Supplemental Information: A Gibbs Sampler for the Identification of Gene Expression and Network Connectivity Consistency

Mark P. Brynildsen, Linh M. Tran and James C. Liao*

Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA 90095, USA

This supplementary information is designed to give details on issues that are not fully discussed in the main article. It does not refer to other concepts than those introduced in the main article.

1 DATA PROCESSING

1.1 Connectivity Data

Using a *p*-value threshold of $1x10^{-3}$, transcriptional regulatory networks were obtained from the ChIP-chip data of (Lee *et al.,* 2002) and (Harbison *et al.,* 2004) (YPD and all conditions). The networks were then merged to obtain a network comprised of all transcription factor-promoter binding relationships known through ChIP-chip experimentation. The total number of genes investigated was 6,229, which is equal to the number of genes in the ChIP-chip binding data from the YPD and condition specific assays from (Harbison *et al.,* 2004). It is worth noting that network connectivity can be constructed from ChIP-chip data in a variety of ways, and that the method of construction may impact an analysis. For instance, more lenient or strict p-value thresholds, or binding ratios could be used. The p-value thresholds used here were $1x10^{-3}$, a very strict, common threshold used in transcription analyses, including those of the developer and authors of the ChIP-chip data used here (Lee *et al.,* 2002; Harbison *et al.,* 2004). Use of more lenient values would increase the connection false positive rate, thereby increasing network error, and pushing ChIP-chip derived network performance towards that of random networks. More restrictive p-values would decrease the connection false positive rate, potentially aiding analysis. However, at a p-value of $1x10^{-3}$ (false positive rate $\simeq$ 4-6%) the majority of uncertainty in the ChIP-chip network would originate from environmental dependence in binding and uncorrelation between binding and regulation, not experimental error. In fact, an overly restrictive p-value could be detrimental since fewer genes could be analyzed, due to more genes being found unbound in ChIP-chip data. Use of raw binding ratios would increase network error and result in similar issues confronted from using more lenient p-values, due to the absence of corrections performed in the process of converting binding ratios to p-values.

1.2 Expression Data

Expression data was gathered from literature (Gasch *et al.,* 2000; Lyons *et al.,* 2000; Gasch *et al.,* 2001; Yoshimoto *et al.,* 2002). The mean normalized expression ratios were extracted from every experiment. The Gene Expression Pattern Analysis Suite v1.1 (GEPAS) was used to process every experiment individually (Herrero *et al.,* 2003; Herrero *et al.,* 2004). Replicates were omitted if they were 1.0 from the median, and merged otherwise. The expression of a gene was the median value of those replicates that were less than 1.0 from the median. After every experiment was individually processed, experiments from the same condition were combined into a single data matrix. For each condition, the data matrix was further processed by GEPAS by filtering out those genes that were absent in more 20% of the experiments, and imputing the missing values for those genes that were present in more than 80% of the experiments by K-Nearest Neighbor (KNN) imputation where K = 15. Every condition data matrix was further processed by omitting those genes that did not show at least a 2-fold change in expression, up or down, in any of the condition's experiments. This is not a necessary step for the method, but allows expedited computation for those genes whose expression was more considerably altered. Lastly, those genes that did not have any regulator bind them in the connectivity dataset were not analyzed, and the remaining expression was z-scored by gene. There were a total of 10 experimental conditions.

2 METHOD COMPARISON

2.1 ChIP-chip, Random Network Comparison

Fixed Strength (type I): ChIP-chip connectivity was obtained as described in section 1.1. To solve Eq. 1 from the main text, non-zero values of $\mathbf{A}$ were specified to be 1, and $\mathbf{\Gamma}$ was minimized via Ordinary Least Squares (OLS). Randomized connectivity was obtained by shuffling the connections (1's and 0's) in $\mathbf{A}$. Eq. 1 was then solved in the same manner. Ten different randomized connectivities were performed for every experimental condition (10 each, 100 total), and each data point in Figure 4 was the average of these ten analyses.

Variable Strength (type II): ChIP-chip connectivity was obtained as described in section 1.1. To solve Eq. 1 from the main text, non-zero values of $\mathbf{A}$ were initialized by random numbers and a bi-linear optimization was performed to minimize $\mathbf{\Gamma}$. Ten different initializations were performed for every experimental condition, and the average of their residuals was used in Figure 4. Randomized connectivity was obtained by shuffling the connections (nonzero values) in $\mathbf{A}$. Eq. 1 was then solved in the same manner. Ten different randomized connectivities were performed for every experimental condition, and each data point in Figure 4 was the average of these ten analyses.

2.2 Average Relative Residual

Calculated as:

$$Average\ Relative\ Residual = \frac{\sum_{i=1}^{N} \frac{\|\Gamma_i\|_F}{\|e_i\|_F}}{N} \qquad \text{(S1)}$$

where $\|X\|_F$ is the Frobenius norm of X, Γ_i is the residual of the i^{th} gene, and e_i is the expression of the i^{th} gene. Results in Figure 5 were calculated as such, and averaged over 10 different initial conditions for every environment.

2.3 Synthetic Data Comparison

Using synthetic data the accuracy and effectiveness of our method was compared to both type I and type II model-fitting procedures. Since both types attempt to minimize $\mathbf{\Gamma}$ from Eq. 1, four different weighting schema were tested on each type. These include ordinary least squares (OLS), and Huber, Cauchy, and fair weighted robust regression. In order to follow biological data closely but enhance computation speed, the proportionate size of the transcription network was kept the same and fewer genes were analyzed. Synthetic systems consisted of 500 genes and 50 synthetic transcription factors. For the 10 real biological conditions analyzed, after processing the data as described in section 1.2, the median network proportion between genes and regulators was 9.99. The average experiment set size was 9.8 over all 10 conditions, so all synthetic data contained 10 experiments. In addition, to keep with current opinion that there is a physical limit associated with how many regulators can control gene expression from a single promoter, regulation in the synthetic data was constrained to a maximum of 3 regulators per gene, with edge densities ranging from 1.6 to 1.74 for the whole system. When *S. cerevisiae* data was analyzed under this constraint, the average transcription network edge density was 1.69.

Synthetic data was created by populating the transcription factor activity matrix $\mathbf{P}$ (Eq. 1, 50x10) with random numbers, specifying an edge density for the transcription network $\mathbf{A}$ (Eq. 1, 500x50), selecting edges at random from a pre-specified edge distribution (3:2:1, singly:doubly:triply regulated genes), populating the non-zero entries of $\mathbf{A}$ (edges) with random numbers, and generating $\mathbf{E}$ from Eq. 1 (500x10). The pre-specified edge distribution was used to approximate the real distribution found in *S. cerevisiae* which on average over the 10 experimental conditions was 2.5:1.5:1. White noise was added to produce a signal to noise (S/N) ratio of 1 or 2. The percentage of erred genes was set to either 40, 50, 60, 70 or 80%. The percentages of genes within that percentage that had erred expression was 67% and erred connectivity was 67%. Those genes specified to have erred expression had their values replaced by random numbers. Genes specified to have erred connectivity had their connectivity shuffled amongst each other, and edges shifted at random between columns. This results in an identical edge density, and an identical number of genes with 1, 2,… regulators as the true network. It was possible that genes could undergo this shuffle-shifting procedure and remain consistent, and we have accounted for that in our analysis. The network and expression, $\mathbf{A}$ and $\mathbf{E}$, were then analyzed with the Bayesian statistic used in our Gibbs algorithm (section 3.2) . This was done to determine the maximum number of regulators per gene that could be analyzed. Genes with too many regulators were omitted from the analysis, and the trimmed dataset fed to our Gibbs algorithm, and the model-fitting procedures.

For comparative purposes the number of consistent genes, cg, was determined by the Gibbs sampler. Consistent genes identified from the Gibbs algorithm were the cg genes with the highest likelihood to be consistent. From the model-fitting procedures, consistent genes were identified as the cg genes with the smallest residual error, Γ_i (Eq.1). We then determined the false positive (FP) and false negative (FN) rates for each procedures consistent gene set. These are reported in Figure 2. For every erred gene percentage and signal to noise ratio (eg. 40% erred genes at signal to noise ratio of 2) ten different synthetic networks and expression sets were analyzed. Therefore, every data point in Figure 2 is the average of 10 different, independent evaluations. In addition, due the tendency of model-fitting procedures to over fit genes with a larger number of regulators we compared performance on a per regulator basis. Again our Gibbs algorithm out performed the model-fitting procedures (results not reported).

Interestingly at a signal to noise ratio of 1, the false positive rate of 70% erred genes analyses was larger than the false positive rate from 80% erred genes analyses. This is due to two factors. The first is due to the thresholds (see Section 3.2) used by our Gibbs algorithm. As the amount of noise and percentage of erred genes increases our method has reduced resolution with genes regulated by a higher number of regulators. In the before mentioned case, three of the ten networks analyzed at signal to noise ratio of 1 and 80% erred genes could not be resolved above single regulation. If these three networks are omitted during comparison the false positive rate would be 6.6 as opposed to 6.2. The second factor stems from the scale and noise. As the noise level increases the number of consistent genes identified by our Gibbs algorithm decreases. In addition, due to the level of noise (80%) even fewer genes are consistent to begin with. As the number of consistent genes identified approaches its limit of zero it is conceivable that the false positive rate may decrease, since mis-identification at such small numbers is unlikely. This is only observed with a signal to noise ratio of 1 and 80% erred genes because all other examples provided in Figure 2 identify ~60-480% more consistent genes. Unfortunately, we could not investigate further at signal to noise ratio 1 since resolution at 90% erred genes was too often compromised to single regulation.

3 GIBBS DETAILS

3.1 Network Relationships

Our approach is based on concepts presented in (Brynildsen *et al.,* 2006). The goal of this method is to identify genes with accurate gene expression and accurate network topology (consistent genes). If accurate network topology is present, the network will require the expression of genes to conform to a set of constraints delineated by Theorem 2 of (Brynildsen *et al.,* 2006). These constraints form a series of relationships that should exist between the expres-

sions of different genes. If gene expression is accurate, it will abide by relationships dictated by accurate topology. However, if either gene expression or topology is erred constraints will be violated. Therefore, identifying relationships that are intact will provide us with a means to identify consistent genes. Following Appendix B of (Brynildsen *et al.,* 2006), we see that:

$$\mathbf{E}_1 = \mathbf{Z}_{\mathbf{A}_1}\mathbf{Z}_{\mathbf{A}_2}^{-1}\mathbf{E}_2 \qquad \text{(S2)}$$

where $\mathbf{Z}_{\mathbf{A}_2}$ (LxL) is an invertible subnetwork of $\mathbf{Z}_{\mathbf{A}}$, $\mathbf{E}_2$ (LxM) is the expression data of genes whose topology comprise $\mathbf{Z}_{\mathbf{A}_2}$, $\mathbf{Z}_{\mathbf{A}_1}$ $((N-L)xL)$ is the network of the remaining genes not within $\mathbf{Z}_{\mathbf{A}_2}$, and $\mathbf{E}_1$ $((N-L)xM)$ is the expression data of genes whose topology comprise $\mathbf{Z}_{\mathbf{A}_1}$. We use $\mathbf{Z}_{\mathbf{A}}$ here to indicate that we do not know the strength with which TFs control their targets, and that we leave these interactions unconstrained within our analysis. Eq. (S2) relates the expression of L genes in $\mathbf{E}_2$ to the expression of the remaining genes in $\mathbf{E}_1$ through the network term $\mathbf{Z}_{\mathbf{A}_1}\mathbf{Z}_{\mathbf{A}_2}^{-1}$. To test if network relationships remain intact we rearrange Eq. (S2):

$$\mathbf{0} = \begin{bmatrix}\mathbf{I} & \mathbf{Z}_{\mathbf{A}_1}\mathbf{Z}_{\mathbf{A}_2}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{E}_1\\ \mathbf{E}_2\end{bmatrix} \qquad \text{(S3)}$$

where $\mathbf{I}$ is an $(N-L)x(N-L)$ identity matrix. To satisfy Eq. (S3), for every row_i of $\begin{bmatrix}\mathbf{I} & \mathbf{Z}_{\mathbf{A}_1}\mathbf{Z}_{\mathbf{A}_2}^{-1}\end{bmatrix}$ the matrix composed of those rows of $\begin{bmatrix}\mathbf{E}_1\\ \mathbf{E}_2\end{bmatrix}$ that multiply against nonzero entries in row_i of $\begin{bmatrix}\mathbf{I} & \mathbf{Z}_{\mathbf{A}_1}\mathbf{Z}_{\mathbf{A}_2}^{-1}\end{bmatrix}$ must be rank deficient. Since gene expression data is fairly noisy rank calculations are insufficient, therefore condition numbers can be used to give a measure of how close a matrix is to rank deficiency. To determine the condition number thresholds that are indicative of rank deficiency we use a Bayesian statistic described in section 3.2.

Every gene of $\mathbf{E}_1$ is analyzed against the genes of $\mathbf{E}_2$ (seeds) to determine whether the connectivity data of the genes is consistent with their gene expression. If $\mathbf{E}_2$ were completely populated with erred genes, all genes would violate Eq. (S3) whether or not they are consistent. If $\mathbf{E}_2$ were populated with consistent genes, consistent genes would satisfy Eq. (S3). Therefore, we want genes in $\mathbf{E}_2$ to be consistent. However, we do not have any prior information pertaining to which genes would be best suited for $\mathbf{E}_2$. To determine those genes best suited for $\mathbf{E}_2$ we have devised a Gibbs sampler, which will be described in section 3.3.

3.2 Bayesian Statistic for Condition Number Threshold

We employed a Bayesian statistic to determine condition number thresholds indicative of rank deficiency for matrices of varying sizes. It takes the form of:

$$\Pr(\mathbf{X}_{rd}) = \frac{L_{rd} * prior_{rd}}{L_{rd} * prior_{rd} + L_{rf} * prior_{rf}} \qquad \text{(S4)}$$

where $\Pr(\mathbf{X}_{rd})$ is the posterior probability that $\mathbf{X}$ is rank deficient, L_{rd} is a likelihood function for rank deficient matrices, L_{rf} is a likelihood function for full rank matrices, $prior_{rd}$ is the rank deficient prior probability, and $prior_{rf}$ is the full rank prior probability.

We generated two likelihood functions, one intended to be representative of rank deficient data matrices and one intended to be representative of full rank data matrices. Since condition number is dependent on matrix size, we obtained likelihood functions for every matrix size analyzed. The likelihood functions were populated by 5000 appropriately sized matrices, and were constructed for every experimental dataset individually. It should be noted that these likelihood functions were constructed from all genes within a dataset, and not just those genes whose binding profile matched their gene expression. This will have important implications for the type of likelihood function selected.

Rank deficient likelihoods, L_{rd}, were constructed by selecting gene expression data from genes whose $\mathbf{Z}_{\mathbf{A}}$ indicated that a rank deficiency should be present. Inevitably, since genes have been included whose binding profile cannot accurately describe their expression, some matrices indicated by $\mathbf{Z}_{\mathbf{A}}$ to be rank deficient will undoubtedly be full rank. This does not concern us much since we are looking to identify rank deficiencies with our algorithm. Even if some matrices used to populate the L_{rd} are full rank it should not affect our threshold much as long as we use a high $\Pr(\mathbf{X}_{rd})$ for selection.

Full rank likelihoods, L_{rf}, were constructed by selecting at random by experiment. For instance, if we were generating a data matrix that has N genes and M experiments to populate L_{rf}, for each experiment N values would be chosen at random to inhabit the corresponding column in the NxM data matrix. This is performed independently for each experiment. We chose to use this method instead of one based off of indications from $\mathbf{Z}_{\mathbf{A}}$, because we are using the combined ChIP-chip data. It is reasonable to assume that not all potential regulators of a gene compiled over all known conditions will be acting on a promoter at a given time. Therefore, if we used our combined connectivity data to select for full rank matrices, we would more often than not end up with rank deficient matrices populating our full rank likelihood. This would severely impact the resolution with which we could identify rank deficiencies necessary to perform our algorithm. Therefore, we sought to include the experimental variability of the dataset without relying upon $\mathbf{Z}_{\mathbf{A}}$ to provide us with full rank matrices, and thus settled on our current approach.

When selecting priors for the statistic we sought to find a measure that would reflect current understanding of TF-promoter relationships. It is generally appreciated that the number of regulators that can act on a promoter at any given time is bounded, and is generally much smaller than the number of regulators known to bind the promoter. Therefore, we chose priors that identify rank deficiencies at a higher rate for matrices expected to be controlled by fewer regulators. The prior for the data to be rank deficient can be written as:

$$prior_{rd} = \left(\frac{1}{2}\right)^{reg} \qquad \text{(S5)}$$

where reg is equal to the total number of regulators supposedly acting on the promoters of genes whose expression data was used to compile the rank deficient likelihood function. For instance, a data matrix composed of expression from two genes whose only known regulator is the same would have $prior_{rd} = .5$, while a data matrix composed of expression from three genes who share 2 regulators, have $prior_{rd} = .25$. The full rank prior was simply, $prior_{rf} = 1 - prior_{rd}$.

Finally, to determine the rank deficiency threshold for matrices of varying sizes, we searched for the condition number which yielded a posterior probability, $\Pr(\mathbf{X}_{rd}) = .9$ from Eq. (S4), starting at a condition number of 1 and going up. If a given matrix size could not reach the $\Pr(\mathbf{X}_{rd}) = .9$, then we would not be able to resolve whether a rank deficiency was present. Genes with ≥ the number of regulators expected to be acting in the rank deficient matrix where could not be reached $\Pr(\mathbf{X}_{rd}) = .9$ were then omitted from analysis.

3.3 Sampling

We look to populate $\mathbf{E_2}$ with genes that when analyzed with Eq. (S3) yield the correct list of consistent genes. We begin by recognizing that if the proper genes inhabit $\mathbf{E_2}$ the number of consistent genes will be maximized. We define r as a vector of length N (N = # *genes*), populated by 0's and 1's.

$$r = \begin{bmatrix} 1, & 0, & ..., & 1 \end{bmatrix}$$

where a 1 designates a consistent gene, and 0 designates an erred gene (0 can indicate an error in expression, connectivity, or both). We postulate that the correct r contains the most 1 entries, since introduction of erred genes into $\mathbf{E_2}$ would create 0's from entries that would otherwise be 1's. Thus, if we maximize $\sum_{i=1}^{N} r_i$ we should obtain the correct r. However, r is not a simple function:

$$r = f(\eta_1, \eta_2, ..., \eta_L, \mathbf{Z_A}, \mathbf{E}) \qquad \text{(S6)}$$

$$r_i = \begin{cases} 1 & \begin{cases} cond(\mathbf{R}_i) \geq thresh_{\mathbf{R}_i} \ \& \\ cond(\mathbf{B}_i) < thresh_{\mathbf{B}_i} \ \& \\ cond(\mathbf{B}_i) < cond(\mathbf{R}_i)*.9 \\ \quad for\ i \notin [\eta_1, \eta_2, ..., \eta_L] \end{cases} \\ 0 & \begin{cases} cond(\mathbf{R}_i) < thresh_{\mathbf{R}_i}, \ \&/or \\ cond(\mathbf{B}_i) \geq thresh_{\mathbf{B}_i}, \ \&/or \\ cond(\mathbf{B}_i) \geq cond(\mathbf{R}_i)*.9 \\ \quad for\ i \notin [\eta_1, \eta_2, ..., \eta_L] \end{cases} \\ 1 & f^c(\eta_j) > 0, \ i = \eta_j \\ 0 & f^c(\eta_j) = 0, \ i = \eta_j \end{cases} \qquad \text{(S7)}$$

$$f^c(\eta_j) = \sum_{k=1}^{sj} r_{k,i}^j \qquad \text{(S8)}$$

$$r_{k,i}^j = \begin{cases} 1 & \begin{cases} cond(\mathbf{R}_k) \geq thresh_{\mathbf{R}_k} \ \& \\ cond(\mathbf{B}_k) < thresh_{\mathbf{B}_k} \ \& \\ cond(\mathbf{B}_k) < cond(\mathbf{R}_k)*.9 \ \& \\ cond(\mathbf{N}_j) < thresh_{\mathbf{N}_k} \ \& \\ cond(\mathbf{N}_j) < cond(\mathbf{R}_k)*.9 \\ \quad for\ k \in Sj,\ k \neq \eta_j \end{cases} \\ 0 & else \end{cases} \qquad \text{(S9)}$$

Here η_j is gene j of $\mathbf{E_2}$, r_i is the r value (1 or 0) for the i^{th} gene, $cond(\mathbf{X})$ is the condition number of matrix $\mathbf{X}$, $\mathbf{R}_i$ is the matrix composed of the rows of $\begin{bmatrix} \mathbf{E_1} \\ \mathbf{E_2} \end{bmatrix}$ that multiply against nonzero entries in the row of $[\mathbf{I} \quad \mathbf{Z}_{\mathbf{A_1}} \mathbf{Z}_{\mathbf{A_2}}^{-1}]$ that corresponds to the i^{th} gene, $\mathbf{B}_i$ is the same as $\mathbf{R}_i$ except that the row of $\begin{bmatrix} \mathbf{E_1} \\ \mathbf{E_2} \end{bmatrix}$ that corresponds to the i^{th} gene is substituted with a random vector (background), $threshold_{\mathbf{X}}$ is the condition number threshold for rank deficiency of a matrix the size of $\mathbf{X}$ derived from the before mentioned Bayesian statistic ($\Pr(\mathbf{X}_{rd}) = .90$), f^c is as described, Sj is the set of all possible candidates for η_j that are not already $\eta_h, h \neq j$, sj is equal to the number of genes in Sj, $r_{k,i}^j$ is a sj vector populated as described, the i and j in $r_{k,i}^j$ stand for the i^{th} gene being η_j while k refers to the k^{th} entry of Sj, and $\mathbf{N}_j$ is the same as $\mathbf{R}_k$ except that the row of $\begin{bmatrix} \mathbf{E_1} \\ \mathbf{E_2} \end{bmatrix}$ that corresponds to η_j is substituted with a random vector.

We can view $\sum_{i=1}^{N} r_i$ as our joint distribution:

$$\sum_{i=1}^{N} r_i = f_\Sigma(\eta_1, \eta_2, \eta_3, ..., \eta_L, \mathbf{Z_A}, \mathbf{E}) \qquad \text{(S10)}$$

To explore the parameter space of $\sum_{i=1}^{N} r_i$ we have chosen to employ a Gibbs sampler. Our goal is to maximize $\sum_{i=1}^{N} r_i$ by sampling from the conditional distributions:

$$\begin{gathered} f_\Sigma(\eta_1 \mid \eta_2, \eta_3, ..., \eta_L, \mathbf{Z_A}, \mathbf{E}) \\ f_\Sigma(\eta_2 \mid \eta_1, \eta_3, ..., \eta_L, \mathbf{Z_A}, \mathbf{E}) \\ \vdots \\ f_\Sigma(\eta_L \mid \eta_1, \eta_2, ..., \eta_{L-1}, \mathbf{Z_A}, \mathbf{E}) \end{gathered}$$

For each η_j there is a list of genes that are possible candidates. This list is denoted as Sj, and we strive to identify the members of Sj that will maximize $\sum_{i=1}^{N} r_i$ when positioned as η_j. Since the set of genes, $[\eta_1, \eta_2, ..., \eta_L]$, needs each regulator to partially control at least one gene (full rank of $\mathbf{A}_2$ requirement), any candidate for η_j must be regulated by TF_j. Therefore, every Sj is composed of all the genes regulated by TF_j that are not currently one of the $\eta_h, h \neq j$. This implies that a gene can be a member of multiple Sj's if it is regulated by more than one TF, which is indeed true.

For every η_j, we set all other variables ($\eta_h, h \neq j$) constant, and construct the conditional distribution, $f_\Sigma(\eta_j \mid \eta_{h \neq j}, \mathbf{Z_A}, \mathbf{E})$, by evaluating $\sum_{i=1}^{N} r_i$ once for every member of Sj as η_j. With the results for these sj evaluations we construct a probability based off of the values from a collapsed version of $f_\Sigma(\eta_j \mid \eta_{h \neq j}, \mathbf{Z_A}, \mathbf{E})$, $r_{k,i}^j$, and select the update for η_j by sampling from this probability distribution. The collapsed version of $f_\Sigma(\eta_j \mid \eta_{h \neq j}, \mathbf{Z_A}, \mathbf{E})$ includes those genes that are related to η_j through $\mathbf{Z}_{\mathbf{A_1}} \mathbf{Z}_{\mathbf{A_2}}^{-1}$. Genes whose

consistency is not impacted by η_j are omitted from $r_{k,i}^j$. The probability is constructed by:

$$\Pr\left(Sj_m = \eta_j \mid \eta_{h,h\neq j}, \mathbf{Z_A}, \mathbf{E}\right) = \frac{\sum_{k=1}^{sj} r_{k,m}^j}{\sum_{t=1}^{sj}\sum_{k=1}^{sj} r_{k,t}^j} \qquad \text{(S11)}$$

We then select results from those $[\eta_1,\eta_2,...,\eta_L]$ that yield a $\sum_{i=1}^{N} r_i$ value of at least 80% of the $\max\left(\sum_{i=1}^{N} r_i\right)$. The algorithm is said to converge once a convergence criteria based on general rank invariability of the genes has been attained. After the number of $[\eta_1,\eta_2,...,\eta_L]$ that yield a $\sum_{i=1}^{N} r_i$ value $\geq \max\left(\sum_{i=1}^{N} r_i\right)$ is greater than or equal to 1000 the sampler is tested for convergence. Convergence is tested every iteration until the convergence criteria is met or $\max\left(\sum_{i=1}^{N} r_i\right)$ changes.

3.4 Convergence

To check for convergence the set of $[\eta_1,\eta_2,...,\eta_L]$ with $\sum_{i=1}^{N} r_i$ values $\geq .8*\max\left(\sum_{i=1}^{N} r_i\right)$ is split in half, where one half (Hr) contains the most recently found $[\eta_1,\eta_2,...,\eta_L]$, and the other half (Hl) the latter $[\eta_1,\eta_2,...,\eta_L]$. Each half is then analyzed to produce Gr and Gl, which are vectors of length N populated by integer values. If $Gr_i = 11$ it means that r_i was found to be 1 in 11 $\left[\eta_1,\eta_2,...,\eta_L\right]$ from Hr, and if $Gl_i = 12$ it means that r_i was found to be 1 in 12 $\left[\eta_1,\eta_2,...,\eta_L\right]$ from Hl. The larger the value of Gr or Gl the more likely a gene is to be consistent. We then parse Gr and Gl into:

$$Gr \rightarrow Gr^0,\ Gr^{10},\ ...,\ Gr^{90} \qquad \text{(S12)}$$
$$Gl \rightarrow Gl^0,\ Gl^{10},\ ...,\ Gl^{90} \qquad \text{(S13)}$$

where Gr^X contains those genes with a Gr value $\geq (X/100)*\max(Gr)$, and the same said for Gl. In this formalism $Gr^{90} \in Gr^{80} \in ... \in Gr^0$, $Gl^{90} \in Gl^{80} \in ... \in Gl^0$, and $Gl^0 = Gr^0$ since they would include all genes. We then check:

$$CP_X = \frac{size\left(Gr^X \cap Gl^X\right)}{\min\left(size\left(\left[Gr^X, Gl^X\right]\right)\right)} \qquad \text{(S14)}$$

where CP_X is the percentage of genes in common between Gr^X and Gl^X, corrected for a possible difference in size between Gr^X and Gl^X. If all $CP_X > .9$, the Gibbs sampler is said to be converged.

REFERENCES

Brynildsen, M. P. *et al.,* (2006) Versatility and Connectivity Efficiency of Bipartite Transcription Networks. *Biophys. J.*, **91**, 2749-59.

Gasch, A. P. *et al.,* (2001) Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell.*, **12**, 2987-3003.

Gasch, A. P. *et al.,* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, **11**, 4241-57.

Harbison, C. T. *et al.,* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature.* **431**, 99-104.

Herrero, J. *et al.,* (2003) GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic. Acids. Res.*, **31**, 3461-7.

Herrero, J. *et al.,* (2004) New challenges in gene expression data analysis and the extended GEPAS. *Nucleic. Acids. Res.*, **32**, W485-91.

Lee, T. I. *et al.,* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science.* **298**, 799-804.

Lyons, T. J. *et al.,* (2000) Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. *Proc. Natl. Acad. Sci. U S A.* **97**, 7957-62.

Yoshimoto, H. *et al.,* (2002) Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in Saccharomyces cerevisiae. *J. Biol. Chem.*, **277**, 31079-88.

Table S1a: Comparison of False Positive Rates for Gibbs and Model-fitting Techniques

	40%, SN:2	50%, SN:2	60%, SN:2	70%, SN:2	80%, SN:2	40%, SN:1	50%, SN:1	60%, SN:1	70%, SN:1	80%, SN:1
Gibbs	0.31%	0.65%	1.09%	1.25%	4.36%	1.64%	1.57%	4.18%	7.55%	6.24%
Huber (II)	0.12%	0.74%	2.17%	8.96%	18.32%	1.64%	2.32%	4.74%	10.99%	19.74%
Cauchy (II)	0.16%	0.56%	1.83%	7.82%	19.43%	1.64%	2.33%	5.80%	9.69%	20.69%
Fair (II)	0.22%	0.60%	2.36%	7.56%	19.06%	1.53%	3.01%	6.23%	9.67%	19.07%
OLS (II)	0.85%	1.44%	3.05%	10.52%	21.32%	1.40%	2.94%	6.43%	10.62%	23.33%
Huber (I)	34.16%	37.94%	46.19%	51.41%	61.14%	28.00%	31.40%	40.80%	45.17%	55.02%
Cauchy (I)	34.12%	37.95%	45.64%	51.24%	61.41%	28.01%	31.36%	40.59%	45.41%	55.19%
Fair (I)	34.30%	37.91%	46.40%	50.87%	61.95%	28.07%	31.71%	41.66%	46.38%	55.30%
OLS(I)	35.92%	40.09%	48.79%	53.19%	63.58%	30.19%	35.15%	44.66%	49.25%	56.3%

Table S1b: Comparison of False Negative Rates for Gibbs and Model-fitting Techniques

	40%, SN:2	50%, SN:2	60%, SN:2	70%, SN:2	80%, SN:2	40%, SN:1	50%, SN:1	60%, SN:1	70%, SN:1	80%, SN:1
Gibbs	26.49%	22.57%	18.50%	14.50%	12.53%	43.81%	36.98%	29.83%	26.92%	19.75%
Huber (II)	26.27%	22.64%	19.08%	17.43%	15.97%	43.80%	37.27%	30.20%	27.70%	21.84%
Cauchy (II)	26.31%	22.50%	18.90%	16.99%	16.15%	43.80%	37.31%	30.63%	27.40%	21.96%
Fair (II)	26.41%	22.53%	19.17%	16.87%	16.12%	43.71%	37.59%	30.71%	27.43%	21.54%
OLS (II)	27.07%	23.17%	19.54%	17.97%	16.65%	43.66%	37.53%	30.86%	27.64%	22.09%
Huber (I)	63.40%	50.58%	42.73%	33.27%	26.33%	60.02%	49.99%	43.16%	34.98%	26.59%
Cauchy (I)	63.36%	50.58%	42.41%	33.22%	26.39%	59.99%	49.98%	43.14%	35.03%	26.60%
Fair (I)	63.59%	50.54%	42.85%	33.08%	26.53%	60.04%	50.14%	43.46%	35.23%	26.64%
OLS(I)	65.35%	52.20%	44.11%	33.93%	26.95%	61.42%	51.68%	44.56%	35.79%	26.65%

a) False positive and b) false negative rates used in Figure 2 of main text, plus those from type I model fit. Type I and type II refer to fixed and variable strength methods respectively. SN:2 and SN:1 refer to signal to noise ratios of 2 and 1.

Table S2: Comparison of ChIP-chip derived and Randomized Connectivities

	Diamide	DTT	Gamma	H_2O_2	Menadione	MMS	Salt	N_2	Zinc	Calcium
ChIP-chip, typeI	0.89	0.94	0.88	0.94	0.87	0.81	0.93	0.94	0.93	0.96
Random, typeI	0.91	0.94	0.9	0.94	0.88	0.83	0.93	0.94	0.94	0.96
ChIP-chip, typeII	0.46	0.62	0.65	0.76	0.7	0.35	0.49	0.6	0.68	0.73
Random, typeII	0.44	0.61	0.68	0.74	0.72	0.41	0.42	0.56	0.61	0.66

Specific values used in Figure 4 of main text. Type I and type II refer to fixed and variable strength methods respectively.