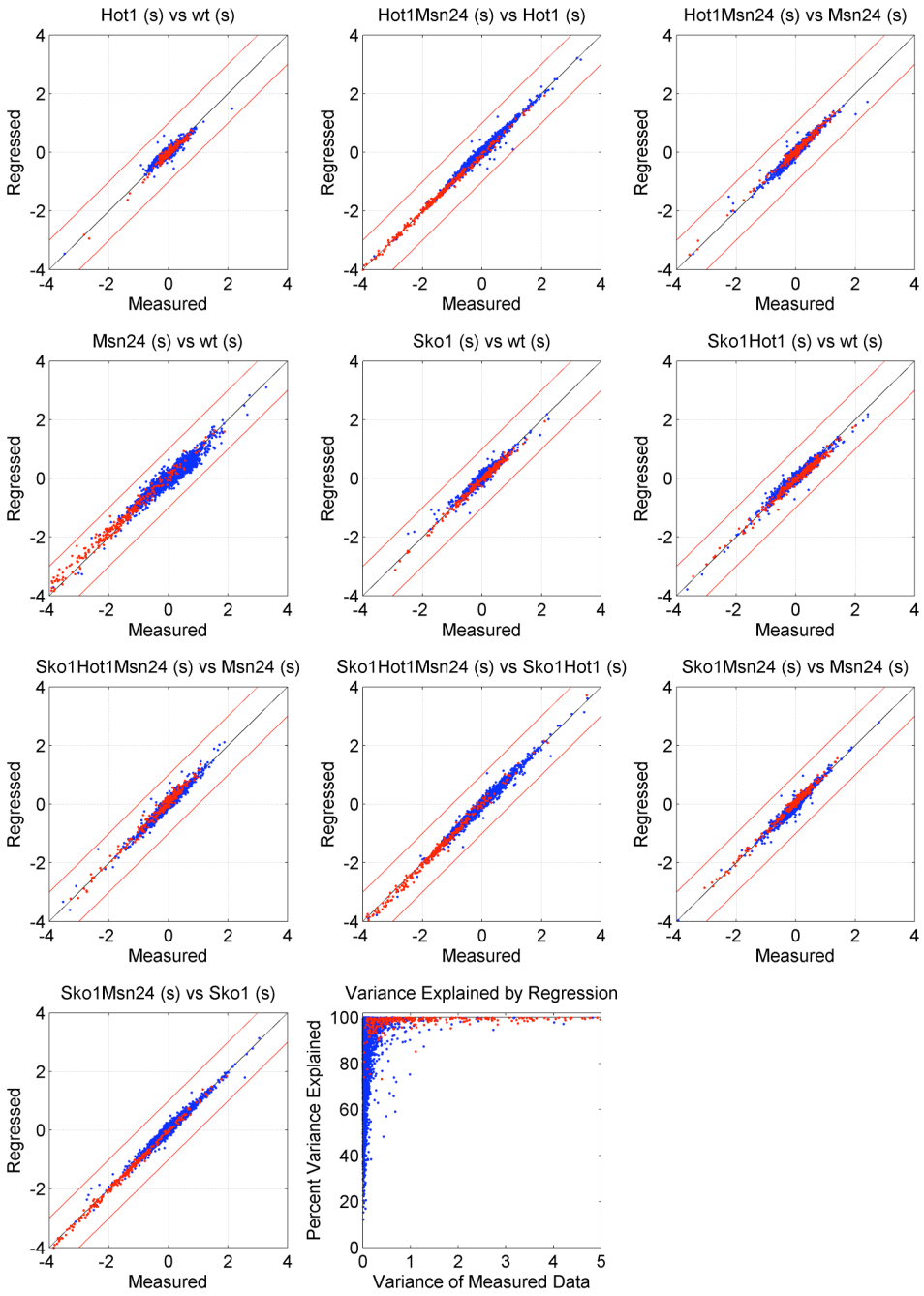



**Fig. S3** Analysis of the fit to a Sko1-Hot1-Msn2/4 linked cycles (Fig. S6). Comparison of measured and expected data, as described in Fig. S2.

**Comparison of Protein Stability and Microarray Mutant Cycles**

Quantitative epistasis analysis has been used previously to dissect interactions within, or between, proteins[5]. For example, to probe the interaction (int) between two amino acids (A and B) in a given protein, the stability (ΔG) of the wt, mutant in A, mutant in B, and mutant in AB protein are measured and the values compared. This analysis takes the form $\Delta G_{int} = (\Delta G_{wt} - \Delta G_A) - (\Delta G_B - \Delta G_{AB}) = (\Delta G_{wt} - \Delta G_B) - (\Delta G_A - \Delta G_{AB})$. Note here, that while there are two distinct ways to calculate $\Delta G_{int}$, the values of the wt and all three mutants are used in both calculations. Thus, either path through the mutant cycle will, by definition, will give the same value and error. The situation is different in the mutant cycle approach introduced here since each two-color microarray reports on the difference between two strains (a ΔΔG by analogy). Following the cycle in Fig. 2a we can see that Co = (Array C - Array E) = (Array D - Array F), and thus there are two distinct ways (C-E and D-F) to calculate the interaction or what we call the cooperative component (Co). Therefore, in the case of the gene expression mutant cycles it is most accurate to use a global fit (as described in the previous section) so that Arrays C-F all help to constrain the value of Co. This is particularly important as the errors in microarrays are substantial (as high as 2-fold) and are compounded in comparative analysis. Similar arguments hold for calculating the values of H and M as well, and thus we use the fitting procedure listed above to solve the value of the $\boldsymbol{\beta}$, or **[H M Co]**, vector.

Since the arrays in Fig. 2a form a closed cycle it is also possible to solve for the values of H, M and Co using the data from any three of the arrays C through F. Again, however, this approach would be less accurate than that described above, even if one of these arrays was measured twice. As the errors for microarrays are log-normal (e.g.

multiplicative), analyzing the three arrays with the smallest sum value will give the smallest errors in the fit. The three arrays that have the smallest sum will vary from gene to gene and thus we measure all four arrays C-F to accurately calculate the component values for all genes.

**Threshold Effects**

Msn2/4 is activated to a lower level in the absence of Hog1 than in the wt strain (Figs. 2c and 3a). Therefore, if the threshold for activation by Msn2/4 varies from gene to gene, so should the impact that Hog1 has on Msn2/4 dependent gene activation. This is precisely what we find. Among the genes in Class II (Fig. 3b), we find three distinct groups of genes (Groups 1-3, Fig. 2c). The first group of genes (Group 1) appears to have a high threshold for activation as they absolutely require Hog1 for Msn2/4 dependent gene activation (Fig. S4, top). By contrast, the genes in Group 2 appear to have a lower threshold for activation by Msn2/4 as they are partially activated even in the absence of Hog1 (Fig. S4, bottom). Finally, genes in Group 3 appear have the lowest threshold for activation by Msn2/4, as they are fully activated even in the absence of Hog1 (M but not Co component). Similar logic applies to Class III, Groups 5-7, except here genes are also induced by Hog1 through Sko1 and Hot1 and thus also have an H component. Group 8 is different in that the induction from Hog1 alone and Msn2/4 alone is redundant and thus there is a negative Co component.

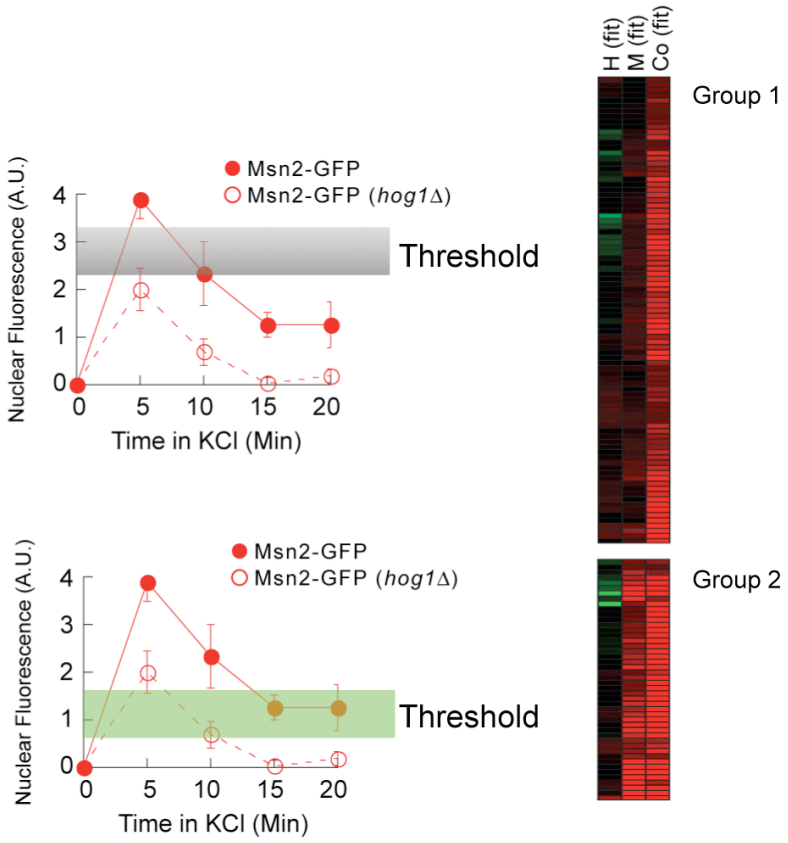



**Fig. S4** Variable threshold for activation by Msn2/4. The expression component data from Groups 1 and 2 in Fig. 2c (right) are shown next to the microscopy data for the nuclear import of Msn2/4 from Fig. 3a (left). Genes in Group 1 appear to have a high threshold for activation by Msn2/4 (grey bar, upper left panel) as there is little or no induction at the Msn2/4 level found in the absence of Hog1 (no M component). By contrast the genes in Group 2 all show significant induction by Msn2/4 in the absence of Hog1 (significant M component) and thus appear to have a lower threshold for gene activation (green bar, lower left panel).

**Hog1 Activates Sko1 and Hot1**

Hog1 is known to phosphorylate, activate and/or bind to the TFs, Msn1, Smp1, Sko1, Hot1 and Cin5, but few target genes have been identified for these factors[6-9]. By analyzing strains lacking one or more of these factors, we found that only Sko1 and Hot1 play a significant role in the osmotic stress response (Table S1 and Fig. S5). At one group of genes (23 genes in the Hog1 network, >2-fold expression change, Fig. S5) Sko1 acts as a repressor in the absence of stress (log growth phase, YEPD medium) and is either inactive (8/23, Sko1 de-repression, Fig. S5) or switches to an activator in osmotic stress (11/23, Sko1 de-repression and activation, Fig. S5). This repressor/activator switch was found previously for Sko1 at a few of the genes in this group (most notably *GRE2,* ref 8,10). Surprisingly, we also discovered a new class of genes (18 in the Hog1 network, >2-fold expression change, Fig. S5) where Sko1 acts as an activator in salt stress, but does not influence expression in pre-stress conditions. We therefore find that Sko1 acts in three distinct ways: (i) as a repressor in pre-stress conditions, (ii) as a repressor in pre-stress conditions that switches to an activator in post-stress conditions, and (iii) as an activator in post-stress conditions; and not in a single mode (repressor/activator) as suggested previously. Moreover, Sko1 activates many more genes than previously appreciated. Finally, we find a small group of genes that are activated by Hot1 in salt stress (6 genes in the Hog1 network, Fig. S5), many of which were known from previous studies (most notably *STL1,* ref 11,12). We observe that Hog1 activity is required for stress dependent activation of Sko1 and Hot1 (compare columns 1 and 2 to column 3 in KCl, Fig. S5) and derepression of Sko1 (compare column 3 in YEPD to column 3 in KCl, Fig. S5).

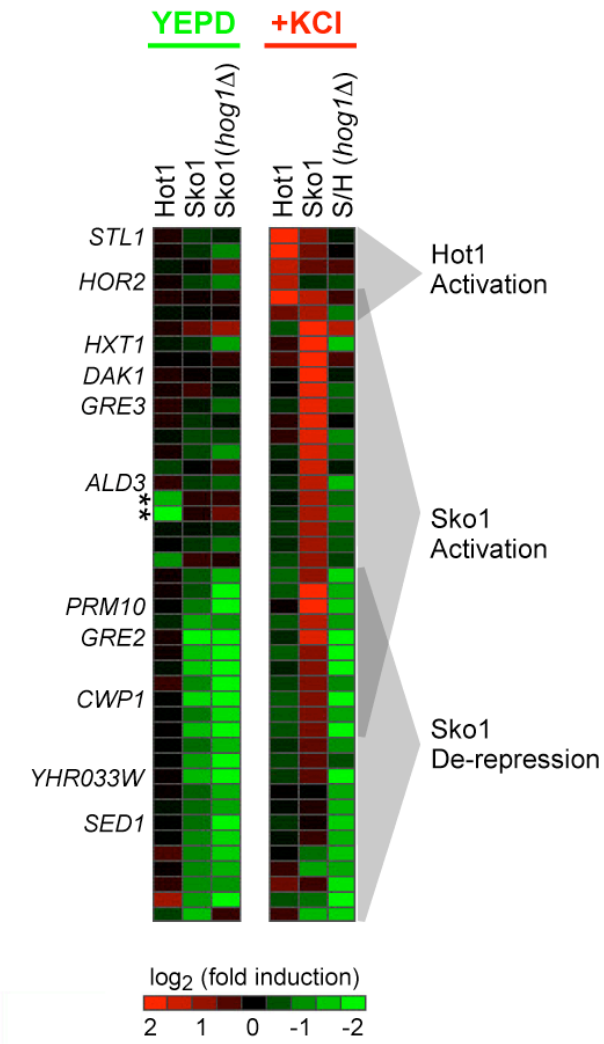



**Fig. S5** Sko1 and Hot1 are Hog1 dependent gene activators. The influence that Sko1 and Hot1 have on gene expression was determined by comparing the gene expression levels in *sko1*Δ and *hot1*Δ to the wild-type strain in both pre and post stress conditions (YEPD and YEPD + 0.4M KCl, repectively). To examine the role of Hog1 in Sko1 de-repression/activation and Hot1 activation we also compared the expression in *sko1*Δ*hog1*Δ to *hog1*Δ in YEPD (Sko1(*hog1*Δ)) and *sko1*Δ*hot1*Δ*hog1*Δ to *hog1*Δ in

YEPD + 0.4M KCl (S/H(*hog1*Δ)). These data show that Sko1 and Hot1 are trapped in their repressing and inactive states, respectively, in stress without Hog1 activity. The data shown is the average from three repeat experiments, and only includes genes that are in the Hog1 network (273 gene is Fig. 2) and are 2-fold induced or repressed by Sko1 or Hot1. The data for six genes (with >2-fold change) that appear to be repressed by Sko1 in KCl (*YPR160W, YNR034W-A, HSP26, UIP4, PDR15* and *YNR014W;* 2.6-fold average change) are left out of the figure. These genes are activated by Msn2/4 and the apparent repression by Sko1 depends on the presence of these TFs (Table S1). It is therefore likely that the repressing influence of Sko1 on these genes (and in salt in general) is indirect and is actually due to increased Msn2/4 activation in the *sko1*Δ strain. *These genes appear to be repressed by Hot1 but are two overlapping genes (*YMR173W* and *YMR173W-A*) that neighbor Hot1 (*YMR172W*) and were likely disrupted by the replacement of Hot1 with *HIS3*. The data for these arrays are listed in Table S1.

**Sko1-Hot1-Msn2/4 Interactions**

To determine the exact role that Sko1 and Hot1 play in the Hog1-dependent stress response, and quantify the logic gates at the promoters of Hog1 dependent genes, we used the mutant cycle approach to examine the influence of, and interaction between, Sko1/Hot1 (together) and Msn2/4 (see red cycle and components, Fig. S6a). Interestingly, we find few positively cooperative (AND) interactions between these factors (Fig. S6b). In fact, the number of genes with Co(Sko1/Hot1-Msn2/4)>1.5-fold is similar to the number expected by chance (5 observed *versus* 2 false positives expected, and 9 observed *versus* 9 false positives expected, at $p<0.01$ and $p<0.05$, respectively, see Fig. S6 legend for details). By contrast, we find a significant number of negatively cooperative (OR) interactions at $p<0.01$ (13 *versus* 2 false positives expected, with no increase in the number of true positives at $p<0.05$). At these genes, Sko1 and Hot1 have little or no influence on expression in the presence of Msn2/4 (and were therefore not identified in the simple analysis of Fig. S5), but induce transcription up to 100-fold in the absence of Msn2/4 (Fig. S6b, lower panel). We also find a further 10 genes that are activated by both Sko1/Hot1 and Msn2/4 ($p<0.01$), but without significant cooperativity ($p>0.01$). At these genes the promoters act primarily as a SUM gate in log terms (some also have low level negative cooperativity).

To examine the influence that Sko1, Hot1 and Msn2/4 have on gene expression individually, and quantify the interaction between Sko1-Msn2/4, Hot1-Msn2/4, Sko1-Hot1 and Sko1-Hot1-Msn2/4, we extended our mutant cycle approach to look at three-way interactions (black and red cycles and black components, Fig. S6a). This allowed us

to fully dissect the interactions at genes with SUM (additive in log space) or OR gate logic (Fig. S6b and Table S2) and accurately measure the influence of Hot1 and Sko1 (separately), even where they act redundantly with Msn2/4 (Table S2 and S3). We then used these component values to expand our initial Hog1-Msn2/4 network model (Fig. 3b) and assess its accuracy and completeness.

In our initial model (Fig. 3b and top of Fig. S6c), Hog1 activity is split into two mechanisms, Hog1 activation through Msn2/4 (Co component) and Hog1 activation that is independent of Msn2/4 (H component). Our further analysis revealed that Hog1 activates genes through at least three Msn2/4-independent mechanisms: Sko1 derepression, Sko1 activation, and Hot1 activation (Fig. S5). Therefore, if our model of Hog1-dependent gene activation is correct, the H component should be the sum of the activation from Sko1 and Hot1 (both in the absence of Msn2/4). Conversely, there should not be significant activation (or derepression) from Sko1 and Hot1 at genes where Hog1 acts exclusively through Msn2/4. Indeed, there is excellent agreement between the genes activated by Sko1 and/or Hot1 and the genes with a significant H component (Fig. S6c; compare the influence of Sko1 and Hot1 in Groups I and II, to that in III). Moreover, there is a strong correlation ($R = 0.90$) between the level of gene induction calculated by summing the log influence of Sko1 derepression and Sko1/Hot1 activation (in the absence of Msn2/4) and that of the fitted H component (Fig. 3d). This consistency between the two levels of the model (signaling and promoter) demonstrates the accuracy of the mutant cycle approach taken to identify the Hog1, Sko1, Hot1 and Msn2/4 components.

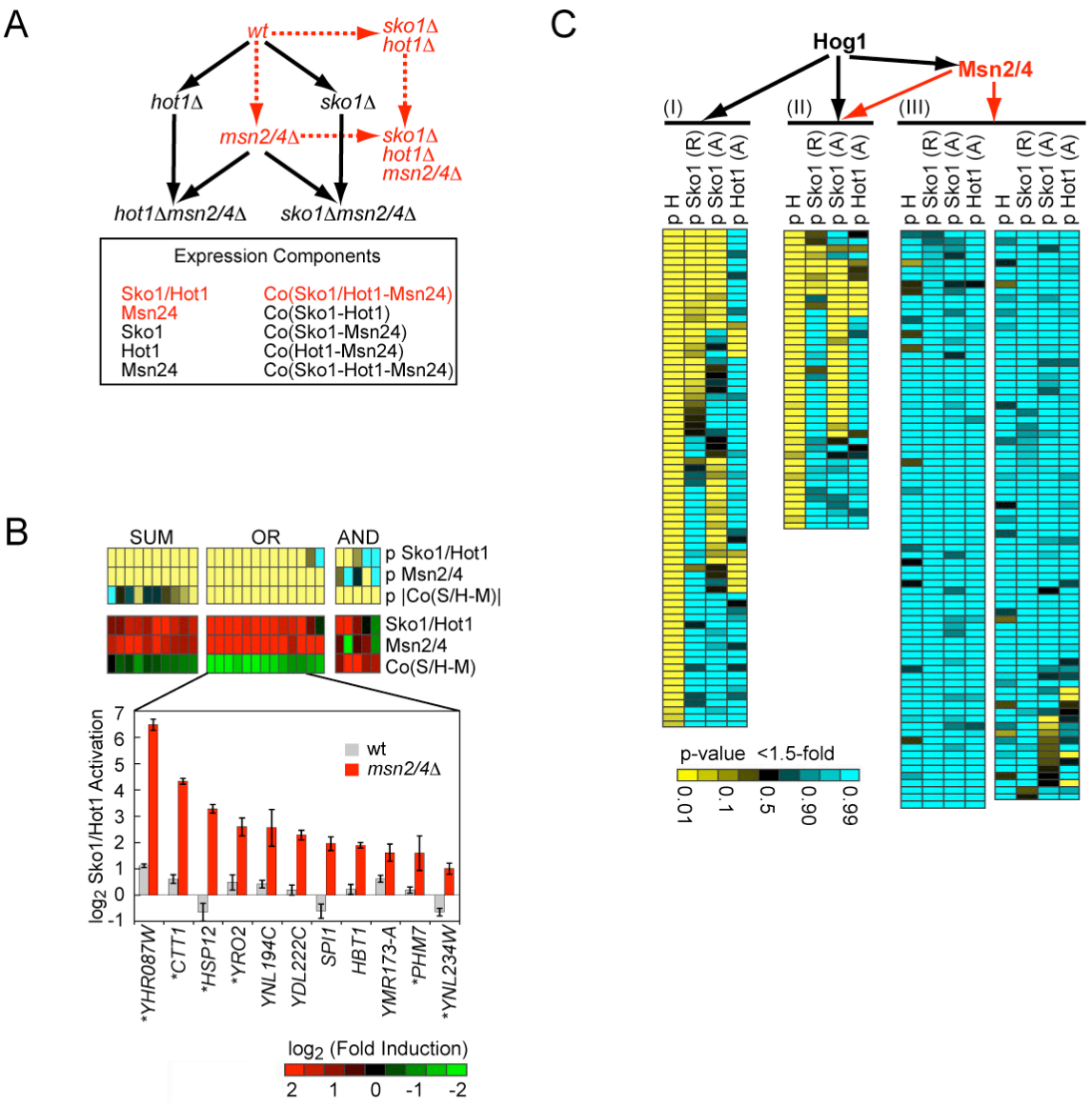



**Fig. S6** Role of Sko1 and Hot1 in Hog1 dependent gene activation. **(A)** Schema describing the experiments and equations used to break the influence of Sko1, Hot1 and Msn2/4 into components. Each arrow represents a single microarray (measured in triplicate) comparing gene expression in two strains. The terms listed below the diagram are the components that were extracted by fitting the cycle data. **(B)** Interaction between Sko1/Hot1 and Msn2/4. Heat map showing the best-fit value of the expression

components, and their statistical significance, for genes with a significant influence from both Sko1/Hot1 and Msn2/4, $p<0.01$. The bar graph shows the raw data for the *sko1*Δ*hot1*Δ/wt and *sko1*Δ*hot1*Δ*msn2*Δ*msn4*Δ/*msn2*Δ*msn4*Δ data for 11/13 OR gate genes where Sko1/Hot1 activity is redundant in the wild-type strain. Gene names highlighted by a star are activated by both Sko1 and Hot1 (in some cases redundantly), other genes are just activated by Sko1. **(C)** Correlation between the H component and Sko1 and Hot1 activity in the absence of Msn2/4. The heat map shows the significance of Sko1 repression (Sko1 R), Sko1 activation (Sko1 A) and Hot1 activation (Hot1 A), in the absence of Msn2/4, from the fit to all the cycles in (A) and the Hog1 component (H) from Fig. 2. The genes are organized into groups based on the mechanism of activation defined in Fig. 3b.

## ChIP Analysis

### Genome-wide ChIP

To validate our network model, and gain insight into its structure, we used genome-wide ChIP analysis[13,14] to identify the DNA binding sites for Sko1 and Hot1 *in vivo*, both before and after the addition of 0.4 M KCl to the medium (we were unable to collect high-quality ChIP data for Msn2/4, see Methods). The binding sites for Sko1 and Hot1 are clearly distinguished by peaks of enrichment that extend across as many as eight consecutive probes on the 42,000-probe tiling array (Fig. 4a). Fitting the ChIP data using a peak-shape model (inset Fig. 4a, and section on data fitting below) allowed us to determine the position and significance of each TF binding site with high accuracy and dramatically reduced the number of false positives from noisy probes (see Methods). This fitting approach revealed that most Sko1 and Hot1 binding sites are in promoter regions (80% of the 100 peaks with an enrichment ratio >5 in KCl, Table S3). Focusing our analysis on promoter regions, we identified all of the binding sites that are above the background noise and estimated their statistical significance based on the conservative assumption that all binding to the ~6000 genes outside of the Hog1 network is spurious (where the Hog1 network is defined as the 273 genes in Fig. 2c). To do this, the peaks from each ChIP experiment were ranked according to their fitted height and then the significance (p-value) of each binding site was set equal to the fraction of genes outside the network assigned to a higher-ranking peak.

Based on this analysis we find excellent agreement between the Sko1/Hot1 target genes identified through expression ($p<0.05$ in Groups I and III of Fig. 3c) and ChIP: 65-80% of the genes repressed by Sko1 (27 total), activated by Sko1 (52 total), or activated by Hot1 (15 total) are bound by the factor in the appropriate condition at $p<0.05$ (Fig. 4b). There is also higher than expected binding of Sko1 and Hot1 at other genes within the Hog1 network. In fact, we find 42 additional Sko1 target genes and 23 additional Hot1 target genes (accounting for the expected number of false positives), within the 273 gene network (Figs. 4c and d). While some of these genes are regulated by Sko1 or Hot1 but just missed the $p<0.05$ cut-off, most of these binding sites are "silent" (24/42 Sko1 and 17/23 Hot1) as >1.5-fold activation or repression (redundant or otherwise) is unlikely (defined by $p>0.80$).

The ChIP data also provides important insight into the function of Sko1. We find three classes of Sko1 binding behavior within the Hog1 transcriptional network (Fig. 4c). In the first small class (6 genes total), Sko1 binds to the promoter in pre-stress conditions (YEPD), but is released within 5 min of KCl stress (*e.g. FSH1* and *HXT4*, Fig. 4a). In the second class (45 genes total) Sko1 is constitutively bound to the promoter (*e.g. YHR033W*, *HXT1* and *HXT5*, Fig. 4a). In the third class (37 genes), Sko1 is only recruited to the promoter after stress treatment (*e.g. YHR087W*, Fig. 4a). This variable Sko1 binding behavior is functionally important since Sko1 acts as a repressor in YEPD and an activator in KCl. Indeed, genes that are only bound by Sko1 in KCl are activated, but not repressed, by this factor (top graph Fig. 4c). By contrast, the genes constitutively

bound by Sko1 are repressed, repressed and activated, or just activated (bottom graph, Fig. 4c).

Hot1 binding falls into two classes (Fig. 4d): constitutive (7 genes) and inducible (28 genes, e.g. *YHR087W* and *HXT1* bottom panel Fig. 4a), but here there is no clear expression difference between the groups. Sko1 binds to most of the Hot1 target genes (75%, Table S3), as seen for the two examples, *YHR087W* and *HXT1*. Therefore, Hog1 influences gene expression in five distinct ways by combining different classes of Sko1 and Hot1 binding: constitutive Sko1 binding with/without Hot1; inducible Sko1 binding with/without Hot1; and pre-stress only binding of Sko1 (see *YHR087W* and *HXT 1, 4* and *5,* in Fig. 4a, for four of these regulatory modes). The short DNA sequences that direct these binding events, as well as those for Msn2/4, were uncovered using a new motif identification pipeline (see next section). In each case we find highly significant overlap between the gene-set with a known (Sko1 and Msn2/4), or a putative (Hot1), DNA binding motif and the targets identified by the ChIP data (Sko1 and Hot1) and the mutant cycle analysis (Sko1, Hot1 and Msn2/4), (Fig. S7 and S8).

**ChIP Data Analysis**

Due to long sheared DNA fragments in the ChIP experiments, a binding event at position ***x*** results in high enrichment of surrounding probes. This effect decays as the distance between the probe and the binding position increases. In general, the probability of a probe located $\Delta_x$ bases away from the binding location is proportional to the integration over all fragment lengths (from length $\Delta_x$ or more), times the number of possible alignments of the DNA fragments that allow both the binding of the fragment by the target transcription factor and its hybridization to the reporting probe), times the relative abundance of DNA fragments of such length, denoted by ***c(l)***.

Thus, the estimate for a peak's shape is given in the following equation:

$$F(\Delta_x) \propto \int_{l=\Delta_x}^{\infty} (l - \Delta_x) c(l) dl$$

The distribution of sheared fragment lengths ***C(l)*** depends on the sonication protocol. We measured the fragment size distribution created by our protocol (described in the Methods) using an agarose gel and found a broad distribution of fragments (200-2000 bp) that is well described by a Gamma distribution. This distribution has two parameters that control the mean and standard deviation of fragment length. In subsequent experiments we used these two parameters to define the entire fragment length distribution ***c(l)***.

**The Peak Fitting Algorithm**

We have developed an iterative algorithm to identify all significant binding events that appear in the probes. Briefly, this is done by identifying stretches of enriched probes and attempting to explain (at least part of) their values using the peak model. We choose the most probable values for center position and peak height, and compute the statistical significance of this peak. If its p-value falls below 0.01, and its height exceeds 1.5, we call it a binding event, and subtract its effect (i.e. predicted enrichments for the probes in ***S***) from the actual ratios. This enables us to identify overlapping peaks one at a time (starting from the strongest one), until the remaining data cannot be distinguished from noise. Our model-based approach also allows us to naturally integrate data from different replicates, computing the likelihood of the peak based on *all* enrichment values of its probes. We now expand on the relevant steps.

**Optimization of Peak Parameters**

Once a window **S** of consecutive probes with enriched values is found, we search for optimal peak parameters to fit the enrichment rations in **S** in a two-step manner. We enumerate over the peak center point ***x*** in a 10 bases resolution and find the height ***alpha*** which minimizes the sum of squared deviations between the log (base 2) of the measured enrichments and ones predicted by peak's shape. This is done using Brent's method for one-dimensional minimization. Finally, we report the position ***x*** and height ***alpha*** whose fit was the optimal.

**Estimating the statistical significance of binding events**

The statistical significance of a binding event is estimated by computing an empirical log-likelihood ratio (*LLR) p-value***.** We compute the likelihood of the set of probes ***S*** given the null model $\boldsymbol{L_0}$, assuming that the values are normally distributed around the median enrichment ratio of the array. We then compute the likelihood of the same probes given a peak model with center ***x*** and height ***alpha***, denoted by $\boldsymbol{L_{peak}}$. We use the log-likelihood-ratio (LLR) $\boldsymbol{L_{peak}/L_0}$ to score the peak. We then compute 1000 shuffling-based LLR scores in the following way: we replace the measured enrichment values for each probe in ***S*** with a randomly chosen probe from the array, find the optimal height as described above, and calculate the log likelihood ratio for this set. Finally, we calculate the empirical LLR-based p-value of the original peak by computing the percentile of the rank of the true LLR score among the 1000 shuffling-based scores.

**Computing a Bayesian Confidence Interval around binding event**

In addition to the estimating the peak's center position and height, we also compute a 99%-confidence interval around the binding position of each peak. This is done by considering the likelihood $\boldsymbol{L_{peak}(x)}$ of the peak's probes ***S***, when centered at position ***x***. We then use Bayes' rule to compute the posterior probability of the center being at each position ***x***, and define the Bayesian Confidence Interval as the region covering 99% of the posterior probabilities.

The entire peak fitting process is sketched in the algorithm below:

**Input:**

- Genomic location of probes
- Distribution of DNA fragment lengths.
- ChIP data (enrichment ratios (IP vs input))

**Algorithm:**

- Estimate shape of a peak $F(\Delta_x)$
- Initialize enrichment ratio threshold ***T*** to 10
- Set cooling factor ***K*** to 0.99
- Let ***B*** be the set of binding events
- Begin main loop:
  1. For every consecutive set ***S*** of probes above threshold ***T***
     - **a)** Add flanking probes (up to 2.5Kb away) into set ***S***
     - **b)** Find center position ***x*** and height of peak ***alpha*** to fit **S** best.
     - **c)** Calculate likelihood-based p-value of peak
     - **d)** If peak is significant, and its estimated height is above 1.5:
       - Calculate 99% Bayesian Confidence Interval
       - Add peak into set of binding events B
       - Predict values for probes in S using B, and subtract from data
  2. Update the enrichment threshold ***T*** = ***T*** * cooling factor ***K***
  3. Repeat main loop until no new significant binding events are found

**Definition of Yeast Promoters**

Promoter regulatory sequences were defined using sequences and annotations from UCSC (sacCer1). For genuine genes (UCSC track "sgdGene"), promoters were defined

as the regions upstream to the translation start site, up to 1 Kb or up to the coding regions of upstream genes. As for pseudo and dubious genes (UCSC track "sgdOther"), we considered 500-bp upstream to the translation start site, regardless of overlapping coding regions.

**Genome-Wide Analysis of Bound Genes**

Once the genome-wide ChIP data was analyzed for peaks, we compute the enrichment of each gene by summing the heights of all significant peaks whose 99% Bayesian confidence interval overlaps its promoter region (as defined below).

Genes whose enrichment was above a threshold corresponding to a 5% false positive rate were declared as bound genes. To calculate the false positive rate for each threshold, we considered as non-targets all the genes outside of the Hog1 network (as defined in Fig. 2), apart for genes whose promoter regions overlap a true positive gene (i.e. a gene within the Hog1 network, whose enrichment is above the threshold). Those (negative) overlapping genes were discarded from the false positive rate computation.

## MOTIF DISCOVERY AND ANALYSIS

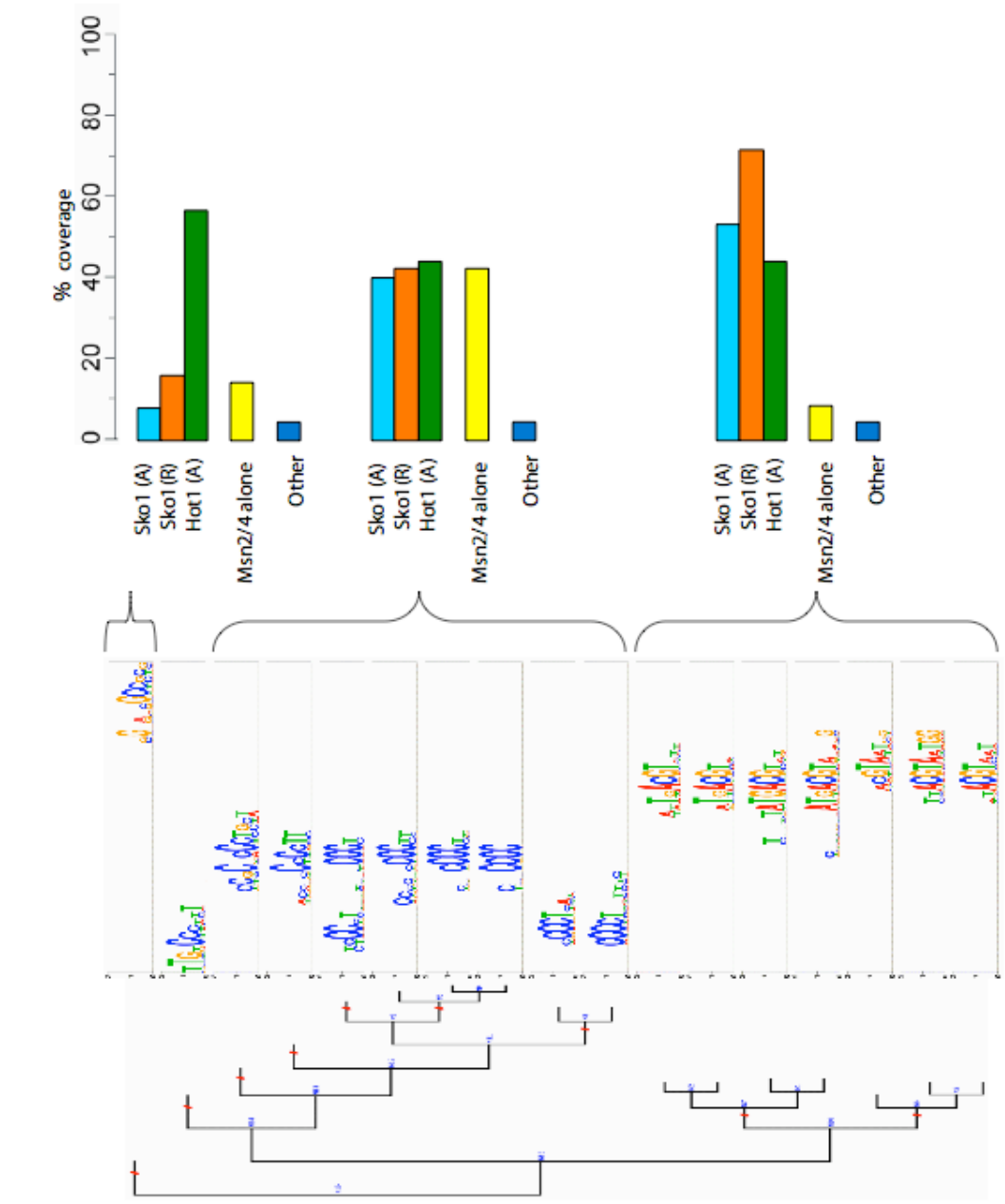
### Overview and Results

To identify the binding sites for Sko1, Hot1 and Msn2/4 we searched the promoters of the gene-sets described in Figs. 2 and 3, for short, overrepresented, DNA sequences using a new motif analysis pipeline (see Methods and ref 15). This led to the discovery of 31 motifs, clustered according to similarity using the BliC score (see Methods, left side of Fig. S7a and b). We found two distinct motifs that are significantly enriched ($p=1.3x10^{-10}$ and $1.5x10^{-9}$) in the promoters of Hot1 activated/bound genes (top, Fig. S7a and top Fig. S7b). This is the first data regarding the DNA binding site for Hot1 and thus further experiments are needed to determine whether Hot1 binds directly to one or both of these sequences. We also identified a family of motifs enriched ($p=1.6x10^{-31}$, for the combined targets of the top three motifs) in the promoters of genes regulated/bound by Sko1 (bottom group, Fig. S7a and S7b). These motifs are grossly similar to known Sko1 binding sites[8,10] and the consensus motif AT(G/T)ACGT(A/C)A, identified previously for binding in pre-stress conditions[16]. Finally, we found a family of motifs, enriched ($p=4.7x10^{-29}$, for the top motif) in the promoters of Msn2/4 target genes (middle group, Fig. S7a and b). These motifs are very similar to, or contain, the known *in vitro* binding site for Msn2/4 (CCCCT, ref 17,18).

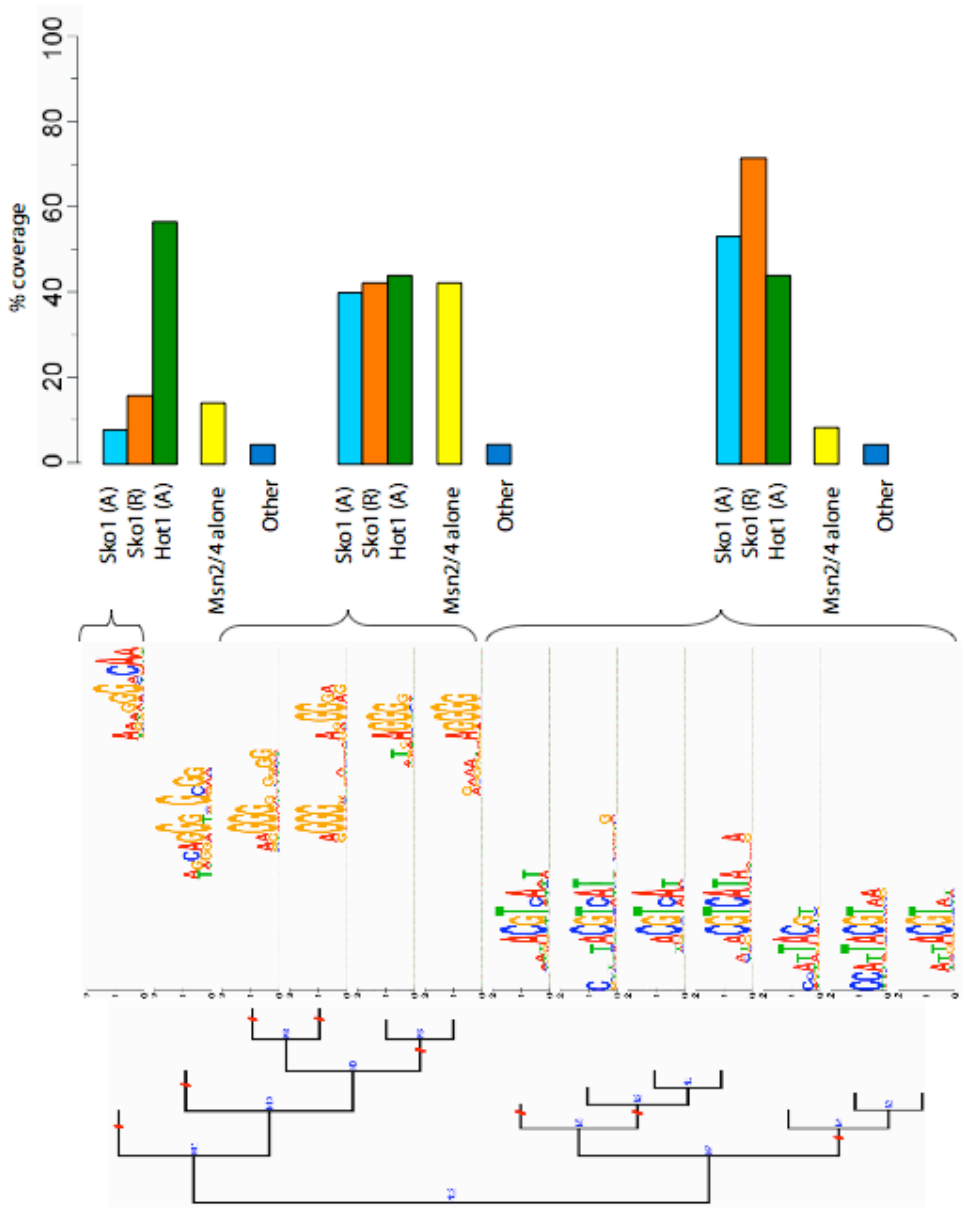
Overall, therefore, we found highly significant overlap between the gene-sets identified by motif analysis and microarray analysis (ChIP data and or expression analysis). This correlation further supports our Hog1 network model, particularly for Sko1 where the overlap between the gene-sets identified by ChIP, expression analysis and motif analysis is excellent (Fig. S8). However, our analysis also suggests that higher order promoter

structure plays an important role in at least Sko1 and Msn2/4 binding behavior since we could not find significant differences in the motifs at genes with different modes of Sko1 regulation/binding or different thresholds for activation by Msn2/4 (data not shown). Future work will therefore focus on building a more sophisticated promoter model that includes long distance interactions between transcription factor binding sites and takes into account nucleosome positions. We anticipate that this model will help us to describe the subclasses of Sko1 and Msn2/4 binding behavior, but also improve the prediction of Msn2/4 binding sites, since the short sequence (CCCCT) known to be sufficient for *in vitro* binding is found in over half of yeast promoters (>4000 sites total).

**Fig. S7** (on the next two pages) Constructing a motif library for Hog1 regulatory network. **Left A and B**: The clustering tree produced by our motif clustering algorithm (see Methods). The motifs are associated with, from top to bottom: two Hot1 related motifs, a cluster of Msn2/4 variants, and a cluster of Sko1 variants. The automatic trimming of the tree into sub-trees is indicated with short red lines. **Right A and B**: Overlap of selected motifs' targets and expression-based groups (see Methods). The target genes shown in Fig. S6c for Sko1 activation, Sko1 repression and Hot1-activation (p-value<0.01), Msn2-activation (p-value<0.0001), and in all the genes outside the Hog1 pathway (Other), were compared with putative target genes based on motif analysis for the best motif of (From top to bottom) Hot1, Msn24 and Sko1.


A
% coverage
0
20
40
60
80
100
Sko1 (A)
Sko1 (R)
Hot1 (A)
Msn2/4 alone
Other
Sko1 (A)
Sko1 (R)
Hot1 (A)
Msn2/4 alone
Other
Sko1 (A)
Sko1 (R)
Hot1 (A)
Msn2/4 alone
Other

**B**


% coverage
0
20
40
60
80
100
Sko1 (A)
Sko1 (R)
Hot1 (A)
Msn2/4 alone
Other
Sko1 (A)
Sko1 (R)
Hot1 (A)
Msn2/4 alone
Other
Sko1 (A)
Sko1 (R)
Hot1 (A)
Msn2/4 alone
Other

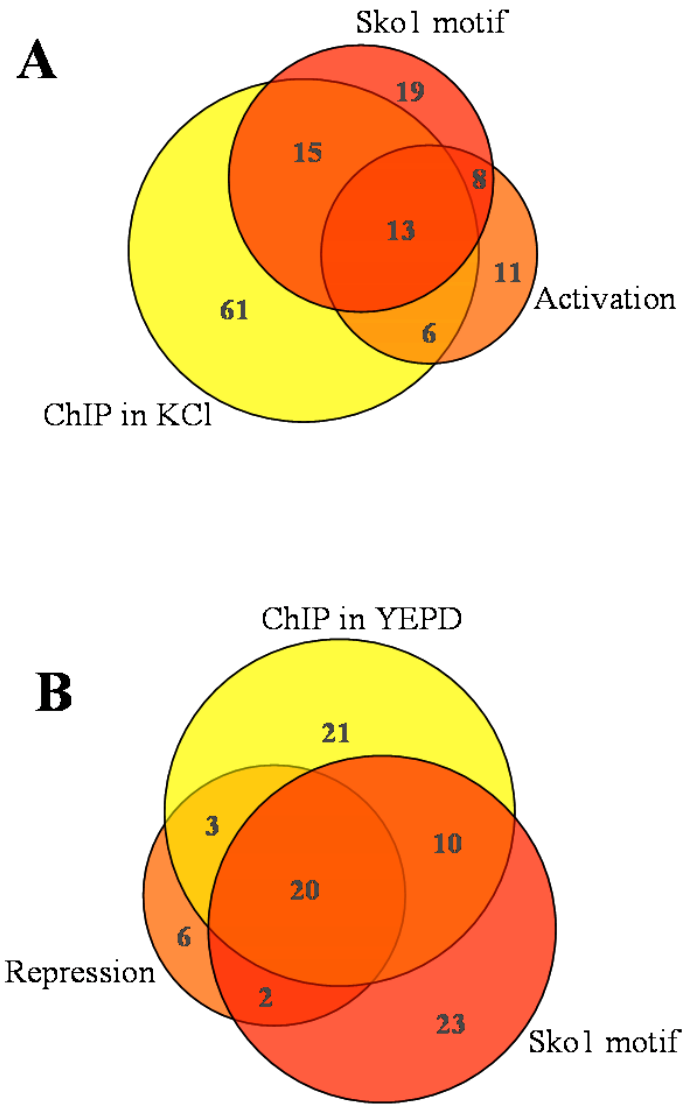



**Fig. S8** Significant overlap between motif targets and ChIP binding results. Venn diagrams showing the overlap between ChIP, expression and motif analysis for Sko1. (**A)** ChIP targets in KCl ($p<0.05$), Sko1 activated genes ($p<0.05$), and the putative target genes based on motif analysis using the best combination of three variants of the Sko1 binding motifs (combined FPR<0.05). (**B)** ChIP targets in YEPD ($p<0.05$), Sko1 repression ($p<0.05$), and the putative target genes based on motif analysis using the best combination of three variants of the Sko1 binding motifs (combined FPR<0.05).

## Methods

We use a three-step method for processing and integrating newly discovered DNA motifs into coherent and reliable sets of non-redundant motifs (Ref 15). The first step of the pipeline includes motif discovery and filtering. The second step includes clustering and merging of similar motifs, and the final step includes identifying the putative target genes for motifs (using genomic scans) and ranking them based on their enrichment in the genes of interest.

### Motif Discovery

On each set of potential co-regulated genes (or target bound sequences) we applied several motif discovery algorithms in parallel. MDscan[19], AligneAce[20], and MEME[21] were run using the TAMO package[22]. In addition, we used the SeedSearcher motif discovery algorithm (www.psi.toronto.edu/~yoseph/YosephBarashThesis.pdf). We applied these discovery algorithms to fifteen groups of genes based on gene expression of Sko1, Hot1, Hog1 and Msn2, and the ChIP data of Sko1 and Hot1. We converted the motifs to a position-specific count matrix representation for the further analysis.

### Construction of Non-redundant Motif set

The discovered motifs were compared and clustered using a dynamic programming algorithm based on the BLiC (Bayesian Likelihood 2-Component) score[15]. Roughly, the BLiC score measures similarity by evaluating to what extent two analogous positions in the two motifs have similar nucleotide preferences and are distinct from the background distribution. We clustered motifs using a hierarchical agglomerative clustering[15]. In each iteration of this procedure the two most similar motifs were merged. The merging operation involves aligning the two motifs and creating a new count matrix that

represents the sum of the two aligned count matrices. These iterations are repeated to form a dendrogram that contains all of the original motifs. A final set of non-redundant motifs was generated by trimming the dendrogram. The trim operation identifies sub-trees containing similar motifs within them and that are sufficiently different from each other. The motif at the root of each such sub-tree merges redundant motifs. Roughly, we require that a sub-tree is trimmed if it is sufficiently different from the next motif (or sub-tree) it is merged with. We used a stringent threshold for this step – we trim if the score of the two merged motifs is below 60% of their maximal score. This threshold was chosen as the optimal split threshold based on hand-curated splits of ten clustering trees.

**Identifying Target Genes and Scoring the Motifs**

For each motif we performed a genome wide scan of all the yeast promoters using the CIS program[23]. This program returns potential targets and their p-value (using a background distribution of promoters). We use a permissive threshold of CIS p-value < 0.1 to select potential sites. From these we take the targets that pass one of the following three criteria:

- **The promoter contains a binding site with a good sequence match to the motif, measured according to the p-value (after applying a Bonferroni correction for multiple hypotheses according to the average length of the scanned sequences).**
- The promoter contains a less significant occurrence of the motif (p-value threshold of 0.1), which is highly conserved among seven

species of the genus Saccharomyces. For this, we used the average conservation of the motif, according to the UCSC conservation track, with a threshold on the conservation[24]. **In situations where a promoter overlaps with a coding sequence, we assigned the overlapping regions a minimal (0) conservation score to avoid bias.**

- The promoter may contain a less significant occurrence of the motif in a specific position, but there is a high probability for a motif somewhere in the promoter (based on multiple weak sites along the entire promoter sequence). For this, two factors are taken into consideration – a threshold over the Bayesian posterior probability that the factor binds to the promoter sequence, which integrates over all possible sites, and the conditional posterior probability of finding a binding site at this specific location**.**

For each motif we choose the relevant threshold values separately by performing a search to find the values that maximize the true positive rate while allowing up to 5% false positive calls. That is, we choose threshold so that the at most 5% of promoters of genes

genes outside of the Hog1 pathway contain a positive call (as done in assigning p-values to peaks in the ChIP analysis).

In finding these thresholds, we used true positive groups according to the gene expression data as: Sko1 activation (p-value<0.01), Sko1 repression (p-value<0.01), Sko1-activation and repression (p-value<0.01), Hot1-activation (p-value<0.01), Msn2-activation (p-value<0.0001).

For Sko1 and Hot1 we also searched for sites that correspond to several variants of the same motif. To find variants of the same motif that together have a high true positive rate, allowing up to 5% false positive calls on their combined hits, we optimized the threshold parameters for all pairs and triplets of the Sko1 DNA motif variants and the Msn2/4 DNA motif variants.

Finally, we scored motifs according to their enrichment relative to the above groups of DNA sequences (-log of the hyper-geometric p-value). For each positive group we selected the top five scoring motifs (and motif combinations). The selected motifs were then clustered to create the final motif library shown in Fig. S7.

**Supplemental Tables and Flow Chart of Data analysis**

**Table S1** Contains all raw array data for KCl stress and a complete set of YEPD controls (all arrays in triplicate).

Fitting the data in Table S1 using the Matrices 1-3 (pgs 8-10 of the supplement) led to the values and errors of the expression components shown in Figs. 2, 3b, 3c and S6. These values and the associated p-values are listed in:

**Table S2** for genes in the Hog1 induced network (273 genes in Fig. 1) together with ChIP data for comparison

and

**Table S3** for all yeast genes with ChIP data for comparison

**Table S4** Contains representative images from the microscopy presented in Fig. 3.

**Table S5** Lists the strains used in this study.

**Table S6** Contains the raw data for all glucose stress experiments. Fitting these data using matrix 1 (pg 8 of the supplement) led to the expression components used to compare the glucose and KCl stress responses in Fig. 6. These values are listed in:

**Table S7** for all yeast genes.

## REFERENCES


1. Rothstein, R. Targeting, disruption, replacement, and allele rescue: integrative DNA transformation in yeast. *Methods Enzymol* **194**, 281-301 (1991).
2. Longtine, M.S. et al. Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. *Yeast* **14**, 953-61 (1998).
3. O'Rourke, S.M. & Herskowitz, I. Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. *Mol Biol Cell* **15**, 532-42 (2004).
4. DeGroot, M.H. & Schervish, M.J. *Probability and statistics*, 816 p. (Addison-Wesley Pub. Co., Reading, Mass., 2001).
5. Horovitz, A. & Fersht, A.R. Co-operative interactions during protein folding. *J Mol Biol* **224**, 733-40 (1992).
6. de Nadal, E., Casadome, L. & Posas, F. Targeting the MEF2-like transcription factor Smp1 by the stress-activated Hog1 mitogen-activated protein kinase. *Mol Cell Biol* **23**, 229-37 (2003).
7. Nevitt, T., Pereira, J., Azevedo, D., Guerreiro, P. & Rodrigues-Pousada, C. Expression of YAP4 in Saccharomyces cerevisiae under osmotic stress. *Biochem J* **379**, 367-74 (2004).
8. Proft, M. & Serrano, R. Repressors and upstream repressing sequences of the stress-regulated ENA1 gene in Saccharomyces cerevisiae: bZIP protein Sko1p confers HOG-dependent osmotic regulation. *Mol Cell Biol* **19**, 537-46 (1999).
9. Rep, M. et al. Osmotic stress-induced gene expression in Saccharomyces cerevisiae requires Msn1p and the novel nuclear factor Hot1p. *Mol Cell Biol* **19**, 5474-85 (1999).
10. Proft, M. & Struhl, K. Hog1 kinase converts the Sko1-Cyc8-Tup1 repressor complex into an activator that recruits SAGA and SWI/SNF in response to osmotic stress. *Mol Cell* **9**, 1307-17 (2002).
11. Alepuz, P.M., de Nadal, E., Zapater, M., Ammerer, G. & Posas, F. Osmostress-induced transcription by Hot1 depends on a Hog1-mediated recruitment of the RNA Pol II. *Embo J* **22**, 2433-42 (2003).
12. Rep, M., Krantz, M., Thevelein, J.M. & Hohmann, S. The transcriptional response of Saccharomyces cerevisiae to osmotic shock. Hot1p and Msn2p/Msn4p are required for the induction of subsets of high osmolarity glycerol pathway-dependent genes. *J Biol Chem* **275**, 8290-300 (2000).
13. Iyer, V.R. et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533-8 (2001).
14. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-9 (2000).
15. Habib, N., Kaplan, T., Margalit, H. & Friedman, N. A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS Computational Biology* **4(2)**, e1000010 (2008).
16. Proft, M., Gibbons, F.D., Copeland, M., Roth, F.P. & Struhl, K. Genomewide identification of Sko1 target promoters reveals a regulatory network that operates

in response to osmotic stress in Saccharomyces cerevisiae. *Eukaryot Cell* **4**, 1343-52 (2005).
17. Martinez-Pastor, M.T. et al. The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *Embo J* **15**, 2227-35 (1996).
18. Schuller, C., Brewster, J.L., Alexander, M.R., Gustin, M.C. & Ruis, H. The HOG pathway controls osmotic regulation of transcription via the stress response element (STRE) of the Saccharomyces cerevisiae CTT1 gene. *Embo J* **13**, 4382-9 (1994).
19. 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).
20. Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**, 1205-14 (2000).
21. 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).
22. Gordon, D.B., Nekludova, L., McCallum, S. & Fraenkel, E. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics* **21**, 3164-5 (2005).
23. Barash, Y., Elidan, G., Kaplan, T. & Friedman, N. CIS: compound importance sampling method for protein-DNA binding site p-value estimation. *Bioinformatics* **21**, 596-600 (2005).
24. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).