nature biotechnology

Functional annotation and network reconstruction through cross-platform integration of microarray data

Xianghong Jasmine Zhou[1,2], Ming-Chih J Kao[2,3], Haiyan Huang[2,4], Angela Wong[1,5], Juan Nunez-Iglesias[1], Michael Primig[6], Oscar M Aparicio[1], Caleb E Finch[1,5], Todd E Morgan[1,5] & Wing Hung Wong[2,7]

The rapid accumulation of microarray data translates into a need for methods to effectively integrate data generated with different platforms. Here we introduce an approach, 2nd-order expression analysis, that addresses this challenge by first extracting expression patterns as meta-information from each data set (1st-order expression analysis) and then analyzing them across multiple data sets. Using yeast as a model system, we demonstrate two distinct advantages of our approach: we can identify genes of the same function yet without coexpression patterns and we can elucidate the cooperativities between transcription factors for regulatory network reconstruction by overcoming a key obstacle, namely the quantification of activities of transcription factors. Experiments reported in the literature and performed in our lab support a significant number of our predictions.

Microarray gene expression profiling is now done in many laboratories, resulting in the rapid accumulation of data in public repositories[1,2]. Despite recent advances in analysis techniques, several important challenges remain. (i) There is an urgent need for methods to effectively integrate multiple microarray data sets. Gene expression values generated with different platforms (such as spotted cDNA or Affymetrix high-density oligonucleotide arrays) are not directly comparable. Even within the same technology, alternative experimental parameters result in systematic variations among data sets often beyond the capability of statistical normalization. (ii) There is a lack of algorithms that can identify functionally related genes which do not have similar expression patterns. Most methods for functional analysis of microarray data make the implicit assumption that genes with similar expression profiles have similar functions[3,4]. However, among genes involved in the same pathway, many gene pairs do not show similar expression profiles[5]. (iii) The reconstruction of transcriptional regulatory networks remains the key challenge for microarray analysis. A major issue is the measurement of transcription factor activities because changes in their expression are often subtle and their activities are often controlled at levels other than expression. This further leads to difficulties in the elucidation of cooperativity between transcription factors. Recently, several approaches have been proposed to address some of these individual problems[5–7], yet there remains a lack of unified frameworks that can simultaneously respond to these challenges.

Here we introduce an approach termed 2nd-order expression analysis, which we will show to be useful in overcoming the three aforementioned problems. We define 1st-order expression analysis as the extraction of expression patterns from one microarray data set, which contains a set of expression profiles measured under relevant conditions. We propose 2nd-order expression analysis as a study of the correlated occurrences of those expression patterns across multiple data sets measured under different types of conditions (e.g., starvation, heat shock). By first extracting expression patterns as meta-information from each data set and then analyzing them comparatively, the results are not affected by variations among data sets. This allows integration of multiple microarray data sets in a platform-independent manner. Here, we apply 2nd-order analysis to 618 yeast expression profiles comprising 39 cDNA or Affymetrix array data sets to group genes that have the same function but may not be coexpressed, to annotate their functions, to quantify the activity profiles of transcription factors and reconstruct regulatory networks.

We illustrate 2nd-order expression analysis with a simple case, the analysis of expression patterns of coexpressed gene pairs. If a pair of genes is tightly coexpressed in multiple data sets, the genes are likely to be functionally linked. We term such gene pairs doublets. Our first objective is to find pairs of such doublets that simultaneously exhibit either high or low expression correlations across multiple data sets, that is, simultaneously turn on or off their functional links over different types of conditions. Such a set of four genes, termed a quadruplet, is likely to be functionally related, even though the global expression profiles of those genes do not exhibit gross similarities (see an example in **Fig. 1**). We identify quadruplets using a two-step procedure: (i) calculate the expression correlations of the doublet in each of the data sets and store them in a vector, termed 1st-order expression correlation profile; (ii) calculate the correlation between two 1st-order profiles to generate the 2nd-order expression correlation, and define those pairs of doublets with high 2nd-order correlations as quadruplets. Throughout the paper, an expression correlation or a

[1]Program in Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089-1113, USA. [2]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [3]School of Medicine, University of Michigan, Ann Arbor, Michigan 48109, USA. [4]Department of Statistics, University of California, Berkeley, California 94720, USA. [5]Andrus Gerontology Center, University of Southern California, Los Angeles, California 90089-0191, USA. [6]Biozentrum & Swiss Institute of Bioinformatics, University of Basel, CH-4056 Basel, Switzerland. [7]Department of Statistics, Harvard University, Boston, Massachusetts 02138-2901, USA. Correspondence should be addressed to X.J.Z. (xjzhou@usc.edu) or W.H.W. (whwong@stanford.edu).

Published online 16 January 2005; doi:10.1038/nbt1058

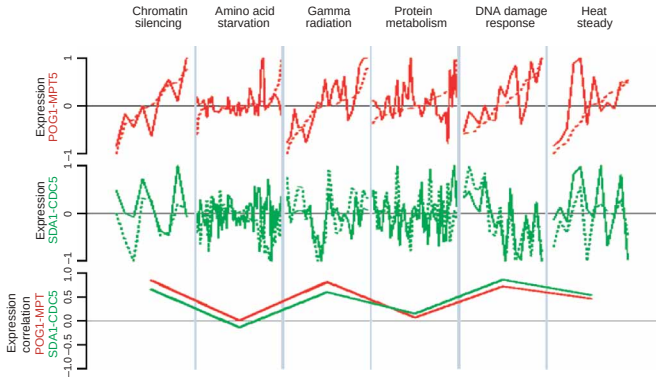

Figure 1 The expression profiles and 1st-order expression correlation profiles of gene pairs *POG1-MPT5* and *SDA1-CDC5* over six microarray data sets (data set details in **Supplementary Methods** online). All four genes are involved in the regulation of the cell cycle. The upper panel shows the normalized expression profiles of *POG1* (red solid) and *MPT5* (red dashed), and the middle panel shows those of *SDA1* (green solid) and *CDC5* (green dashed). The bottom panel shows expression correlation of the two gene pairs across six data sets. It is obvious that the overall expression similarity between the two gene pairs is not significantly high. However, as indicated in the bottom panel, their 1st-order expression correlation profiles exhibit high correlation, that is, the four genes have high 2nd-order expression correlation. To clearly demonstrate the pattern, the six data sets were sorted in decreasing order of the expression correlation difference between the two gene pairs. In each data set, expression profiles were sorted in increasing order of *POG1* expression value.

2nd-order correlation is considered high if it is greater than 0.6, a statistically conservative cutoff value (see **Supplementary Methods** online). Applying the approach to known yeast genes, we identified 5,142 doublets. Among the possible 13 million pairs of doublets, we further identified 278,799 quadruplets, 84% of which contain functionally homogeneous genes ($P < 10^{-5}$ by Monte Carlo simulation).

Furthermore, we compared our method to standard microarray analysis methods[8] by focusing only on cDNA array data sets. In contrast to our two-step procedure, the standard method merges multiple cDNA array data sets and calculates gene expression correlations across all arrays. We found that 83% of the 268,828 quadruplets derived by our method are functionally homogenous, and only 54% of the 4,186 co-expressed gene pairs determined using the standard method are functionally homogenous. To compare these results, we counted gene pairs contained in the functionally homogenous quadruplets. Since each quadruplet {a-b, c-d} yields 4 cross-doublet gene pairs: a-c, a-d, b-c and b-d, altogether the quadruplets give rise to a set of 2,597 distinct and novel gene pairs that are not contained in the set of doublets derived using first-order analysis. Strikingly, 97% of the 2,597 pairs are missed by the standard method. The 2nd-order expression analysis is complementary because (i) for a quadruplet, the cross-doublet gene pairs may be functionally related but may not show high expression correlation; (ii) our method is sensitive to gene pairs which are only coexpressed in a subset of the data sets, while the standard method is limited to gene pairs which are globally coexpressed; and (iii) the standard method is susceptible to variations between data sets, which can bias the estimation of expression correlations in the merged data.

Having validated that the 2nd-order expression analysis can effectively group functionally related genes, we generalize the method from grouping two gene pairs to grouping k gene pairs for functional annotation. This can be achieved by clustering doublets based on their 1st-order expression correlation profiles. Each cluster represents a module of functional links following the same patterns of being turned on or off across multiple data sets. Similar to the case of quadruplets, functionally related genes with low 1st-order but high 2nd-order correlations can be clustered together. As an indication of the power of 2nd-order correlation for functional clustering, using all known doublets, we observed that 72 of the top 100 tightest clusters are functionally homogeneous. We applied 2nd-order clustering to functional annotation, and made a prediction for a doublet only if it is in a tight cluster that includes at least three doublets and in which all remaining doublets shared the same function. Among the 100 tightest clusters of known doublets, 179 doublets satisfy this condition. Of these, 91% has the same function that is shared by the remaining doublets in the cluster. Expanding this approach to unknown genes, we assigned 79 functions to 67 unknown genes (**Supplementary Table 1** online). Many of those predictions are supported by experimental studies in the literature. For example, we predicted that YLR183C participates in mitosis, and a recent study revealed that it is involved in the regulation of G1/S transition[9]. We assigned the function 'cation transport' to YLL051C, which is known to have iron-regulated expression[10]. To validate our prediction of YOR309C as a gene involved in 'rRNA processing' we used northern blot analysis to examine the abundance and the processing of cellular rRNAs in the YOR309C knockout strain (**Fig. 2**). The presence of a strong 35S band and the reduced abundance of other rRNA cleavage products suggest a defect early in the 35S pre-rRNA processing pathway in the knockout strain, which supports our prediction for YOR309C.

Finally, we generalized the 2nd-order method to reconstruct regulatory networks. In fact, the biological basis for 2nd-order correlation is contained in the structure of regulatory networks. Particularly, it is the correlated activities of transcription factors that give rise to the 2nd-order expression correlation between their target gene pairs, or more generally, target gene sets. In the following, we illustrate this mechanism by analyzing the 2nd-order relationship of two

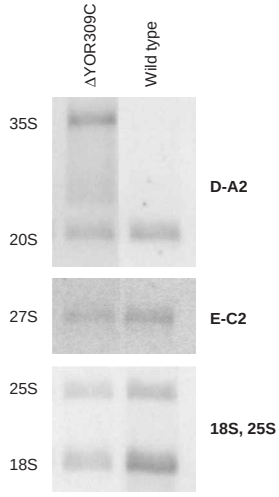

Figure 2 Northern blot analysis showing the abundance of different cellular rRNAs in wild-type and ΔYOR309C cells. The 35S pre-rRNA band is prominent in ΔYOR309C cells. In addition, compared to wild-type cells, ΔYOR309C cells have lower levels of 18S and 25S rRNA, and of their respective precursors 20S and 27S. This suggests a defect early in the 35S pre-rRNA processing pathway in the knockout strain. Oligonucleotide probes specific for different cleavage processing sites (D-A2, and E-C2) and the products 18S and 25S were used[30].

transcription modules. A transcription module is defined to be a set of genes that are regulated by the same transcription factor(s) based on genome-wide location data[11], and are coexpressed in at least m out of a total of n data sets ($m > n/4$ and $m \geq 8$ in this study). If two transcription modules form or do not form two coexpression clusters mostly under the same set of conditions (that is, in the same data sets), it in fact suggests that the two (sets of) transcription factors regulating the two modules are mostly active or inactive simultaneously. Such cooperativities between two sets of transcription factors can be quantified using 2nd-order expression correlation: (i) We assessed the activities of a transcription factor by the tightness of coexpression among the genes it regulates. Specifically, for a transcription module, we constructed a vector of length n storing the average pairwise expression correlations among its member genes for each data set. This 1st-order average expression correlation profile can be interpreted as the activity profile of the transcription factor(s) that regulate the module. (ii) We calculated the correlation between two activity profiles of transcription factors, that is, 2nd-order expression correlation, to measure the cooperativity between the two transcription factors.

Based on the genome-wide location data and coexpression clusters recurrent in multiple data sets, we identified 60 transcription modules, all of which demonstrated a high degree of consistency in terms of their known functions and regulations (**Supplementary Table 2** online). Among module pairs controlled by distinct transcription factors, 34 pairs showed high 2nd-order correlation. For these module pairs, we further traced the potential source of cooperativity of their regulators using DNA binding data[11], protein interaction[12,13] and protein complex[14,15] data. Given two modules controlled by respective transcription factor(s) TF_1 and TF_2, which for simplicity are assumed to be individual instead of sets of transcription factors, there are at least three types of direct causes of the cooperativity between TF_1 and TF_2 (**Fig. 3a**): the expressions of TF_1 and TF_2 are activated by a

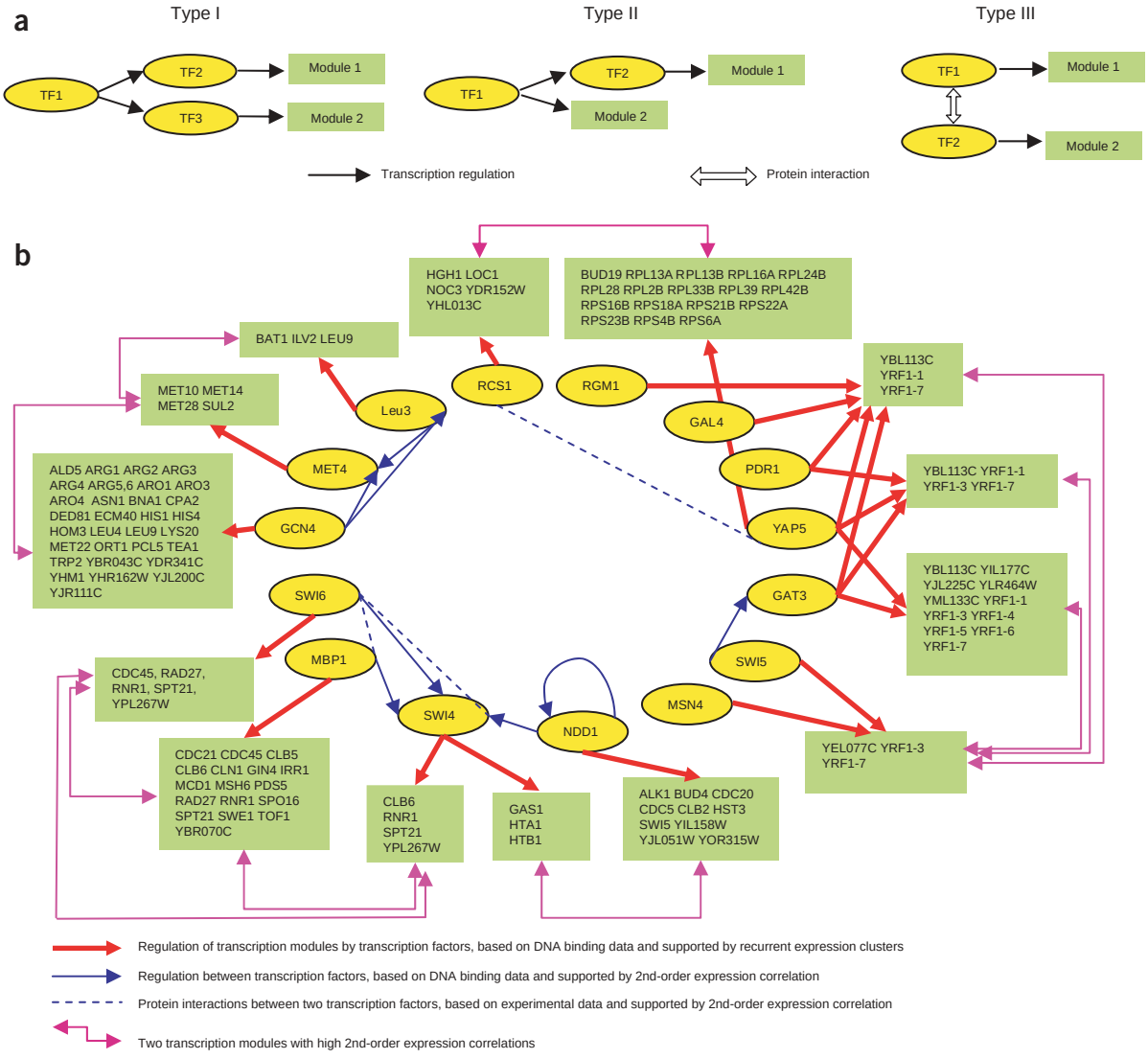

Figure 3 Reconstruction of regulatory networks by 2nd-order expression analysis. (**a**) Three types of possible transcription cascades that could explain 2nd-order correlation of two transcriptional modules. Given two modules controlled by two transcription factors, TF_1 and TF_2, respectively, the coactivation of the two modules implies cooperativity between TF_1 and TF_2, which may be caused by a type I cascade in which the activities of TF_1 and TF_2 are controlled by common transcription factor(s), TF_3; or a type II cascade in which the activity of TF_2 is controlled by TF_1, or vice versa; or a type III cascade in which TF_1 and TF_2 interact at the protein level. (**b**) Regulatory network reconstructed on the basis of the derived transcription cascades. Yellow ovals denote transcription factors; green boxes are transcription modules defined by recurrent expression clusters.

common transcription factor TF_3 (type I transcription cascade), TF_1 activates the expression of TF_2 (type II transcription cascade), or TF_1 and TF_2 interact at the protein level (type III transcription cascade). In the special case where a module pair shares a majority of common genes, the cooperativity between TF_1 and TF_2 is known to be combinatorial control. Note that the three types of transcription cascades are certainly only a few of the many possibilities.

We identified a significant portion (29%, $P < 10^{-5}$ by Monte Carlo simulation) of cooperative module pairs as participants in transcription cascades: 2 pairs in type I, 8 pairs in type II, and 3 pairs in type III cascades. In fact, these transcription cascades interconnect into a partial cellular regulatory network (**Fig. 3b**). A large proportion of identified transcription cascades are involved in cell cycle control. Examples are the type I/II cascade involving SWI4 and NDD1 (the autoregulation of NDD1 leads to both types I and II construction), the type II/III cascade involving SWI4 and SWI6, the type II cascade involving SWI4 and MBP1 and the type III cascade involving SWI6 and MBP1. Many identified cascades can be validated by experiments in the literature[9,16]. In **Supplementary Notes** online, we show a detailed example of building the transcription cascade among the regulator GCN4, the LEU3 module and the MET4 module, the interrelationships of which are not obvious in the 1st-order expression level but can be revealed with the aid of 2nd-order analysis, all with further support in the literature[11,17,18].

We further extend the study from analyzing two transcription modules to *k* modules to capture the relationships among multiple sets of transcription factor(s). Clustering the 1st-order profiles of the 60 modules, we obtain sets of modules the regulators of which are likely to be concurrently active across different conditions. In the majority of the clusters, the member modules participate in the same biological processes (**Fig. 4**), confirming again the utility of the 2nd-order approach in grouping functionally related genes. An additional application of the 2nd-order clustering is to assign transcription factors to biological processes, a difficult task with common clustering methods due to the subtle expression patterns of transcription factors. For an unknown transcription factor in a module cluster, we can annotate its function by integrating the evidence of two dimensions: (i) the functions of known genes in its target module, and (ii) the functions of known genes in other modules in the same module cluster, including both the transcription factors

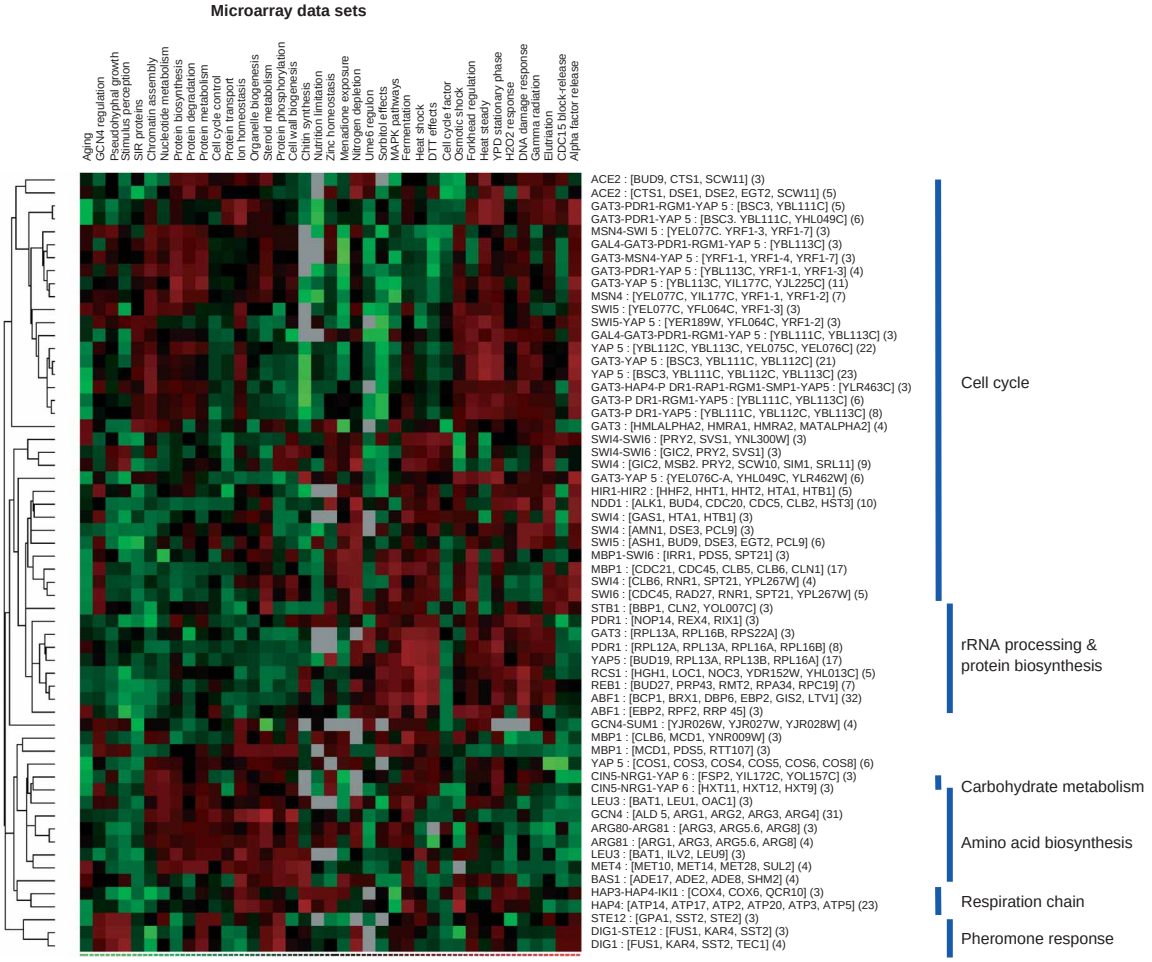

Figure 4 Hierarchical clustering of transcription modules based on their average 1st-order expression correlation profiles. The color of each cell represents the standardized average 1st-order correlation of a transcription module (row) within a data set (column), where red/green indicates high/low correlation, and gray indicates insufficient data. The name of the data set is at the top. In each row to the right, the name of the transcription factor(s) is listed before the colon, multiple combinatorial transcription factors are concatenated with a hyphen, genes in brackets, ordered alphabetically, are in the same transcription module. Because of space limitations, gene lists may not be complete for some transcription modules (for complete gene lists refer to **Supplementary Table 2** online); the total number of genes in each transcription module is indicated in parentheses. For example: "TF_1-TF_2: [$Gene_1$, $Gene_2$, $Gene_3$] (3)".

and their target genes. In **Supplementary Notes** online, we provide a detailed example of combining these two types of evidences to predict that the transcription factor GAT3 plays a role in mitotic and meiotic cell cycles, a prediction that is consistent with the growth pattern of its deletion strain[19] and its cell cycle phase–specific expression in two independent studies[20,21].

We have used 2nd-order expression analysis to heterogeneous data sets, to annotate functions of genes beyond coexpression and to identify cooperativity among transcription factors to reconstruct regulatory networks. Integrating heterogeneous data sets is one of the current challenges in bioinformatics. Previous studies[6,22] have emphasized extracting modules in which genes manifest a common pattern across a significant subset of data. In our approach, we go one step further to consider the high-order relationship between those modules, that is, clustering modules exhibiting similar activity patterns across different types of conditions. Such 2nd-order similarity is not visible at the first-order level examined by current methods, and it allows us to cluster genes with the same function yet not obviously coexpressed. In terms of regulatory network reconstruction, the major difference between our method and other methods[23,24] is that 2nd-order analysis by design not only derives the transcription modules as do 1st-order methods, but also reveals network relationships among different transcription modules, and in turn sheds light on the cooperativity among their regulators. Specifically, we overcome a major obstacle in regulatory analysis—the quantification of the activities of transcription factors—by measuring the coexpression tightness of the transcription module it regulates. We note that a recently published study also addressed the same question, using a different approach[7]. Based on the activity profile of individual transcription factors, the 2nd-order correlation measures the cooperativity between them, which together with DNA binding and protein interaction data, provides inference on transcription cascades. In doing so, it becomes possible to trace the signals one-step upstream along transcription cascades, where signals are influential yet generally subtle to measure, and where relationships are normally difficult to infer. Furthermore, when larger collections of data sets become available in the near future, 2nd-order analysis can be extended to study patterns among transcription factors or transcriptional modules that are more complex than correlation. For example, we can adapt a variety of approaches currently applicable to gene expression profiles to the activity profiles of transcription factors to elucidate their complex interrelationships. In this way, 2nd-order analysis can reveal the structure of hierarchical regulatory networks one level higher than current methods.

METHODS

Microarray data. We integrated all yeast microarray data sets (up to January 2003), each containing at least eight experiments, from the Stanford Microarray Database[2] and from the NCBI Gene Expression Omnibus[1]. We included two additional cDNA array data sets[25,26] and also the Rosetta Compendium data[27] to gain broader coverage of experimental conditions. Note that the power of the 2nd-order method can be maximized if each data set contains a set of coherent biological conditions (some data sets need additional processing to meet this requirement, e.g. the Rosetta compendium data), and if the collected data sets together cover a greater range of perturbations (details in **Supplementary Methods** online).

Construction of functional categories. We constructed 43 functional categories covering 2,429 known yeast genes based on the biological process ontology of gene ontology[28] as previously described[5]. Each functional category contains more than 60 genes and none of its subcategories contains more than 60 genes (details in **Supplementary Methods** online).

Computing the 1st-order expression and 2nd-order expression correlation. For any two genes a and b from the same functional category, we define their 1st-order expression correlation $J_{a,b,k,}$ in the data set k as the leave-one-out Pearson's correlation coefficient with the minimum absolute value, that is, the jackknife correlation. This estimate is a measurement robust against single experimental outliers and sensitive to overall similarities in expression patterns. We thus obtain a 1st-order expression correlation profile ($J_{a,b,k}$), $k = 1, \ldots, n$, for genes a and b. Provided with n data sets subjected to different types of perturbations, a gene pair is defined as a doublet if it demonstrates coexpression ($J > \tau$) in at least m out of n experimental groups ($m > n/4$ and $m \geq 8$). Note that n may vary for different gene pairs because of missing data, and a data set is included only if it contains at least eight experiments that simultaneously measured the expressions of genes a and b. We choose $\tau = 0.6$ because it is a statistically conservative cutoff value, which nonetheless retains a sufficient number of functional links (details in **Supplementary Methods** online). Between the 1st-order expression correlation vectors of any two nonoverlapping doublets *a-b* and *c-d*, we compute jackknife correlation to identify quadruplets with high (>0.6) 2nd-order expression correlation. A quadruplet is defined to be functionally homogeneous if all four genes come from the same gene ontology functional category.

In this study, we treat each data set equally in terms of its contribution to 2nd-order expression correlation. However, as our method will be applied more generally to larger collections of microarray data sets, there may be redundancies in some of the experimental conditions. Additional weighting schemes may be applied to offset such redundancies.

Functional annotation. For a functional category i, our functional annotation scheme consists of three steps. (i) For a unknown gene a, we count the number λ_i of the doublets in which a is paired with genes from the functional category i. Modeling λ_i as a hypergeometric random variable, we compute the statistical significance $P(a, i)$ of associating gene a with the functional category i. If $P(a, i) < 0.05$, we include the λ_i doublets into the collection U_i. We repeat the step for all unknown genes to construct the collection U_i. This step serves as a prefiltering of doublets that are highly likely to represent true functional links. (ii) For the functional category i, we cluster all doublets in U_i and all doublets of known functions using the TightCluster algorithm[29] (a resampling-based approach that produces stable and tight clusters without forcing all points into clusters), and select the top 100 tightest clusters for further analysis. We repeat the step for all functional categories. Note that the 2nd-order correlation was used as the similarity measure during the clustering. (iii) If a tight cluster contains at least three doublets and all of its known genes fall into the same functional category, we assign this function to those unknown genes. Because the information used in step i (1st-order expression information) and that in steps ii and iii (2nd-order expression information) are complementary, satisfying both criteria imposes very strong constraints on our predictions.

Assess the statistical significance of the number of identified quadruplets. To demonstrate the power of 2nd-order expression analysis in grouping functionally related doublets into quadruplets, we evaluate the number of functionally homogenous doublet pairs expected under the null hypothesis. We construct 1,000 random pairs of doublets from the 5,142 doublets derived from known genes, and calculate the ratio γ of functionally homogeneous pairs. This is done for 10^5 iterations. The observed quantity is compared to the distribution of γ generated under the null to derive the P value.

Identification of transcription modules. Given n expression data sets and all genes known to be under the regulation of a transcription factor based on genome-wide location data[11] (we used a P value cutoff of 10^{-3}), we searched for a subset of genes $G = \{g_1, g_2, \ldots, g_l\}$, where $l \geq 3$ and the jackknife expression correlation of all gene pairs in G were greater than 0.6 in the same m out of n data sets, where $m > n/4$ and $m \geq 8$. For each gene cluster, we constructed the 1st-order average expression correlation profile by computing the average pairwise expression correlations of its member genes in each of the n data sets. We merged two gene clusters if (i) the two gene clusters were under the regulation of the same transcription factor based on DNA binding data, (ii) they differed only by one gene, and (iii) they showed high correlation of their

1st-order average expression correlation profile. In this way, we determined 60 nonredundant gene clusters, defined as transcription modules.

Assessing the statistical significance of the number of identified transcription cascades. To statistically evaluate the power of our method in revealing transcription cascades, we investigated the expected numbers of such cascades discovered under the null hypothesis. We randomly picked 1,000 pairs of transcription factors from the 106 transcription factors for which genome-wide location data is available. We then counted the percentage of transcription factor pairs, denoted as ρ, which participated in transcription cascades types I, II or III, based on DNA binding and protein interaction/complex data. We repeated the procedure 10^5 times, and generated a distribution for ρ. We compared the observed 29% to the distribution of ρ generated under the null hypothesis to derive the *P* value.

Yeast cell culture and northern blots. *S. cerevisiae* strains ΔYOR309C and BY4741 (wild-type) were purchased from Open Biosystems. Cells were cultured in YPD medium (1% yeast extract, 2% peptone, 2% glucose) at 30 °C. RNA was isolated from cells at mid-log phase using glass beads and phenol/choloroform and precipitated with ethanol. We loaded 5 μg RNA per lane on a 1% agarose formaldehyde gel and transferred to Biodyne B nylon membranes (Millipore). DNA oligonucleotides specific for D-A2, E-C2, 18S and 25S[30] were labeled with ^{32}P and hybridized overnight to membranes with ULTRAhyb Ultrasensitive Hybridization Buffer (Ambion) at 37 °C. After washing with 2× SSC, 0.1% SDS, membranes were exposed and scanned using a Phosphoimager (Molecular Probes).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank Robert Gentleman for making his computer resources available for part of this project, Timothy Hughes for technical advice and Michelle Arbeitman for sharing her lab space. We also thank two anonymous reviewers for their helpful comments. The work of X.J.Z. was supported by the National Science Foundation grant DMS0090166 to W.H.W., the Faculty Setup Grant from USC and the National Institutes of Health (NIH) grant R01GM067243 to Simon Tavaré. The work of M.-C.J.K was supported by a Howard Hughes Pre-doctoral Fellowship. The work of H.H. was supported by the NIH grant P20CA96470 to W.H.W. and the Faculty Setup Grant from UC Berkeley. The work of W.H.W. was supported by the NIH grant R01HG02341.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 21 July; accepted 1 November 2004
Published online at http://www.nature.com/naturebiotechnology/

1. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
2. Gollub, J. *et al.* The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* **31**, 94–96 (2003).
3. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
4. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
5. Zhou, X., Kao, M.C. & Wong, W.H. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. USA* **99**, 12783–12788 (2002).
6. Rhodes, D.R. *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. USA* **101**, 9309–9314 (2004).
7. Gao, F., Foat, B.C. & Bussemaker, H.J. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* **5**, 31 (2004).
8. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
9. Horak, C.E. *et al.* Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.* **16**, 3017–3033 (2002).
10. Martins, L.J. *et al.* Metalloregulation of FRE1 and FRE2 homologs in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **273**, 23716–23721 (1998).
11. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
12. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
13. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
14. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
15. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
16. Futcher, B. Transcriptional regulatory networks and the yeast cell cycle. *Curr. Opin. Cell Biol.* **14**, 676–683 (2002).
17. Mountain, H.A., Bystrom, A.S. & Korch, C. The general amino acid control regulates MET4, which encodes a methionine-pathway-specific transcriptional activator of Saccharomyces cerevisiae. *Mol. Microbiol.* **7**, 215–228 (1993).
18. Zhou, K., Brisco, P.R., Hinkkanen, A.E. & Kohlhaw, G.B. Structure of yeast regulatory gene LEU3 and evidence that LEU3 itself is under general amino acid control. *Nucleic Acids Res.* **15**, 5261–5273 (1987).
19. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
20. Primig, M. *et al.* The core meiotic transcriptome in budding yeasts. *Nat. Genet.* **26**, 415–423 (2000).
21. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
22. Tanay, A., Sharan, R., Kupiec, M. & Shamir, R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genome-wide data. *Proc. Natl. Acad. Sci. USA* **101**, 2981–2986 (2004).
23. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
24. Bar-Joseph, Z. *et al.* Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* **21**, 1337–1342 (2003).
25. Natarajan, K. *et al.* Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Mol. Cell. Biol.* **21**, 4347–4368 (2001).
26. Roberts, C.J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
27. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
29. Tseng, G. & Wong, W. A Method for Tight Clustering: with Application to Microarray. *Proc. 2nd IEEE Computer Society Bioinformatics Conference*, 396–397 (2003).
30. Peng, W.T., Krogan, N.J., Richards, D.P., Greenblatt, J.F. & Hughes, T.R. ESF1 is required for 18S rRNA synthesis in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **32**, 1993–1999 (2004).