Supporting Text

Determination of Affected Genes.

We identified genes affected by the following environmental and genetic inputs:

1) carbon source change from raffinose to galactose, and
2) deletion of known *GAL* genes (*GAL1*, *GAL2*, *GAL3*, *GAL4*, *PGM2*, *LAP3*, *GAL7*, and *GAL10*).

We measured three types of high-throughput data:

1) Gene expression levels at 1, 5, 10, 30, and 60 min, and 3, 6, and 9h after galactose addition using two-channel microarrays (1). (Six or more biological replicates were hybridized at each time point, and half of the arrays were dye-flipped to avoid bias toward a particular dye.
2) Relative abundance of proteins in galactose to those in raffinose (2).
3) Gene expression levels with deletions of the *GAL* genes to assess the role of *GAL* genes in galactose utilization and also in regulating other biological processes (2).

Normalization of microarray data.

Microarray experiments are subject to various sources of variability such as biological variability coming from heterogeneity in samples; and experimental variability occurring during sample preparation, PCR reaction, spotting, hybridization, and scanning. These kinds of variability may affect gene expression analyses such as identification of differentially expressed genes, correlation analysis and regression analysis. We need to normalize microarray data to minimize these kinds of variability.

Here, we used a modified version of the normalization method proposed in Yang *et al*. (3). Our method also adjusts intensities to remove intensity dependent bias using a nonparametric smoother (super-smoothing). Spatial variation is handled by treating each grid location individually using MA-plots (Eq. **1** below). A major departure from Yang *et al*. (3) is that our method removes variation across and within arrays at the same time. Also, we adjusted both cy3 and cy5 channel intensities together instead of adjusting each channel intensity individually. The normalization procedure is as follows. First, we generate a MA plot for each grid location:

$$M_{jk} = \log_2\left(X_{jk}/Y_{jk}\right), \quad A_{jk} = \log_2\left(X_{jk}Y_{jk}\right)/2, \qquad \textbf{[1]}$$

where X_{jk} represents a cy3 intensity for gene j of array k in grid i (this index is not shown for notational convenience, but note that all the computation is done for grid i hereafter).

Second, we compute the median of A_{jk}s in all arrays for all genes (i.e., $med(med(A_{jk}))$, and the deviation of the median of each gene j from the overall median is then defined as $c1_j = med_k(A_{jk}) - med_j(med_k(A_{jk}))$. Third, we apply super-smoothing to estimate the intensity

dependent relation between M_{jk} and A_{jk}s for all genes of all the arrays in grid *i*. Thus, $c2_{jk}$ is the super-smoothing fit to the MA plot for grid location *i*. Super-smoothing adaptively determines the kernel window size depending on the sample distribution instead of fixing it to a pre-defined value (e.g., 0.2) as done in Yang *et al*. (3). Fourth, we adjust M_{jk} and A_{jk}s using $c1_j$, $c2_{jk}$, X_{jk} and Y_{jk}s accordingly.

$$M'_{jk} = M_{jk} - c2_{jk}, \quad A'_{jk} = A_{jk} - c1_j, \qquad \textbf{[2]}$$

$$\log_2\left(X'_{jk}\right) = A'_{jk} + M'_{jk}/2, \quad \log_2\left(Y'_{jk}\right) = A'_{jk} - M'_{jk}/2. \qquad \textbf{[3]}$$

Fifth, for scale normalization, we compute MAD_ks (median absolute deviation) of M_{jk} and A_{jk}s for all arrays, respectively, and their medians: $MAD_k(M) = med_j(abs(M_{jk} - med_j(M_{jk})))$, $MAD_k(A) = med_j(abs(A_{jk} - med_j(A_{jk})))$, $MAD(M) = med_k(MAD_k(M))$, and $MAD(A) = med_k(MAD_k(A))$. Then, the deviation of $MAD_k(M)$ for each array *k* from the overall median $MAD(M)$ is defined as:

$$sM_k = MAD(M)/MAD_k(M) \text{ and } sA_k = MAD(A)/MAD_k(A).$$

Sixth, we adjust M_{jk}' and A_{jk}s$'$ using sM_k and sA_k:

$$M''_{jk} = sM_k\left(M'_{jk} - med_j\left(M'_{jk}\right)\right) + med_j\left(M'_{jk}\right), \quad A''_{jk} = sA_k\left(A'_{jk} - med_j\left(A'_{jk}\right)\right) + med_j\left(A'_{jk}\right), \qquad \textbf{[4]}$$

$$\log_2\left(X''_{jk}\right) = A''_{jk} + M''_{jk}/2, \quad \log_2\left(Y''_{jk}\right) = A''_{jk} - M''_{jk}/2. \qquad \textbf{[5]}$$

Finally, the steps 1-6 are applied to X and Y intensities in all the 48 grid locations (i.e., *i*=1 to 48 in this study). These steps remove intensity dependent and spatial biases across and within the chips. This normalization procedure was applied to microarray data at each time point. Then, the final median values of all the spots for each time point were normalized to ensure that all the $\log_2$ ratio (i.e., $\log_2(X / Y)$) median values at each time point have the same value and MAD.

Determination of P values for each data set using Gaussian kernel density estimation.

For a data set E (called an evidence type hereafter) and a collection G of elements (genes/proteins), we define a global set of observations $O \subseteq \Re \times E \times G$, where $\Re$ is the real number space. An observation $(y, e, g) \in O$ means that evidence type *e* was observed for element *g*, and the value *y* resulted.

First, we defined a measure (*y*) for each evidence type. For example, we defined the following measures for three types of evidence ($E = \{e_1, e_2, e_3\}$) that were used to determine genes (*H*) affected by environmental and genetic perturbations.

1) For time-course array data, integrated $\log_{10}$ ratios of gene expression at each time point, to expression in raffinose: $y_1 = \int_0^{9hr} \log_{10}\left(g_{gal_t} / g_{raf_t}\right)dt$, where g_{gal_t} and g_{raf_t} represent expressions of each gene (g) in galactose and raffinose, respectively, at time = t. Fig. 5A shows how each *GAL* gene varies over 9 hrs (see the shaded area in Fig. 5A for *GAL10*). Trapezoidal method was used for the numerical integration.
2) For deletion array data, maximum $\log_{10}$ ratios of gene expression levels in wild type to those in *GAL* deletion mutants: $y_2 = \underset{k}{Max}\left|\log_{10}\left(g_{wt} / g_{\Delta gal_k}\right)\right|$, where $g_{\Delta gal_k}$ and g_{wt} represent the expression of each gene when *GAL* gene k is deleted, and in wild type, respectively.
3) For proteomic data, $\log_{10}$ ratios of protein expression levels in galactose to those in raffinose (see Fig. 5B): $y_3 = \log_{10}\left(g_{gal} / g_{raf}\right)$, where g_{gal} and g_{raf} represent expression levels of each protein (g) in galactose and raffinose, respectively.

Second, we applied a Gaussian kernel density estimator to each of these measures (y) of all genes (g) to estimate a nonparametric density function $D_e(y)$ for each evidence type. $\int_{-\infty}^{\infty} D_e(y)dy = 1$ for all types of evidence $e \in E$. We assume that $D_e(y)$ represents the distribution of genes not affected given perturbations, which is true when only a small fraction of measured genes are affected in an experiment (generally the case in high-throughput experiments). Fig. 5B shows the estimated density function for each measure.

Third, we performed a two-tailed hypothesis test using the estimated distribution to determine the empirical probability (P value or significance $S_e(y)$) that observe a particular measure (y) by chance when the corresponding gene (g) is actually not affected (i.e., when the null hypothesis, H_0: $g \notin A$ where A is the set of true affected genes, is true). This P value is then defined by $S_e(y) = S_e(y|g \notin A) = 2Min\left(S_e^{'}(y), 1 - S_e^{'}(y)\right)$, where $S_e^{'}$ is defined as $\int_{-\infty}^{y} D_e(y^{'})dy^{'}$ (see also *Eq.* **10** for one tailed test). Note that we used the symbol "*S*" (representing significance) for P value to differentiate it from probability used in Bayesian methods. For example, given the time-course data (e_1), the statistical measure (y_1) for *GAL10* is 5.53, and the corresponding P value $S_e(y_1)$ is defined by 2 times the area under the distribution curve from 5.53 to positive infinite (4).

These steps were also applied to the other two measures to calculate P values for deletion array data (e_2) and proteomic data (e_3) (see Fig. 5B) when the null hypothesis is true.

Estimation of PP and PD interaction maps.

To better understand the interactions among the affected genes, we estimated reliable PP and PD interaction maps by integrating various types of evidence that can help infer the

presence of PP and PD interactions for a PP and PD pair. Sections 2.1 and 2.2 present mainly the types of evidence used and the estimation of a probability (*P* value) of a PP or PD pair not interacting with each other given each type of evidence. In addition to experimental data, we used computational methods to predict both PD and PP interactions. To improve prediction accuracy (or not to bias selection of PP and PD interactions toward predictions from a relatively large number of computational methods), we first integrated the computational predictions and then integrated the overall *P* values from the computational predictions with *P* values from several types of experimental data (see Fig. 1 in the text). Thus our PP and PD interaction result are based on a nested set of integrations, as describe din Sections 2.1.1, 2.1.5, and 2.2.3. Note, however, that we used the same procedure (identical to that used for identifying the affected genes) to perform each PP and PD evidence integration.

Estimation of the PP interaction map.

We determined PP interactions by integrating the following five types of data (see Fig. 1B): *i*) the full set of PP interactions in DIP and BIND, including data from yeast two hybrid and TAP-tag assays, as well as paralog interaction analysis; *ii*) the combined sub-cellular localization data from SGD (GO cellular components) and GFP database (5); *iii*) gene expression correlations estimated from ~1300 gene expression profiles from ExpressDB with additions from our time-course, and deletion gene expression profiles; *iv*) changes in gene expression level due to gene deletions, and *v*) domain-domain (DD) interactions computationally predicted by Multiprospector and InterDom.

P value for PP interaction detection methods.

We used single high-throughput experiments (H), small-scale experiments (S), multiple high throughput experiments (M), and paralogous interactions to verify protein interactions (R). Deane *et al*. (6) analyzed PP interactions in DIP and assessed the reliability of individual detection methods based on an expression profile reliability index as summarized in Fig. 7. We developed a scheme for estimating the *P* value for PP interactions detected by multiple kinds of methods using the reliability measures in Deane *et al*. (6). For instance, for the PP interactions detected by both multiple high throughput experiments (M), and paralogous interaction method (R), the *P* value $S\left(PP \notin H \middle| M, R\right)$ can be estimated (assuming that the *P* value can approximate as the probability of the detection methods (discrete variable) identifying the interactions for protein pairs when the null hypothesis $PP \notin H$ is true):

$$S_e(y) \equiv S\left(K \middle| PP \notin H\right) \approx P\left(K \middle| PP \notin H\right)$$

$$= \prod_l^L P\left(k_l \middle| PP \notin H\right) = \frac{\prod_l^L P\left(k_l\right)}{P\left(PP \notin A\right)^L} \prod_l^L P\left(PP \notin H \middle| k_l\right) \qquad \textbf{[6]}$$

where $K=\{k_1, k_2, \ldots, k_L\}$ is the set of detection methods (either S, H, M, or R), and $P(PP \notin H|k_l)$ is the *P* values for a detection method k_l, as reported in Deane *et al*. (6): for example, $P(PP \notin H|M)$ =0.22 and $P(PP \notin H|R)$ = 0.135 (see Fig. 7). The prior probabilities are estimated as the numbers of PP interactions in DIP detected by the corresponding methods: N_i represents the number PP interactions detected by method *i*. Also, the total number of noninteracting proteins is estimated by $\lfloor N_p \times (N_p + 1)/2 - t_p N_p/2 \rfloor$, following Deng *et al*. (7), where N_p is the number of proteins being considered (6307 in this study), and t_p represents an estimated number of interactions per protein (50 in this study). For instance, $S(M, R|PP \notin H)$ = 3,350 × 1,922 × 0.22 × 0.135 / (6,308 × 6,307 / 2 – 50 × 6,307 / 2) = 4.910×10^{-10}.

We summarized the probabilities $P(PP \notin H|k)$ derived from Deane *et al*. (6) in Fig. 7, and also summarized the *P* values for all the combinations of detection methods in Table 1A. They were assigned to all 15118 PP interactions in DIP (December 20, 2003). BIND (August, 2003) includes 763 yeast two hybrid data not found in DIP. For these data, we assigned a *P* value of 2.921×10^{-4} (corresponding to the detection method H in Table 1A). For the rest of PP combinations, we assigned a conservative *P* value of 1. This issue is further addressed below (see the second paragraph in *P value for expression changes in deletion experiments*).

P value for cellular compartments.

Cellular compartments of the interacting proteins were used as the second source of evidence in determining PP interactions. This is based on the assumption that interacting proteins should be in the same cellular compartment. We estimated the *P* value for this evidence based on the Minimum Number of Transport Processes (MNTP) required for the interacting proteins to be in the same compartment: $S(MNTP_{CC}|PP \notin H)$ for any pair of cellular compartments. First, we grouped all the annotated cellular compartments into 26 groups: ER, ER-Goli, Golgi, Golgi membrane, bud, cell, cell cortex, cell fraction, cell wall, cytoplasm, cytosol, endosome, extracellular, intracellular, membrane, mitochondrion, nucleoplasm, nucleus, periplasmic space, peroxisomal membrane, peroxisome, plasma membrane, septum, unknown, unlocalized, vacuolar membrane, vacuole. Then, the MNTPs were rationalized for all the pairs of these 26 components.

Second, we derived the *P* value for each of these MNTPs:

$$S_e(y) \equiv S(MNTP_{CC} = l|PP \notin H) \approx 1 - P(MNTP_{CC} \geq l|PP \notin H)$$
$$= 1 - \sum_{k \geq l} \frac{\left[N_{MNTP_{CC}=k}\left(N_{MNTP_{CC}=k} + 1\right)/2 - t_p N_{MNTP_{CC}=k}/2\right]}{\left[N_a(N_a+1)/2 - t_p N_a/2\right]} \approx 1 - \frac{\sum_{k \geq l} N_{MNTP_{CC}=k}}{N_a}, \qquad \textbf{[7]}$$

where N_a and $N_{MNTP_{CC}=k}$ represent the number of annotated proteins and the number of protein pairs that requires a *k* MNTP to be in the same cellular compartment, respectively. As an estimate for N_{MNTP} (when MNTP>0), we counted the number of proteins localized in MNTP compartments (see *Eq.* **6** for the definition of t_p). This is the one-sided (left-sided) test of a discrete statistical measure (MNTP): it should be left-sided because PP$\in H$ should be at the same compartment (MNTP=0). For the pairs of cellular components with MNTP=0 (e.g., nucleus and nucleoplasm), we assigned the *P* value of 0.05 to allow for annotation errors. See Table 2 for the actual *P* values used. Note that proteins in the membrane interface with two cellular compartments, and thus require no transport to interact with proteins in either of the two compartments (e.g., MNTP = 0 for proteins in vacuole membrane and cytoplasm).

The *P* value $S\left(MNTP_{CC}\middle|PP\notin H\right)$ reflects just the chance that two noninteracting proteins can be in a particular compartment, not the chance that we observe a specific pair of cellular compartments when two proteins do not interact, i.e., $S\left(CC\middle|PP\notin H\right)$. We used the *P* value $S\left(MNTP_{CC}\middle|PP\notin H\right)$ for the variable MNTP, derived from GO cellular components (GOCC) and Huh *et al*. (5), because it is a more appropriate measure for our assumption that two interacting proteins should be in the same compartments. The latter *P* value $S\left(CC\middle|PP\notin H\right)$ does not reflect how easy two protein pairs can be colocalized in the same compartments for several reasons (e.g., incompleteness in cellular component annotations).

P value for expression correlations.

It has been suggested that the expression levels of interacting proteins tend to correlate with each other (6). This motivated us to use the correlation between gene expressions as the third type of evidence in inferring PP interactions. First, we transformed all the expression data obtained from Affymetrix and two-channel cDNA or oligoarrays into the $\log_{10}$ ratio format (conditions of interest versus controls) and, then, these $\log_{10}$ ratios for each chip was normalized to ensure that their median and MAD (median absolute deviation; see above) is the same as the overall median and MAD of all the arrays in the data set, respectively, which includes ≈ 1300 arrays. We then combined all the replicate arrays into one representative array by taking the medians of all the same genes in multiple replicate arrays.

Second, we computed the correlations of all the possible interacting protein pairs (i.e., 6,308 × 6,307 / 2 for the total 6,307 proteins). Third, we calculated the *t* statistic for the correlation of each protein pair:

$$t = r\left(N-2\right)\Big/\sqrt{1-r^2}\,, \qquad \textbf{[8]}$$

where r is the correlation, and N is the number of samples used to compute the correlation. This t statistic theoretically follows t_{N-2}, but the actual data does not meet the underlying assumption, thus causing the estimated P values to be often underestimated. We therefore applied kernel density estimation to directly estimate the distribution from all the actual t statistic values. Fourth, the P values for each protein pair was determined by a two-tailed test (assuming that all possible PP pairs represent the characteristics of noninteracting proteins due to the small fraction of interacting proteins). This is the probability of observing an expression correlation by chance when the corresponding protein pair does not interact (i.e., when the null hypothesis is true).

P value for expression changes in deletion experiments.

Following Deane *et al*. (6) and Zhang *et al*. (8), we used expression change after the deletion of one of two genes as the fourth type of evidence:

$$y = \underset{k}{Max}\left|\log_{10}\left(\Delta g_k / wt\right)\right|, \qquad \textbf{[9]}$$

where Δg_k and wt represent the expression of each gene after the deletion of gene k and in wild type, respectively.

First, we calculated the measures (*Eq*. **9**) for gene pairs for which there are knock-out gene expression data available (deletion data for 294 genes were used in this study; ExpressDB). Then, we applied Gaussian kernel density estimation to obtain a distribution of the measure, and P value was then estimated by a one-tailed test:

$$S_e(y) \equiv S\left(y \middle| PP \notin H\right) = \begin{matrix} 1 - S_e^{'}(y) & right-sided, \\ S_e^{'}(y) & left-sided \end{matrix} \qquad \textbf{[10]}$$

where $S^{'}_e$ is defined as $\int_{-\infty}^{y} D_e(y^{'})dy^{'}$ for the empirically estimated distribution $D_e(y)$ of y.

We used a P value of 1 for pairs for which no knockout expression data was available for one of the interaction partners. Thus, in effect, if data is not available, we assumed the interaction is not supported by this assay, making a conservative decision about network membership. However, we could instead decide not to estimate the P value for these PP interactions and then integrate the other four types of evidences only. If so, the types of evidences being integrated vary depending on data availability, thus the estimation of the P value involving the change in the degree of freedom for each data element based on data availability for that element. In this study, we avoid this complexity by assuming the P value of 1 for nonavailable data. However, Pointillist allows any value (e.g., 0.5) to be used for missing values.

P value for DD interactions.

Computational predictions for DD interactions were used as the fifth type of evidence. We have two sets of predictions for DD interactions, which were generated by InterDom and Multiprospector, respectively. These two types of evidence were integrated together first (in a nested manner; see Fig. 1) to estimate the overall *P* values for computational predictions. We used this nested approach because there are often many more computational predictors available than experimental sources of data, which leading data integration to being biased toward the computational methods. Although this issue may not seem serious with two sources of computational predictions, it can be serious with increasing the number of types of predictions as described below.

To determine the representative *P* value value for each DD interaction, we first applied kernel density estimation to the prediction scores from each method: InterDom (e_{DD1}), and Multiprospector (e_{DD2}: interfacial energy). We then estimated a *P* value $S_{e_{DD}}(y) = S(e_{DD}|DD \notin H_{DD})$ for each of the two types of prediction scores using the corresponding estimated distribution (one-sided test) using *Eq.* **10** (assuming that a majority of domain pairs do not interact). Finally, the two types of *P* values were integrated as explained above. The trustworthiness weights for the InterDom and Multprospector prediction scores were 0.491 and 0.509, respectively. Then, the overall *P* value for two data sets was calculated by applying the one-sided test using the empirical density function $F(\tilde{s}'_{DD}, H^c_{DD})$ for the combined *P* value:

$$S_{E_{DD}}(y) = S(E_{DD}|DD \notin H_{DD}) = 1 - \int_{-\infty}^{\tilde{s}'_{C_{DD}}} F(\tilde{s}'_{C_{DD}}, H^c_{DD}) d\tilde{s}_{C_{DD}} \qquad \textbf{[11]}$$

where E_{DD}={e_{DD1}, e_{DD2}}, and the symbols sub-indexed by *DD* represent their corresponding parts described above for the case of DD interactions. Finally, we estimated the overall *P* value $S_e(E_{DD}|PP \notin H)$ for PP interactions by taking the minimum value if there are multiple DD interactions for a single PP pair: $S_e(y) \equiv S_e(E_{DD}|PP \notin H) = \min(S_{E_{DD}}(y))$ for $\forall DD \in PP$.

Finally, Pointillist integrated the five types of *P* values estimated from *P value for PP interaction detection methods* to *P value for DD interactions*, as shown in Fig. 1B (see text).

2.1.1 Estimation of PD interaction map.

We determined protein-DNA (PD) interactions by integrating the following five types of data (Fig. 1C): (*i*) ChIP-chip data for 113 transcription factors in YPD media reported in Lee *et al*. (9), supplemented with Gal4p and Mth1p ChIP-chips in galactose media (1); (*ii*) sub-cellular localization data from SGD and GFP database; (*iii*) gene expression correlation between TFs and their target genes; (*iv*) expression changes after deletion of 23 transcription factors; (*v*) the overall *P* values of five computational TFBS prediction tools (Fig. 1F).

In this study, we only considered 135 transcription factors, which is a combined set of transcription factors obtained from Lee *et al*. (9) and known transcription factor binding sites in Table 6. Also, we included several reported transcription factor binding sites for Adr1p, and Pip2p. Here, we explain the statistical measures used to estimate *P* values for the five types of evidence (see text for other information).

P value for ChIP-chip data.

ChIP-chip data obtained from Lee *et al*. (9) were used as the first type of evidence. To study Gal4p-DNA binding in galactose, we carried out additional ChIP-chip for Gal4p in galactose and replaced the overlapping portion of ChIP-chip data in YPD with galactose-specific data. Later, to validate our model predictions, we performed additional ChIP experiments for Mth1p in galactose to validate the Mth1p binding to the promoter region of *HXT7*.

The $\log_{10}$ ratios of intensities of IP enriched versus whole cell extracts were computed as a statistical measure. We then applied kernel density estimation to estimate the distribution of the ratios and then calculated the *P* value for each ratio using the estimated distribution (one-sided test) as shown in *Eq.* **10**.

P value for transcription factor cellular compartments.

We assigned *P* values $S_e(y) \equiv S(TF_{CC} | PD \notin H)$ =0.05 for proteins annotated with "nucleus" as their cellular location, and 0.95 for those annotated with the other cellular compartments. This is different from how cellular compartments were used above. For PD interactions, it only matters whether transcription factors exist in nucleus for their binding to DNA.

P value for expression correlations.

We computed *P* values $S_e(y)$ for all TF-target pairs as described above using *Eq*s. **8** and **10**. The only difference from that above is that the distribution estimated by kernel density estimation is constructed by only the pairs of transcription factors being investigated and all DNA targets (i.e., 135 × 6,307, which is different from all possible protein pairs, 6,308 × 6,307 / 2).

P values for deletion effects.

For PD interactions, we only considered the effect of the deletion of transcription factors (23 transcription factor deletion experiments) on downstream genes, while all available 294 deletion data were used above. The *P* values were computed in the same way as described above using *Eqs.* **9** and **10**.

P value for Mogul predictions.

We used our in-house software package Mogul for computational predictions of TFBSs (http://labs.systemsbiology.net/bolouri/software/Mogul). Although Mogul includes more than 30 algorithms for TFBS predictions, we used only five relevant, representative algorithms (fuzznuc, AligAce, MEME, Sampler, and MotifSampler) for this study. First, Mogul was run on sets of randomized sequences to generate the basis for statistical tests (i.e., computation of *P* values for the outputs of each algorithm) as follows:

1) Run MEME with four different window sizes (5, 10, 15, and 20).
2) Run AlignAce, MotifSampler, and Motsa 10 times for each of four window sizes as MEME above.
3) Determine the average and the standard deviation of scores from each algorithm for the random sequences. Once the Mogul run is completed, the outputs from each algorithm for each gene were summarized to a null $r \times l$ score matrix where r and l represent the number of runs (i.e., 4 for MEME, and 10 for the other coregulated algorithms) and the length of the intergenic region, respectively. Once the *i-th* run of each algorithm is done, the score values of all the predicted binding sites for each coregulated gene were normalized between 0 and 1. Then, we added the result to the *i*-th row of the score matrix of the gene. For example, if MEME predicted the TGAAACAATA in the coordinate [412,422] of the intergenic region of the coregulated gene *p* in the second run, the normalized score 0.62 of the predicted binding site was added to the elements from 412 to 422 in the second row of the MEME score matrix for the gene *p*. Then, the column-wise mean vector ($1 \times l$) of the score matrix ($r \times l$) of each algorithm represents the averaged importance of each base-pair being a part of the binding sites predicted by each algorithm. The mean vector along the intergenic coordinate reveals the density of predicted binding sites (see the heights of the shaded region in Figs. 8A and 8B). Finally, the average and the standard deviation of the mean vector elements for each algoritm were computed for a statistical test (see below).

Next, Mogul was run on upstream sequences of putative coregulated genes determined from ChIP-chip data.

1) Scan for known TFBSs using fuzznuc.
2) Run MEME with four different window sizes (5, 10, 15, and 20).
3) Run AlignACE, MotifSampler, and Motsa 10 times for each of four window sizes as MEME above.
4) Compute the mean vector from each algorithm for a coregulated gene as described in step 4 above.
5) Test whether each element (i.e., each base-pair) of the mean vector against the average and standard deviations for randomized sequences (see Step 4 above; denoted by MASR in Fig. 8A) using a one-tailed *t* test. The resulting *P* values indicate the probability $S_{e_M}(y) \equiv S(e_M | n \not\subset B)$ of observing a mean score for each base pair (n) by chance, when the null hypothesis ($n \not\subset B$, where B represent a set of binding sites) is true (note that each set of randomized sequences are not coregulated genes, and the distribution used in one-tailed *t* test was determined by these sets of randomized sequences). Fig. 8B shows that the intergenic region of

AGA1 has a sequence corresponding to the known binding site for *Ste12* (marked by an arrow in the bottom panel) between coordinates [412,418]. For Fuzznuc, we assigned the *P* value of 0.05 to predicted binding motifs.

Finally, these *P* values were integrated as described in Section 2.1.5. We formed the initial set *H* by selecting base-pair coordinates whose *P* values were less than 0.05 for at least one of the algorithms. The overall *P* values are computed once we identified the final value of *H* (see above).

Also, we used the following parameters to run Mogul algorithms:

1) For the large sets of coregulated genes (>40), only MEME with four window sizes (5, 10, 15, and 20) was run due to the heavy computational load.
2) Instead of randomizing individual upstream sequences, we generated random sequences with the GC content (35.1%) as the noncoding region of the yeast genome (see http://biochemie.web.med.uni-muenchen.de/Yeast_Biology/052_Genome2.htm).
3) We ran three coregulated algorithms (AlignACE, MotifSampler, and Motsa) 10 times each to ensure the estimation of consistent prediction scores for Gibb's-sampling based algorithms.

Finally, we applied Pointillist to integrate the *P* values for the five types of evidences (see Fig. 1C).

Network Analysis

We developed a recursive algorithm that can build a parsimonious sub-network model by extracting only relevant interactions to the selected genes rather than including all PP and PD interactions in the interaction maps. The algorithm permits the resulting sub-network to include only the interactions branches connecting the affected genes via a certain number of intermediate proteins as follows (see Fig. 9A): (*i*) add the 69 affected genes to the source and sink and then select the interactions (called the interactions at depth=0) between the nodes in the source and sink (see Figs. 9B); (*ii*) identify the first neighbors of the genes in the source and sink from the PP and PD interaction maps; (*iii*) add these neighbors to the source and sink, and then select new interactions (called the interactions at depth=1) between the nodes in the source and sink recursively updated; and (*iv*) repeat steps 2 and 3 until the new interactions identified by adding the first neighbors does not significantly improve the connectivity status of the affected genes among them. As a measure for such connectivity status, we used the ratio of the number of connected clusters of the affected genes to the total number of proteins in the sub-network. The recursive iteration was terminated at depth=1 when this measure reached and remained below the cutoff value of 0.01 (see Fig. 9C). The resulting sub-network is shown in Fig. 10A and compared with the one constructed using the interaction maps composed of the PP interactions in DIP and BIND and the PD interactions identified by ChIP-chip data (see Fig. 10C).

References

1. Weston, A. D., Baliga, N. S., Bonneau, R. & Hood, L. (2003) *Cold Spring Harb Symp. Quant. Biol.* **68,** 345-357.
2. Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R. & Hood, L. (2001) *Science* **292,** 929-934.
3. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. & Speed, T. P. (2002) *Nucleic Acids Res.* **30,** e15.
4. Lock, C., Hermans, G., Pedotti, R., Brendolan, A., Schadt, E., Garren, H., Langer-Gould, A., Strober, S., Cannella, B., Allard, J. *et al*. (2002) *Nat. Med.* **8,** 500-508.
5. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S. & O'Shea, E. K. (2003) *Nature* **425,** 686-691.
6. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. (2002) *Mol. Cell. Proteomics* **1,** 349-356.
7. Deng, M., Mehta, S., Sun, F. & Chen, T. (2002) *Genome Res.* **12,** 1540-1548.
8. Zhang, L. V., Wong, S. L., King, O. D. & Roth, F. P. (2004) *BMC Bioinformatics* **5,** 38.
9. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I. *et al*. (2002) *Science* **298,** 799-804.
10. Ostergaard, S., Olsson, L. & Nielsen, J. (2001) *Biotechnol. Bioeng.* **73,** 412-425.