Nucleic Acids Research, 2005, Vol. 33, No. 2 ***605–615***
doi:10.1093/nar/gki166

Extraction of transcription regulatory signals from genome-wide DNA–protein interaction data

Yael Garten, Shai Kaplan and Yitzhak Pilpel*

Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

Received August 5, 2004; Revised November 26, 2004; Accepted December 15, 2004

ABSTRACT

Deciphering gene regulatory network architecture amounts to the identification of the regulators, conditions in which they act, genes they regulate, *cis*-acting motifs they bind, expression profiles they dictate and more complex relationships between alternative regulatory partnerships and alternative regulatory motifs that give rise to sub-modalities of expression profiles. The 'location data' in yeast is a comprehensive resource that provides transcription factor–DNA interaction information *in vivo*. Here, we provide two contributions: first, we developed means to assess the extent of noise in the location data, and consequently for extracting signals from it. Second, we couple signal extraction with better characterization of the genetic network architecture. We apply two methods for the detection of combinatorial associations between transcription factors (TFs), the integration of which provides a global map of combinatorial regulatory interactions. We discover the capacity of regulatory motifs and TF partnerships to dictate fine-tuned expression patterns of subsets of genes, which are clearly distinct from those displayed by most genes assigned to the same TF. Our findings provide carefully prioritized, high-quality assignments between regulators and regulated genes and as such should prove useful for experimental and computational biologists alike.

INTRODUCTION

Regulation of gene expression patterns represents one of the most complicated computational tasks for living cells. Often, the determination of gene expression patterns is the result of highly intricate calculations that a gene network performs on the cell's environment. The computational task is to reflect as accurately as possible a multi-component cellular environment and to integrate a diversity of inputs into a decision about the preferred level of activity of each gene at a given external condition, internal state, developmental stage, etc. A convenient way to construct such a detailed multi-dimensional transcription regulatory input function is to use multiple transcriptional regulators that, when wired by the appropriate logical gates, may provide reliable mapping of the environment onto gene expression patterns.

The modern genomic era provides opportunities to study gene network architectures in an unprecedented throughput. Comprehensive characterization of networks amounts to identifying the individual regulators, followed by identification and characterization of functional combinations they form, along with the identity of the regulated genes. In the present study, we have focused on the yeast *Saccharomyces cerevisiae* and have taken several steps toward obtaining this goal.

We departed from a recently published most valuable resource of functional genomic data (1), which consists of 113 yeast transcription factor (TF) proteins whose spectrum of promoter-binding sites was determined *in vivo* (for further detail see Materials and Methods). This so-called 'location data' thus provides a matrix of regulatory connections between the regulatory proteins in the cells and their regulated genes. Following the original publication of the location data, a gene is considered 'assigned to a TF' if the *p*-value on the interaction between the TF and the gene's promoter is below a threshold. Although most highly valuable and comprehensive, the location dataset contains an amount of noise that has yet to be determined. False positives in the location data may come at two levels. First, it is not certain that all 113 proteins examined indeed serve as TFs. Second, it is likely that some portion of the genes assigned to true regulators are not actually regulated by them (e.g. binding to regions of the chromosome in which protein binding to DNA is not followed by transactivation). We used the multiple hypotheses assessment algorithm False Discovery Rate (FDR) (2) to roughly assess the amount of noise in the location data with the strictest *p*-value originally used (p-value = 0.001). Our calculation suggests that using this

*To whom correspondence should be addressed. Tel: +972 8 9346058; Fax: +972 8 9344108; Email: pilpel@weizmann.ac.il

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

threshold, the expected amount of false TF assignments is ~18% (see details below in Results and in Materials and Methods). Indeed, our observations show that while genes associated with many of the TFs in the data show significant coherence in their mRNA expression profiles, some TFs appear to be associated with genes that show close-to-random expression profiles in a variety of conditions.

As in many other genome-scale methodologies, the ChIP–chip technology which produced the location data is likely to generate false-positive observations that may be filtered out with the aid of additional sources of information. Interestingly, such additional sources may themselves be noisy and yet serve the purpose of filtering, provided that the noise in the different methods is not trivially correlated. We hypothesized that the analysis of the following may serve for filtering noise in the location data: (i) the coherent expression of genes regulated by the TFs, (ii) the identification of regulatory motifs among the genes assigned to each TF [in similarity to other recent publications (3,4)] and (iii) combinatorial partnerships between sets of TFs. For each of the filtration methods, we present (i) the rational underlying the method, along with the relevant algorithm or computation, (ii) exemplifying figures, (iii) a 'birds-eye view' of the results of application of the method to the entire dataset and (iv) the subset of the TF-gene assignments reported by the location data, which is supported by this method (on the website). In the datasets we provide in our website, genes are assigned to TFs only if expression, sequence and/or combinatorial TF interaction support such assignments. In addition, we provide designation of the conditions in which the TFs appear to be functional. The result is not only a more reliable set of TF-to-promoter assignments, but, most crucially, higher-level genetic network wiring information.

MATERIALS AND METHODS

Location data

In vivo TF location data lists, for each TF, the promoters that are detected to be bound by it *in vivo*. This is a result of an immunoprecipitation assay in which DNA-binding proteins are allowed to bind their target sites along the genome, followed by detection of the sites bound by each protein individually. We used the data produced by Lee *et al*. (1) obtained for yeast cells grown in rich medium. The TF-promoter assignments in that data are provided in the form of a *p*-value on the hypothesis that there exists an interaction between a TF and a promoter. In all analyses reported here, we adopted the most restrictive *p*-value as suggested in the original publication (1). For the purpose of our analysis, we only used intergenic region bindings that occurred upstream of an open reading frame (ORF).

The data of Lee *et al*. was downloaded in April 2003 from: http://staffa.wi.mit.edu/cgi-bin/young_public/navframe.cgi?s=17&f=downloaddata. It provides data for 113 TFs, while the original Lee *et al*. paper included genome-wide location analysis experiments performed for only 106 yeast strains that expressed epitope-tagged regulators.

Since about a quarter of yeast genes are arranged in pairs transcribed from divergent promoters, the number of intergenic regions is considerably smaller than the number of ORFs. On the other hand, long intergenic regions were segmented in the location data chips. In total, the number of printed probes was 6756 and the number of ORFs in this dataset is 6270.

Calculating FDR

In many instances of the present analyses, we generate a multiplicity of hypotheses. We adjust *p*-value thresholds on the generated hypotheses by controlling the rate of false discovery as follows:

Let R denote the number of hypotheses rejected by a procedure.
Let V denote the number of true null hypotheses erroneously rejected (type I error).
q denotes V/R when $R > 0$ and 0 otherwise.

The FDR (or *q*-value) is the expected proportion of false positives (type I error) among the rejected hypotheses. It is given by the following FDR theorem formula (2): $\mathrm{FDR} = E(q)$.

In the context of the location data analysis, V is the number of expected false-positive binding predictions. At a given *p*-value threshold p, $V = \mathrm{p} \times$ number of hypotheses.

R is the number of predicted binding events at the current *p*-value threshold. In the location data, the probability that $R > 0$ is effectively equal to 1 at the *p*-value of 0.001. Of the 4177 TF-intergenic region assignments predicted by the location data, 763 are expected to be false discoveries (0.001 *p*-value × 113 TFs × 6756 intergenic regions = 763). Therefore, calculation of the FDR *q*-value for a *p*-value of 0.001 yields a *q*-value of 0.18, or 18% (*q*-value = 763/4177 = 0.18).

Supplementary Figure S1 shows an analysis of the effects of using various *p*-value thresholds to capture true assignments. Supplementary Figure S2 contains a thorough analysis of the trade-off between FDR and the location data *p*-value threshold. We also used positive FDR analysis (5) that changed only negligibly the *q*-value estimation (see our Supporting website). The code and further FDR analysis of the location data are available in our Supporting website.

We have fixed a permissive expected rate of false discovery of 10% in all FDR analyses in this paper.

Statistical significance of expression coherence (EC) score and TF synergies

The EC score is a measure of the extent to which a set of genes is clustered into one or more clusters in expression space, and is equal to the fraction of gene pairs in the set whose normalized Euclidean distance (between expression profiles) falls below a threshold (6).

The significance of the EC score of a set of X genes is measured by randomly sampling 10^5 sets of X genes and calculating the EC score for each sampled set. The fraction of sets that have an EC score greater than or equal to the score of the original set of genes is the *p*-value on the significance of the score (7).

Significantly, synergistic TF pairs were detected in a way similar to our original definitions (6,8). Let G1 be the set of genes assigned to TF1, and let G2 be the set of genes assigned to TF2. G12 is the set of genes in the intersection of G1 and G2, i.e. the set of genes assigned to both TF1 and TF2. We define and calculate the 'intersection set EC score' as the EC

score of *N* genes in G12. We then randomly sample 1000 sets of *N* genes from G1, and also 1000 sets of *N* genes from G2, and calculate their EC scores. A pair of TFs is synergistic if its intersection set EC score is at the top 5% of each of the two distributions of random EC scores. The use of a relatively permissive 5% threshold is justified for two reasons. First, each of the two sets of genes already has a relatively high EC score, since each set is bound by a regulating TF. Therefore, a subset of these coherent sets, which is at their top fifth percentile in EC score, is even more significant. Second, since the score of the intersection set, in synergistic pairs, is at the top of both gene sets' distribution of scores, the effective significance of two events with *p*-value of 0.05 should typically be even better than 0.05.

mRNA expression data

Whole-genome mRNA expression data of 40 time series in yeast were obtained from ExpressDB (9). These time series represent a wide range of natural (e.g. cell cycle) (10–12) and perturbed (13–17) conditions. Detailed description of all analyzed conditions is available on our Supporting website.

Clustering parameters

In clustering the expression profiles of the genes assigned to each TF using the QT_clust algorithm in each condition, we explored various thresholds of cluster diameters, defined as the maximal Euclidean distance between any two genes within a cluster. The diameters were chosen as follows: for each condition, all pairwise distances between each pair of genes measured were calculated, yielding a distribution of pairwise distances. The 6 diameter thresholds we explored were the distances associated with the top 5, 10, 20, 30, 40 and 50 percentile of this distance distribution. An adaptive clustering algorithm was recently developed for similar purposes, see Supporting website for application of this algorithm (18).

RESULTS

The *p*-value trade-off in the original location data

The location data assigns a *p*-value on the hypothesis that TF X binds to promoter Y and thus contains *p*-values for a set of multiple hypotheses. In order to determine which hypothesis is true, a *p*-value threshold is selected and only those hypotheses that pass this threshold are assumed to be correct. This thresholding, while probably capturing true assignments of TFs to promoters, results in a yet-to-be-determined amount of false assignments. We started by statistical assessment of the FDR in the location data with the strictest *p*-value used by its authors (*p*-value = 0.001). Our calculations suggest that using this threshold the expected amount of false assignments of intergenic regions to TFs is 763 of the 4177 assignments (i.e. the FDR, known as *q*-value, equals 0.18), hence there are 3414 expected true positives (4177 − 763 = 3414). For details on calculations, see Materials and Methods. Without additional sources of information, it is thus impossible to establish which of the added hypotheses are likely to be true TF-promoter assignments.

With the goal of a cleaner version of the location data in mind, which will allow better deciphering of the genetic regulatory map, we applied a number of methods which each produces a matrix, identical in size to that of the location data, of regulatory connections between the regulators in the cells and their regulated genes. These matrices were then used in combination to produce the noise-filtered version of the location data. In each matrix, a gene *i* was marked as regulated by TF *j* if and only if it was (i) assigned to TF *j* in the original location data and (ii) it also had evidence strengthening this assignment as found by at least one of the following methods of detection: (a) clustering of gene expression profiles, (b) regulatory motif detection, (c) synergy interactions or (d) co-localization of TFs.

Method (i): Decomposition of expression profiles of the regulated genes

We begin by inspection of the mRNA expression profiles of genes associated by the location dataset to each of the 113 TFs in a diverse set of 40 conditions. Here and in all subsequent analyses we refer to a whole time series (such as exposure to heat shock or progression through the cell cycle) as a 'condition', which is composed of 3–28 time points. An intuitive expectation from a set of genes that are indeed regulated by a shared TF is that they display similar expression profiles at least in the conditions in which the TF exerts a significant regulatory effect. Yet, we need not necessarily anticipate one coherent cluster; an alternative may be that some TFs will give rise to several distinct expression patterns. The EC score is thus a suitable measure of the extent to which a set of genes is clustered into one or more clusters in expression space (for definition of EC score, see Materials and Methods). We explored various thresholds that correspond to different extents of expression similarities that may be dictated by various regulators.

We have examined the expression profiles of the genes assigned to each TF in the location data in 40 time-series experiments that span a broad range of natural (10–12) and perturbed (13–17) conditions. We performed EC analyses (6,7) on each gene set in each condition and evaluated their statistical significance using our recently proposed formalism (7). We used the FDR theorem (2) to account for the multiplicity of hypotheses tested and determined a *p*-value threshold that guaranteed a desired FDR.

Figure 1A is a matrix depicting significant EC of particular TFs in particular conditions (for details on statistical significance of EC score, see Materials and Methods). We assume that a TF regulates the gene set assigned to it in the location data in a particular condition if these genes are significantly coherent in that condition. Figure 1B and C depicts distributions of the number of TFs regulating each condition and number of conditions regulated by each TF, respectively. The conditions that are controlled by the largest number of TFs are the Cho cell-cycle experiment (10,12), the MAPK signaling experiment (11) and the nitrogen depletion experiment (13). These conditions are subject to the regulation of 30–34 TFs. The two ribosomal protein regulators, Rap1 and Fhl1, show regulation in many of the conditions. Conversely, 16 out of the 113 TFs in the dataset, which had three or more genes assigned to them, had no condition in which the genes assigned to them show significant coherence. Some of these TFs may be involved in AND-gated combinatorial regulation, and only when inspecting them along with their partners may coherence emerge. Alternatively, it may be that such TFs represent multiple

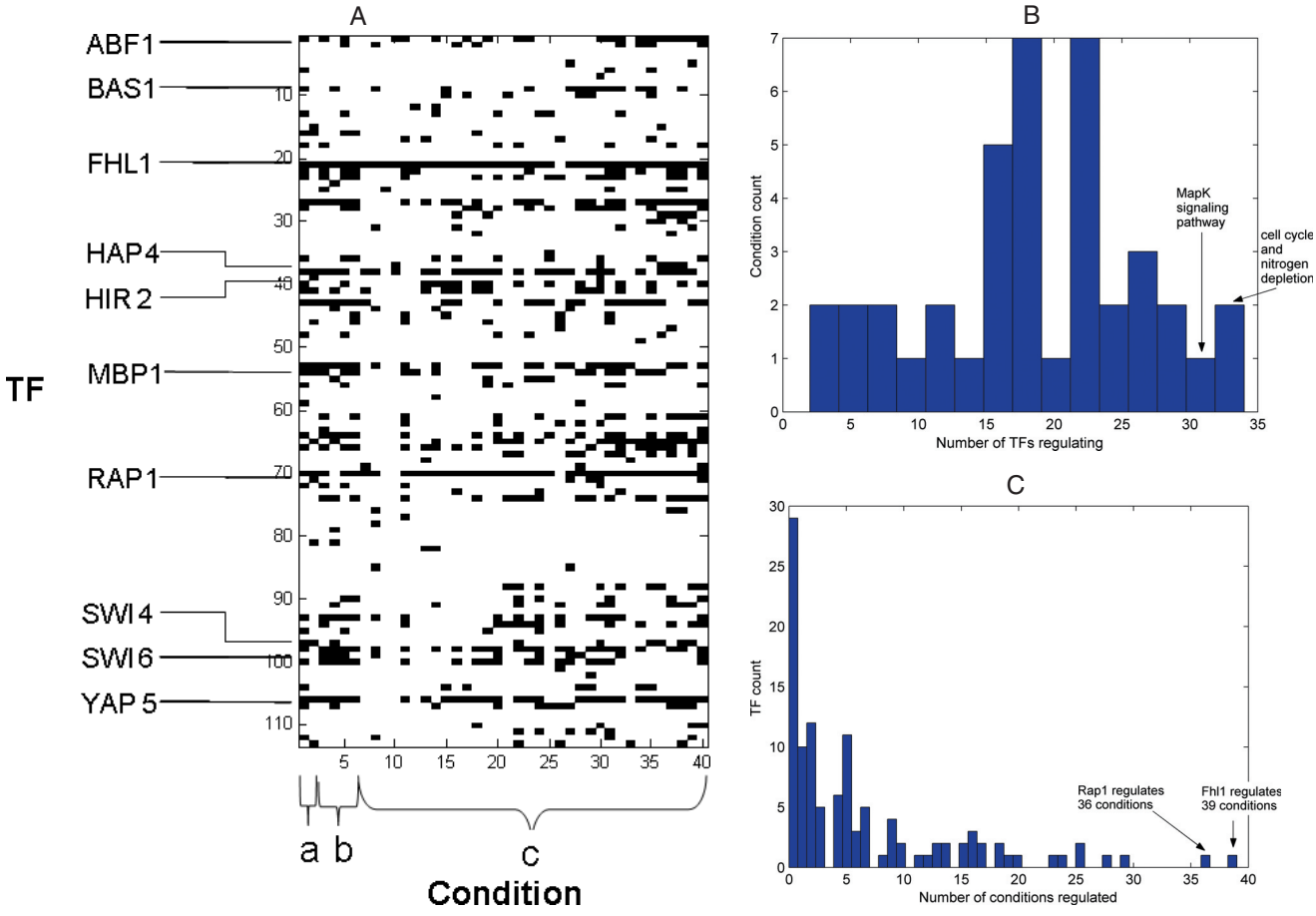

Figure 1. (**A**) A matrix depicting EC of each TF in each condition. An *ij*-th entry in the matrix is colored black if the *i*-th TF was significantly coherent in the *j*-th condition, and white otherwise. Conditions marked as 'a' are Cho's cell cycle (12) and Chu's sporulation (16), in 'b' are Spellman's four cell-cycle conditions (10) and in 'c' are predominantly stress responses (13–15,17). Selected TFs are designated by their names; all TF and condition names, along with textual version of the matrix, are available online. (**B**) A histogram with the number of TFs regulating each condition. (**C**) A histogram depicting the number of conditions regulated by each TF.

cases of false TF-promoter assignments. It is also possible that some of the low-scoring TFs are in fact regulating conditions not examined here.

The EC score was deliberately designed such that TFs that predominantly give rise to one or a few tight clusters of genes (when clustered by expression profiles) can score highly, while a significant amount of genes with no clear cluster-assignment may be tolerated. In order to detect such behaviors, we have subjected the mRNA expression profiles of genes assigned to each of the TFs to decomposition by the QT_clust clustering algorithm (19). Unlike many clustering algorithms, such as k-means, that require in advance the determination of number of clusters and that give rise to clusters of various extents of tightness (20), in this algorithm the only input is the minimal cluster tightness, and the output is the number of clusters along with the gene-cluster assignments. In all present analyses we used a relative, rather than absolute, measure of cluster tightness. The distance between each two genes in a cluster was required to be lower than a distance D, such that the probability of two random genes from that experiment to be at distance D or lower was p. For each TF, we experimented with a range of values of p, from 0.05 to 0.5. Figure 2A shows the result of running QT_clust over the set of genes assigned to Abf1 using Chu's sporulation expression data (16). This is a clear example of a TF whose associated genes display various different expression patterns (colors of expression profiles are only relevant later in the text). Our analyses show such situations where the genes regulated by a TF may be decomposed into several distinct expression profiles, even in conditions in which the genes assigned to the TF are significantly coherent. For example, in only 288 of the 738 cases in which a TF scored highly at a condition, the most populated cluster is at least three times larger than the second largest cluster; the rest of the TFs represent cases in which the genes assigned to the TF give rise to several sizeable well-separated clusters.

Clustering of the expression profiles often yields several major clusters that are highly populated and have a clear, distinct expression pattern, and many more clusters that are lowly populated, often containing only one or few genes whose expression profile was dissimilar from that of all other genes. Additionally, it is possible and even likely that the lowly populated clusters consist of genes that were misassigned to the TF, since they have a profile so different than that of the genes that appear to be tightly regulated by the TF. Such a decomposition of the expression signal allows us to view the genes assigned to each TF and distinguish the signal from noise in the data. For example, refer again to Figure 2A. It seems that this TF gives rise to several different temporal patterns. In addition, 8 out of 28 of the clusters not shown contain only one or two genes. Thus, expression-based data

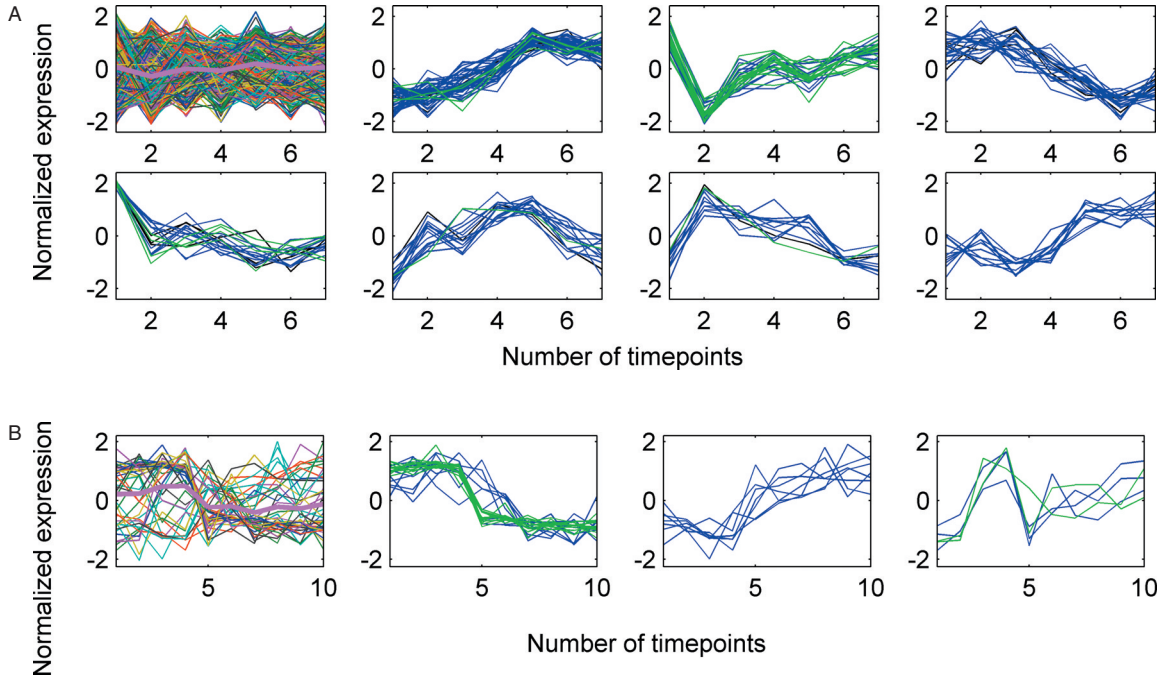

Figure 2. Expression profiles of genes regulated by Abf1 during sporulation (**A**) and Bas1 during nitrogen depletion (**B**). The first box on the left in each panel represents the expression profiles of all the genes assigned by the location data to the respective TFs. The rest of the boxes represent the results of decomposition of these genes into the most populated clusters generated by QT_clust. In (A), genes containing NCGTNNNNARTGAT and CGATGAGMTK are colored green, genes with only the first motif are colored blue, genes with only the second motif are colored red and genes with none of the motifs are black. In (B), genes containing the RNMRGAGTCA motif in their promoter are colored green, the rest are blue.

cannot support the proposed assignments of these genes to the corresponding TFs. On the other hand, the genes in the substantially populated clusters are the most likely true assignments, and recalculating the EC score of these genes alone may show that the TFs do in fact give rise to significantly coherent expression profiles, which were undetectable amidst the noise.

Hence, filtration method (i) yields a matrix in which gene *i* is assigned to TF *j* only if it was assigned in the original location data, and also if it belongs to a cluster of at least three genes, in at least one of the conditions in which the set of genes assigned to TF *j* is significantly coherent. This choice of minimal cluster size reflects a balance between the desire to include as many assignments as possible and the tendency to remove those that seem to be outliers (see our Supporting website).

Following this filtration, the EC score of each TF's genes was recalculated. Of the TF-condition pairs that did not pass the EC significance test on the original data, 96 TF-condition pairs were significantly coherent after this filtration.

We examined the relationship between the results of clustering the genes assigned to a TF, in multiple experimental conditions. For each TF, we calculated the number of common genes in the largest cluster, in all pairs of conditions in which the EC score of that TF was significantly coherent. Supplementary Figure S3 shows a plot of the distribution of these pairwise overlaps, which are significantly larger than that expected at random. Thus, the sets of coherent genes of the same TF in different conditions significantly overlap.

It is noteworthy that the location data was generated from yeast grown in rich medium, a growth condition quite different from many of the conditions for which we have expression data. Yet, our analysis shows that often genes associated with many of the TFs display good coherence in multiple, diverse conditions. This may be taken to indicate that the TF is localized in the vicinity of its binding site, perhaps somewhat statically, and some additional modifications may render it active in the appropriate condition.

We next turned to investigate reasons underlying the existence of one or several large tightly controlled clusters of genes for each TF. In order to detect such behaviors, we inspected the results of the clustering of the mRNA expression profiles of genes assigned to each of the TFs. In order to understand what may be responsible for a unique expression pattern of a subset of the genes assigned to the TF, we turned first to analyze regulatory sequence motifs.

Method (ii): From location data to regulatory motifs

We used AlignACE (21), a Gibbs sampler that searches for over-represented motifs in a set of DNA sequences, to derive regulatory motifs from promoters of sets of genes assigned to each TF in the location data. We identified a total of 567 significant motifs for 61 of the TFs. (A significant motif is one that had a MAP score >10, as proposed in the original AlignACE paper (21), and a group specificity score $<10^{-6}$. In addition, we required that the ratio between the number of consecutive gaps and the nucleotides in the consensus sequence be $\leqslant 0.4$, a threshold that reconciles removal of false motifs with maximization of the number of TFs for which motifs are derived.) We then turned to recalculate the EC score of genes assigned to each of these TFs, this time considering only a subset of these genes, namely the ones that contain the significant motif in their promoter. For each TF for which a motif was found, we compared

each such EC score with a distribution of 10 000 EC scores of random samples of genes assigned to the TF, but that do not necessarily contain the motif. The sample size of each such random gene set was the number of genes assigned to the TF that also contained the motif in their promoter. We say that the motif improves the EC score of the TF in a given condition if its EC score is at the top 5% of the random scores distribution for that condition. Of the 738 TF-condition pairs shown as significant in Figure 1A across all 113 TFs, 641 pairs represent 61 TFs for which we found significant motifs. For 421 out of the 641 TF-condition pairs, we obtained a motif that significantly improves the EC score. In these cases, we hypothesize that the genes that contain a motif and belong to the cluster it dictates are the more likely targets of the TF.

An example of such behavior is the histidine and adenine biosynthesis regulator Bas1 that gives rise to incoherent expression profiles in the nitrogen depletion condition (Figure 2B) (13). Yet, a motif we discovered by AlignACE, using the promoters assigned to this TF, whose consensus is RNMRGAGTCA (MAP score 24, group specificity score 3.8×10^{-10}), is most highly over-represented in only one of the two major clusters of this TF. This motif is highly similar to a motif experimentally shown to be bound by Bas1 (22). While it is still possible that some of the genes in the other clusters are also targets of Bas1, by reassigning to this regulator only the genes that contain the motif found, we may have filtered a significant amount of false assignments.

Another interesting behavior is displayed by the genes assigned to the chromatin remodeling factor, Abf1 (see Figure 2A). AlignACE run on the promoters of 282 genes assigned to this TF resulted in two regulatory motifs: NCGTNNNNARTGAT (MAP score 390, group specificity score 1.6×10^{-98}) that occurs in 262 of the TF targets and CGATGAGNTK (MAP score 26, group specificity score 9.9×10^{-6}) that occurs in 37 of the targets. The latter motif is also known as the PAC motif, whose binding TF remains elusive (23). All of the 37 genes that contain the second motif in their promoters contain the first as well, and a possible interpretation is that these genes are under the regulation of at least two TFs. Interestingly, while a significant portion of the genes that have both motifs (Figure 2A, green) co-cluster in the sporulation condition shown here, across many conditions they display more complex behavior (data not shown) that probably reflects condition-dependent dominance of either of the motifs.

When examining the significant motifs found, it is important to bear in mind that not all of the genes assigned a TF in the location data contain the significant motif found for that TF. The average ratio between the number of genes containing the motif and the number of genes assigned to the TF is ~39% (for further detail see Supporting website). This may indicate the level of noise in the data, although alternative motif-searching algorithms may change the exact picture.

The significant motifs discovered gave rise to a matrix, in which gene *i* is assigned to TF *j* only if it was assigned in the original location data, and also the promoter of gene *i* contained a significant motif which was found for TF *j*. This matrix portrays filtration method (ii).

Method (iii): Synergistic interactions between TFs and a synergy map of the yeast TF combinatorial network

Figure 3A shows the expression profiles of the genes assigned to the regulator Ndd1 in the Carbon-1 medium in the environmental stress experiment (13). Here, again the expression of

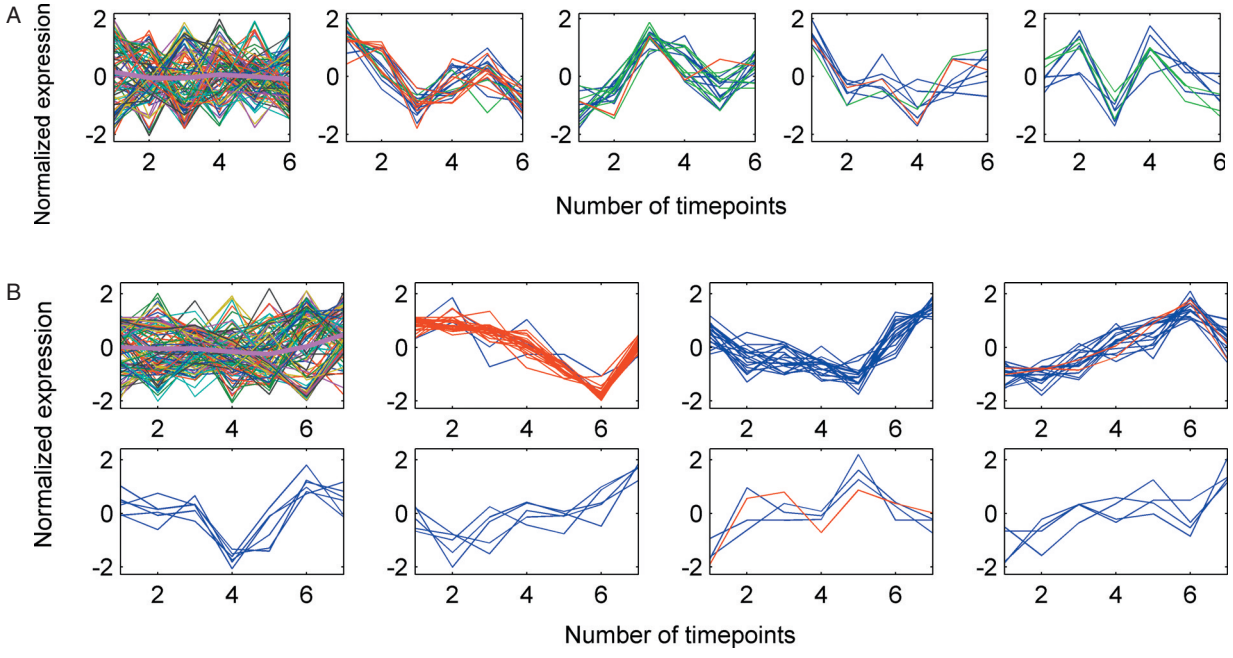

Figure 3. Genes assigned to Ndd1 in the Carbon-1 medium in the environmental stress experiment (**A**) and Yap5 during exposure to the reducing agent dtt (**B**) (13), with the same QT_clust-based clustering as in Figure 2. In (A), genes that are assigned to Ndd1 and Swi5 are colored red, while genes that are assigned to Ndd1 and Mcm1 are colored green. Genes assigned to Ndd1, but not to Swi5 and not to Mcm1 are colored blue. In (B), genes that are assigned also to Fhl1 are colored red, while genes only assigned to Yap5 are blue.

these genes is not coherent, yet the clustering shows that the gene expression profiles may be decomposed predominantly into two coherent clusters. Interestingly, we have identified two TFs, Swi5 and Mcm1, such that half of the genes bound by both Ndd1 and Swi5 fall in the largest cluster, and over a quarter of the genes bound by both Ndd1 and Mcm1 fall in the second largest cluster (see Figure 3A). It thus appears that with alternative partners Ndd1 may participate in regulation of completely different responses. Interestingly, all three regulators in this set are known as cell-cycle regulators, yet we provide here an indication that they are involved in the regulation of the response to nitrogen depletion, a process that evokes meiosis in yeast. This is another demonstration (6,24) of the extensive regulatory cross-talk between the meiotic and mitotic cell-division processes.

In the detection of such combinatorial interactions between regulatory motifs, we have previously defined motif synergy (6,8). A pair of regulatory motifs is considered synergistic if the EC score of the genes containing the two motifs together is significantly higher than that of the genes that contain either of the motifs alone. Zhang and Banerjee (25) have recently adopted this definition and explored synergistic interaction in the location data during cell cycle. We report here the detection of synergistic interactions among all pairs of TFs in the location data, in each of the above 40 conditions. A pair of TFs is considered synergistic if the EC score of the genes assigned to both TFs is significantly higher than that of the genes assigned to either of the TFs alone. We used our previous statistical formalism for calculating a *p*-value on the hypothesis that two TFs are synergistic (6,8). (For details on calculation of significance of synergy see Materials and Methods.) For each condition, we derived a list of all pairs of TFs which are synergistic in that condition. This resulted in a total of 279 unique significant synergistic interactions across all 40 conditions.

An example of two synergistic TFs is given in Figure 3B, which shows the expression profiles of the genes assigned to the regulator Yap5 during exposure to the reducing agent dtt (13). The genes that are also assigned to Fhl1 are colored red, and appear predominantly in the largest cluster. Thus, it appears that when the promoter of a gene is bound by both Yap5 and Fhl1, the expression profile of this gene is likely to be very specific, and distinct from the expression profile of genes bound by Yap5 alone.

Figure 4A is a graph depicting all significant synergistic interactions of one of the 40 conditions, namely exposure to the dtt reducing agent (13), in which synergism between the two TFs described in Figure 3B is highlighted. Similar maps in additional conditions, in addition to a combined map of all conditions, appear on the website.

Synergistic interactions provide us with strengthened evidence of a true regulatory interaction. Thus, the data of synergistic interactions produces the matrix of filtration method (iii). In this matrix, gene *i* is assigned to TF *j* only if it was assigned in the original location data, and also was assigned to another TF that shows synergy with TF *j*.

Method (iv): Co-localization of TFs in shared promoters

Another means to detect interactions between regulatory proteins, which does not involve expression data, is to detect their degree of co-localization in shared promoters. Two TFs are said to co-localize if they are shown in the location data to bind to the same promoter. Significant co-localization describes cases in which the number of promoters assigned to the two TFs is significantly large, given the number of promoters assigned to each TF alone. The basic premise here is that if two or more TFs co-localize in a significantly high number of gene promoters, the genes in which the TFs co-localize are more likely to be true targets of the respective TFs compared with genes that are associated with each TF alone. However, we note that high rate of co-localization of two TFs does not necessarily imply temporal co-localization, namely it may be that the two TFs are bound to the promoter in different conditions, perhaps even in a mutually exclusive manner.

Analogous to our previous motif co-occurrence calculation (8), we consider two TFs, TF1 and TF2, as potentially functionally interacting if the number of promoters in which they co-localize is significantly high considering the number of promoters assigned to each of them individually. To test the null-hypothesis that the observed or higher rate of co-localization of two TFs could be obtained by chance given the above priors, we use the cumulative hyper-geometric probability distribution.

$$P(X \geqslant \mathrm{tf12}) = \sum_{i=\mathrm{tf12}}^{\min(\mathrm{tf1},\,\mathrm{tf2})} \frac{\binom{\mathrm{tf1}}{i}\binom{g-\mathrm{tf1}}{\mathrm{tf2}-i}}{\binom{g}{\mathrm{tf2}}}$$

where g is the number of promoters in the genome, tf1 and tf2 are the number of promoters assigned to TF1 and TF2, respectively, and tf12 is the number of promoters assigned to both TF1 and TF2.

We have generated a graph of all pairwise interactions in the location dataset (Figure 4B). While co-occurrence analysis of regulatory motifs was introduced before (8), we now provide an analysis at the level of the TFs themselves and show that most such interactions occur within one highly connected graph. The nodes of the graph correspond to TFs, and edges connect between pairs of TFs if the *p*-value on the hypothesis that they significantly co-localize falls below a determined threshold. For clarity of the graph, and due to the high number of significant co-localizations, we set a *p*-value threshold of 10^{-10}. The graph displays several interesting properties. Coloring the graph according to the biological function ascribed to each TF, we discover clustering of TFs according to their annotated function. (For details on derivation of biological functions, see legend of Figure 4B.) In particular, we discern a highly connected cluster of cell-cycle regulatory TFs (see cluster I in Figure 4B). This observation is similar to the one we have initially made with cell-cycle regulatory motifs yet with a completely different criterion for regulatory interactions (6). This is another clear indication that the cell-cycle is one of the most tightly controlled processes in yeast, and that an intricate network of regulators is at work in its regulation. The map shows two other clusters that are also rather homogenous in terms of the functions of the TFs they contain. This clustering by function suggests, as in other biological networks (26), that a 'guilt-by-association' approach may be used for annotating the regulatory role of poorly characterized TFs by ascribing

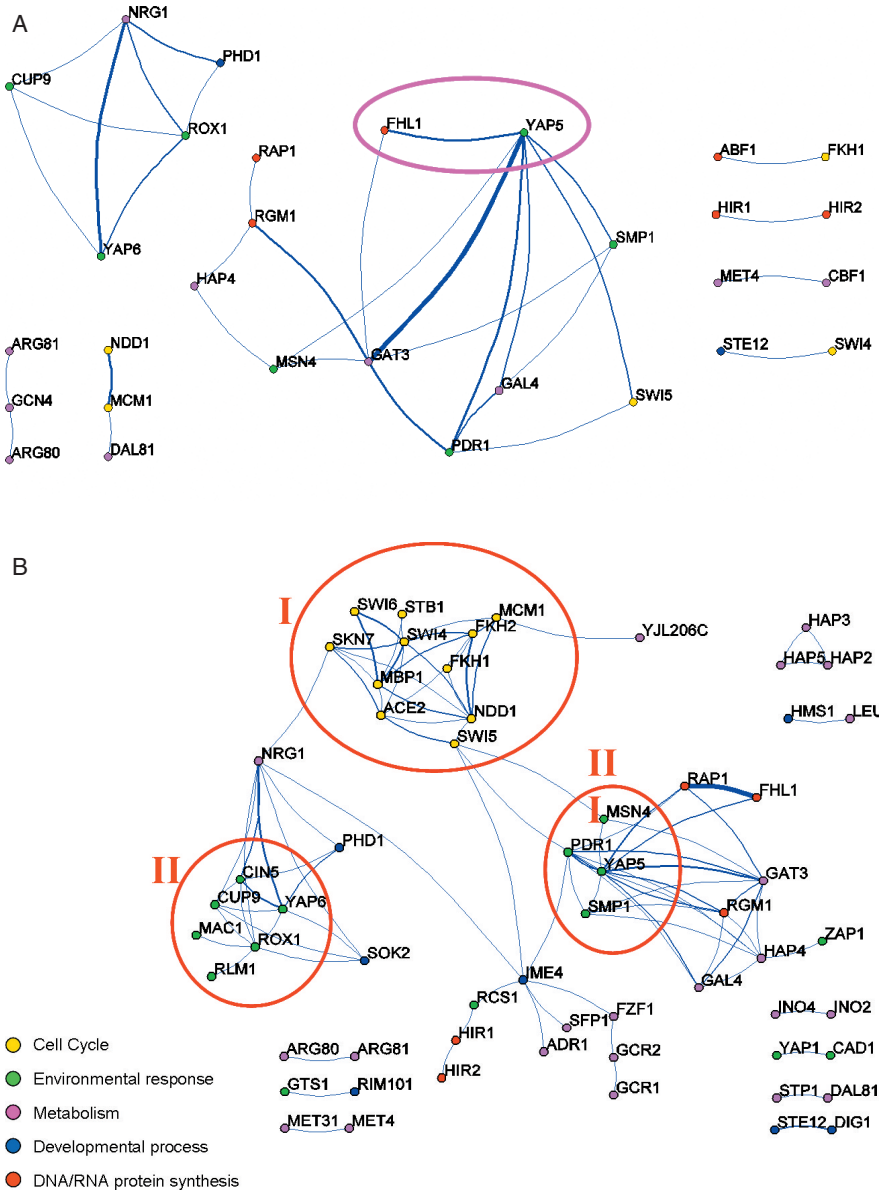

Figure 4. Graphs depicting TF synergy during exposure to the reducing agent dtt (**A**), and co-localization (**B**). The nodes in the maps represent TFs, an edge between two nodes represents significant synergy in (A), and significant (p-value < 10^{-10}) co-localization in (B), between the two corresponding TFs. Graph rendering was performed with Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.htm). Two nodes that are analyzed in detail in Figure 3B, that correspond to Yap5 and Fhl1, are highlighted in (A). The three main clusters of co-localized TFs are circled red in (B). Nodes in (B) are colored according to the regulatory function of the TFs. Such functions were annotated in (1) according to the biological function of genes assigned to the TF. Width of lines connecting two TFs reflects the number of genes assigned to both TFs; size of node reflects the number of genes assigned to the TF.

them the role of their annotated partners, if they occur in such 'functionally coherent' clusters.

Significant co-localization interactions provide us with strengthened evidence of a true regulatory interaction, and thus this data produces the matrix of filtration method (iv). In this matrix, gene i is assigned to TF j only if it was assigned in the original location data, and also was assigned to another TF which co-localizes significantly with TF j.

Relationship among the four methods of filtration

The four filtration methods discussed here each served to produce a higher quality matrix of TF-gene interactions. The numbers of interactions predicted by the single methods are 4044, 2795, 2313 and 2418 for clustering of gene expression profiles followed by filtration of lowly populated clusters, motif detection, synergy and co-localization analysis, respectively. Altogether, 1487 interactions were predicted by all four filtration methods presented in this study. Figure 5A shows the number of TF-gene assignments supported by each unique combination of methods. Supplementary Figure S4A and B show in each of the four filtration methods, and in their union and intersection, per TF, the percent and absolute numbers of gene assignments not supported by the method, relative to the total number of genes assigned to the TF in the location data.

Figure 5B shows an analysis of three of these methods: motifs, synergy and co-localization. Each of the three methods

A

# assignments		coherence	motifs	synergy	colocalized
AND[a]	OR[b]				
648	648	■			
49	49		■		
3	3			■	
63	63				■
1009	1706	■	■		
23	674	■		■	
106	817	■			■
0	52		■	■	
8	120		■		■
67	133			■	■
124	1856	■	■	■	
78	1961	■	■		■
569	1479	■		■	■
40	230		■	■	■
1487	4274	■	■	■	■

[a] Relationship between shaded methods is that of 'AND'. Number reports the interactions which are supported by all marked methods (but not by unmarked methods).

[b] Relationship between shaded methods is that of 'OR'. Number reports the interactions which are supported by any of the marked methods (but not by unmarked methods).

B

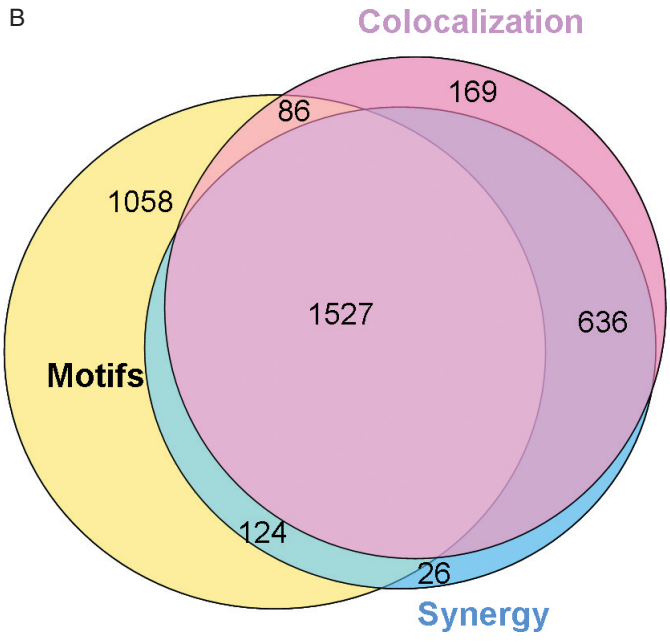

Figure 5. (**A**) Table displaying number of TF-gene assignments supported by all possible combinations of methods. For each combination marked by gray boxes, the number of assignments supported by this unique combination of methods is reported. The first column reports the number of assignments supported by all of the methods marked in each row (an 'AND' relationship between the methods), while the second column reports assignments supported by any of the methods (an 'OR' relationship). For example, the 5th row reports that there are 1009 interactions supported by both the coherence and the motif method (and not supported by the other methods), and 1706 interactions supported by either the coherence or the motif method (and not supported by the other methods). For further details, see Supporting website. (**B**) Venn diagram depicting the relationships among the TF-gene interaction predictions of three methods of filtration: motif detection, synergy and co-localization. A total of 3626 unique interactions were predicted by at least one of the three methods, and 1527 interactions were predicted by all three methods.

utilizes a different type of data source, sequence analysis, expression data or statistical analysis of common gene sets. The Venn diagram portrays the relationship between the cases of TF-gene interactions and the methods that predict each interaction. For instance, there is a significant overlap between those interactions predicted by the synergy method, with the interactions predicted by the co-localization method. In addition, there are a significantly large number of TF-gene interactions, which were predicted by all three methods. A total of 3626 unique TF-gene interactions were predicted by the three methods.

Finally, we consider the matrices resulting from each filtration method as part of a more global prioritization scheme. On one extreme, the 1487 predictions supported by all four methods represent the highest-quality set of interactions. Nevertheless, this set has the lowest coverage. The union of all four methods lies on the other end of the scale, and predicts 4274 interactions (all but 159 of the original TF-gene assignments in the location data). Between these two extremes are TF-gene assignments that are supported by various subsets of the filters. We have implemented a relatively simple prioritization scheme, offered on the Supporting website that ranks assignments based on the number of filters supporting them. In the future, more sophisticated means will be offered that prioritize predictions according to the confidence of filter-specific scores supporting each assignment and partial dependencies between the different filters.

DISCUSSION

Functional genomics provides bioinformatics with genome- and proteome-wide data with an unprecedented throughput. Yet, the optimal utilization of these data sources requires establishing efficient means to assess the extent of noise in the data, and potentially also to filter it out. It is desired that in parallel to technological improvements on the experimental side that will reduce the noise level, accompanying computational tools will be developed to provide noise-filtration. It is likely that such tools will have to involve data from other sources (that themselves may be noisy as well). In the current work, we achieve exactly that. By combining the location data with promoter sequence data and extensive information on mRNA expression, we have significantly improved the accuracy of the DNA–protein location data and, through this process, have gained new insights on gene network design principles. An approach developed recently by Bar-Joseph *et al.* (27) is most useful for adopting a more permissive *p*-value threshold on TF-gene assignments in the location data in order to reduce false negatives, when TF combinatorics and expression data support it. On the other hand, our methods are mainly aimed at removing false positives. In that respect, the two approaches are complementary to each other. Another method that prioritized TF-promoter interactions based on the location and expression data was that of Gao *et al.* (3). We have thus performed comparative analysis that gauged the extent of overlap between TF-gene assignments supported by the three studies, using in our study the intersection of assignments derived from all four filters (see Supplementary Figure 5). We found that the three studies produce significantly overlapping sets of assignments, yet each study identifies

unique assignments that the other studies do not provide support for. Among the possible reasons for lack of congruence are Bar-Joseph's algorithm's sensitivity toward assignments with higher than 0.001 *p*-value, our explicit reliance on TF synergies, co-localization and sequence motifs, and Gao's emphasis on contribution of the TF to the expression fold change of regulated genes at individual time points (as opposed to effect across an entire time series).

It should be noted that while we have considerable confidence in TF-gene assignments supported by at least one of the four filters presented here, it is entirely possible that additional filters may be proposed that would support additional such assignments. Such filters may include functional annotations or genome-wide transcription response to deletion of TFs. In addition, we stress that supporting evidence for assignments in this work is mainly proposed for cases in which the DNA–protein interaction data show regulatory effect on gene expression. It is possible that some assignments represent true binding events that resulted in no detectable transcriptional effects.

In this study, we used motif-finding algorithms for binding site predictions and microarray expression data, both of which are noisy techniques which will be further refined in the future. However, as long as the situation of noisy genome-wide technologies prevails, the course of action must be cleaning of one noisy data by intersection with other, potentially noisy, data sources. This is of course legitimate only in cases where there is no correlation between noise in the different technologies, and there is no reason to assume that noise in expression, sequence and location data should be correlated. Thus, our final products are rigorously statistically prioritized observations for which support comes independently from multiple sources that each by itself may be noisy, yet their concurrence is unlikely by chance.

In the present analysis, subsets of co-expressed genes assigned to a TF are considered true positives even if they display expression profiles completely un-correlated with that of the TF itself. This reflects the notions that not all TFs vary at the mRNA expression level, that TF-gene interaction may include negative effects, and that various logical interactions may be used to combine multiple regulators. While the basic building blocks of transcription regulatory networks are the TFs and the regulatory motifs they bind, this work also provides the next level in gene network deciphering, namely TF-motif combinations. We provided here two largely independent methods for TF interaction, and were encouraged to realize that a sizable number of predictions are in the intersection of the two methods. Proposed interactions that are supported by the two methods constitute our highest quality predictions.

Combinatorial interactions among multiple regulators provide organisms with exponentially growing computational capacity, as well as the potential to respond to their multi-dimensional complex environment. These responses control the level of activity of each gene in the genome. In the present analysis, TF combinatorics plays a dual role. On the technical level, it serves to clean the location data. On the biological level, the discovery (and rediscovery) of combinatorial interactions constitute a crucial step toward full deciphering of the architecture of gene regulatory networks. Might it be that in addition to the role of TF combinatorics in representation of the multi-dimensional cellular environment, they are also employed by biology itself for the task of noise-filtering? Since the DNA-binding sites of most TFs are relatively short [5–20 bp (28)], their specificity toward their actual sites, which reduces sharply with increased genome size, is very low even for small genomes, such as yeast's. A potential solution could be perhaps to increase the size of the individual DNA-binding sites of TFs, but this would probably require a complete redesign of their protein folds. The obvious alternative is to employ simple AND-gated combinatorics, of homo- or hetero-TF combinations, in order to filter out genes that are bound by individual TFs but should not be regulated by them.

Supporting website

Supporting website is available at http://longitude.weizmann.ac.il/TFLocation/TFLocation.html.

Included in the website is a Matlab GUI that allows exploration of the expression profiles of TFs in multiple conditions, detection of combinatorial interactions among them, and effect of regulatory motif on their coherence patterns. In addition, the website provides our noise-filtered version of the location database and various interactive means for user-defined filtering strategies.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank all members of the Pilpel laboratory for many stimulating discussions, and H. Benjamin, G. Getz, M. Lapidot, A. Mitchell, D. Peer, R. Sharan and I. Simon for critically reviewing the manuscript. Y.P. is an incumbent of the Rothstein Career Development Chair in Genetic Diseases and is a Fellow of the Hurwitz Foundation for Complexity Sciences. We are grateful to the Israeli Science Foundation, the Minerva foundation, the Ben May Foundation, and the Leo and Julia Forchheimer Center for Molecular Genetics for grant support. Funding to pay the Open Access publication charges for this article was provided by The Israeli Science Foundation.

REFERENCES

1. 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) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
2. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.
3. Gao,F., Foat,B.C. and Bussemaker,H.J. (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.
4. Segal,E., Yelensky,R. and Koller,D. (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19**, i273–i282.
5. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
6. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet*, **29**, 153–159.

7. Lapidot,M. and Pilpel,Y. (2003) Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription. *Nucleic Acids Res*., **31**, 3824–3828.
8. Sudarsanam,P., Pilpel,Y. and Church,G.M. (2002) Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res*., **12**, 1723–1731.
9. Aach,J., Rindone,W. and Church,G.M. (2000) Systematic management and analysis of yeast gene expression data. *Genome Res*., **10**, 431–445.
10. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
11. Roberts,C.J., Nelson,B., Marton,M.J., Stoughton,R., Meyer,M.R., Bennett,H.A., He,Y.D., Dai,H., Walker,W.L., Hughes,T.R. *et al*. (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, **287**, 873–880.
12. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
13. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
14. Jelinsky,S.A., Estep,P., Church,G.M. and Samson,L.D. (2000) Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. *Mol. Cell. Biol*., **20**, 8157–8167.
15. Causton,H.C., Ren,B., Koh,S.S., Harbison,C.T., Kanin,E., Jennings,E.G., Lee,T.I., True,H.L., Lander,E.S. and Young,R.A. (2001) Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell*, **12**, 323–337.
16. Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.
17. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
18. De Smet,F., Mathys,J., Marchal,K., Thijs,G., De Moor,B. and Moreau,Y. (2002) Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, **18**, 735–746.
19. Heyer,L.J., Kruglyak,S. and Yooseph,S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res*., **9**, 1106–1115.
20. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet*., **22**, 281–285.
21. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol*., **296**, 1205–1214.
22. Springer,C., Kunzler,M., Balmelli,T. and Braus,G.H. (1996) Amino acid and adenine cross-pathway regulation act through the same 5′-TGACTC-3′ motif in the yeast HIS7 promoter. *J. Biol. Chem*., **271**, 29637–29643.
23. Dequard-Chablat,M., Riva,M., Carles,C. and Sentenac,A. (1991) RPC19, the gene for a subunit common to yeast RNA polymerases A (I) and C (III). *J. Biol. Chem*., **266**, 15300–15307.
24. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nature Genet*., **27**, 167–171.
25. Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res*., **31**, 7024–7031.
26. Schwikowski,B., Uetz,P. and Fields,S. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol*., **18**, 1257–1261.
27. Bar-Joseph,Z., Gerber,G.K., Lee,T.I., Rinaldi,N.J., Yoo,J.Y., Robert,F., Gordon,D.B., Fraenkel,E., Jaakkola,T.S., Young,R.A. *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol*., **21**, 1337–1342.
28. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.