Molecular Systems Biology 3; Article number 96; doi:10.1038/msb4100137
Citation: ***Molecular Systems Biology*** **3:96**

www.molecularsystemsbiology.com

molecular systems biology

Prediction of phenotype and gene expression for combinations of mutations

Gregory W Carter*, Susanne Prinz, Christine Neou, J Patrick Shelby, Bruz Marzolf, Vesteinn Thorsson and Timothy Galitski

Institute for Systems Biology, Seattle, WA, USA
* Corresponding author. Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103, USA. Tel.: +1 206 732 1396; Fax: +1 206 732 1299;
E-mail: gcarter@systemsbiology.org

Received 21.11.06; accepted 9.2.07

Molecular interactions provide paths for information flows. Genetic interactions reveal active information flows and reflect their functional consequences. We integrated these complementary data types to model the transcription network controlling cell differentiation in yeast. Genetic interactions were inferred from linear decomposition of gene expression data and were used to direct the construction of a molecular interaction network mediating these genetic effects. This network included both known and novel regulatory influences, and predicted genetic interactions. For corresponding combinations of mutations, the network model predicted quantitative gene expression profiles and precise phenotypic effects. Multiple predictions were tested and verified.
Molecular Systems Biology 27 March 2007; doi:10.1038/msb4100137
Subject Categories: functional genomics; differentiation and death
Keywords: computational biology; data integration; gene expression; genetic interaction; network model

Introduction

Identifying causal links between genetic variation and phenotype is a central challenge of modern genetics. The combination of gene perturbation technology (Fire *et al*, 1998; Winzeler *et al*, 2000) and high-throughput phenotype assays (Drees *et al*, 2005; Dudley *et al*, 2005) enables rapid identification of genes active in a biological response. Linkages between individual genes and specific phenotypes can also be established statistically by detecting quantitative trait loci (QTLs) (Barton and Keightley, 2002). However, formulating biomolecular models based on these techniques is difficult, because most phenotypes are controlled by multiple genes with multiple allelic variants. Moreover, the alleles of these genes often interact in complex ways to affect phenotype (Shook and Johnson, 1999; Steinmetz *et al*, 2002; Carlborg and Haley, 2004; Sinha *et al*, 2006). Thus it is necessary to model functional relationships between relevant (i.e. trait) genes instead of viewing each gene as an independent factor. The resulting network models will have the capacity to predict, systematically and explicitly, the effects of multiple interacting genetic perturbations. This capacity will enable testing of genetically complex hypotheses, prioritization of candidate genes for targeted intervention, and the personalization of prognoses and therapies (Ideker *et al*, 2001; Galitski, 2004).

The identification of functionally relevant interactions in databases of diverse high-throughput data types is a substantial challenge for the construction of predictive network models. Many recent efforts have sought to distill functionally important information by detecting systematic congruences in multiple large data sets (Wong *et al*, 2004; Sachs *et al*, 2005; Segre *et al*, 2005; Workman *et al*, 2006; Zhong and Sternberg, 2006). These approaches have had success in functionally classifying genes and identifying probable candidate gene pairs for the simple presence or absence of genetic interaction, often defined either very broadly as any genetic nonindependence (Zhong and Sternberg, 2006) or very narrowly as one particular interaction mode such as negative synthesis (Tong *et al*, 2004). In contrast, our goals are to infer specific functional relationships to drive network modeling, and to make precise testable predictions for novel combinatorial perturbations of genes.

Accordingly, we developed an analysis of genetic interaction as a quantitative influence and used the results to direct the integration of molecular (physical) interaction data. We define these influences as positive or negative numbers of varying magnitude that account for the fraction of a measurable phenotype (e.g. the expression of a gene) inferred to be caused by a system element (e.g. a regulatory protein). The measured phenotype is modeled by multiple influences acting throughout the inferred network. Our mathematical modeling is based on the classical genetic-interaction approach of observing how genetic perturbations interact to affect phenotypes, thereby revealing functional relationships such as activation, repression, and pathway order (Avery and Wasserman, 1992). However, because mutant phenotypes result from the activities of complex molecular pathways, the biochemical interpretation of a genetic interaction is often ambiguous and frequently involves multiple alternative molecular models and both direct and indirect mechanisms (Kelley and Ideker, 2005; Zhang *et al*, 2005). Conversely, molecular interactions, plentifully generated through high-throughput methods, often lack in functional interpretation or are of uncertain relevance to specific genetic observations (Galitski, 2004). Therefore,

our approach is to exploit this complementarity of genetic and molecular interactions. Our approach is to (i) decompose genetic-interaction data into influences encoding genetically direct and indirect effects and (ii) use the molecular wiring to constrain the molecular interpretation of genetic interactions, and thereby assign function to specific molecular interaction paths.

Results

Our approach required genetic manipulations, genome-scale molecular-interaction data sets, and efficient phenotype assays. Thus we used the filamentous growth response of budding yeast as a model system (Gimeno *et al*, 1992; Lengeler *et al*, 2000). In response to environmental cues, yeast cells switch from their round single-cell growth form to a pathogen-like, adhesive, invasive, filamentous form (also known as pseudohyphal growth). Both the filamentous-growth phenotype (Drees *et al*, 2005) and microarray data (Van Driessche *et al*, 2005) have been shown to be suitable measurements for the study of genetic interactions. We therefore constructed a set of single gene and double gene deletion strains and assayed each for filamentation phenotype and gene expression (thousands of measurements per strain). We inferred specific genetic influences from these data and used the results to guide the integration of molecular-interaction data in a network controlling filamentous growth.

Combinatorial genetic perturbation

Transcription factor genes were chosen for perturbation in our study because they play a direct role in the regulation of gene expression. Thus, they offer a good prospect of modeling their effects on gene expression and phenotype. The genes of five specific transcription factors known to regulate the filamentous growth response were chosen for deletion: *TEC1* (Gavrias *et al*, 1996), *SOK2* (Ward *et al*, 1995), *SKN7* (Lorenz and Heitman, 1998), *SFL1* (Robertson and Fink, 1998), and *CUP9* (Prinz *et al*, 2004). We refer to these starting-point genes as *seed genes*. They were chosen because: (i) they show a full range of single-mutant phenotypes (from strongly hypo-filamentous to strongly hyper-filamentous), creating interesting double-mutant combinations; and (ii) they are downstream of a representative group of major signaling pathways involved in filamentous growth. The inference of a genetic interaction requires comparing the phenotypes of four genotypes: a 'wild type', two single mutants, and a double mutant that carries both mutant genes. Thus, we studied 16 strains: the wild type, five single-gene deletion strains (*tec1*Δ, *cup9*Δ, *sfl1*Δ, *sok2*Δ, *skn7*Δ), and all 10 combinatorial double deletions (*tec1*Δ*cup9*Δ, *tec1*Δ*sfl1*Δ, etc.). Gene expression data were collected for each strain under filamentous growth conditions (Supplementary information). All subsequent analyses were restricted to 1863 genes showing differential expression (Supplementary information). Also, each strain was phenotyped for filamentation (Supplementary information). The results revealed a rich pattern of genetic interactions with frequent occurrences of classical epistasis, in which a double-mutant phenotype is the same as one of the mutations (the epistatic mutation) and the other (hypostatic) mutation is masked (Supplementary Table S1). For example, we found *TEC1* deletion to be epistatic to all other seed gene deletions (i.e. the phenotypes of all *tec1*Δ-containing double-mutant genotypes were like the *tec1*Δ phenotype), in agreement with its known role as a major direct regulator of filamentation genes (Chou *et al*, 2006).

Model of interacting genetic influences on gene expression

We used data-driven linear decomposition to model genomic expression and quantify genetic interactions. Matrix decomposition methods, including singular value decomposition (SVD) (Alter *et al*, 2000; Carter *et al*, 2006) and generalized Network Component Analysis (gNCA) (Yang *et al*, 2005), have proven successful in disentangling multiple overlapping quantitative signals in microarray data. Our decomposition method was designed to dissect the complexities of genetic interactions (Materials and methods). The solution can be represented as a network of influences, as illustrated in Figure 1. This procedure results in the decomposition of an expression data matrix **D** into two matrices: (i) an *influence matrix*, **X**, of coefficients for the genotype-independent influences of the seed genes on target genes; and (ii) a *genotype matrix*, **G**, of inferred activity levels for the seed genes in each genotype. This is concisely written as

$$\mathbf{D} = \mathbf{X}.\mathbf{G} \qquad (1)$$

Thus, the genetically 'direct' (not necessarily molecularly direct) influences from the seed genes to target genes are separated quantitatively from the genetically 'indirect' effects that involve a second seed gene and a genetic interaction. In the genotype matrix, **G**, we define the wild-type activities to be equal to one ($g_A^{wt}=g_b^{wt}=\ldots=1$); activity levels of null alleles are fixed at zero (Materials and methods). Note that other allele types can be accommodated readily with a measured level of activity relative to the wild type. Other genotype matrix elements (capturing genetic interactions) are unknown *a priori*, but they can be calculated as activity changes relative to wild type under perturbations of other seed genes ($g_A^{B\Delta}$, $g_B^{A\Delta}$, $g_C^{A\Delta B\Delta}$, etc.).

We performed a least-squares best-fit solution for the decomposition defined by Equation (1) (Supplementary information). For our data set, the resulting model showed high correlations with the observed expression profiles of all genes across all experimental conditions (Materials and methods). The inclusion of genetic interactions in the model accounted for much of this correlation (Supplementary Figure S1).

We next integrated molecular interaction data with our decomposition results to construct regulatory network models (Materials and methods). Figure 2 illustrates the strategy with a small network for the transcriptional regulation of the gene *DDR48*, encoding an ATPase involved in stress response, cell wall organization, and flocculation (Tonouchi *et al*, 1994). This strategy was applied genome wide. As an example, Figure 3 shows the inferred molecular network transmitting positive influences from *SKN7* to 54 genes. Networks were constructed from molecular interactions that were specifically selected based on the presence of quantitative influences (with

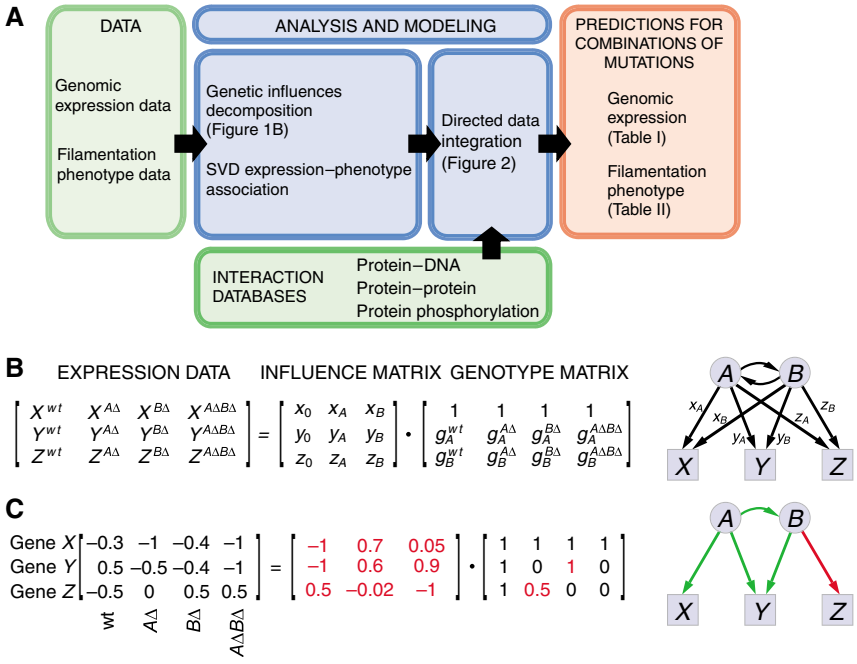

Figure 1 Outline of modeling strategy. (**A**) Overall modeling strategy. (**B**) Genetic influences decomposition and corresponding networks for a simplified system of two seed genes, *A* and *B* (circles), influencing the expression of three target genes, *X*, *Y*, and *Z* (boxes). The data matrix is decomposed in terms of influence and activity variables, corresponding to all possible influences in the network. (**C**) Illustration with synthetic expression data of three genes in four strain backgrounds. Positive and negative numbers in the influence matrix with magnitude greater than the significance cutoff (0.1) map to green and red edges in the network, respectively. Black numbers correspond to expression data and activities fixed by genetic backgrounds (all wild-type activities are 1, the activity of the deleted gene *A* is g_A^A=0, etc), while the red numbers are the best-fit solution of the system. In this example, the genotype matrix element of gene *B* in the *A*-deletion strain is reduced (g_B^A=0.5), from which a positive influence from *A* to *B* is deduced and shown in the network.

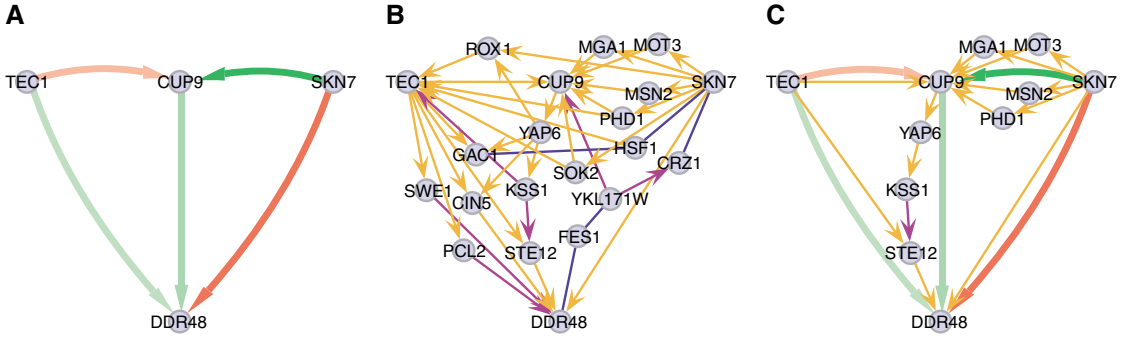

Figure 2 Directed data integration: modeling how *TEC1*, *CUP9*, and *SKN7* control the expression of gene *DDR48*. (**A**) Network of inferred influences that cause genetic interactions. Edges indicate the direction of positive (green) and negative (red) influences, with intensity indicating magnitude of influence. (**B**) Network of physical interactions connecting the four genes from high-throughput data sets. This network is too dense and disorganized to identify functional pathways. Interactions are protein–protein (blue), protein–DNA (orange), and protein phosphorylation (violet). (**C**) Integrated network constructed from the subset of molecular paths in (B) that are specific candidates for transmission of influences in (A). Influences from the remaining seed genes (*SOK2* and *SFL1*) have been omitted for clarity.

sign, magnitude, and direction) inferred in the model. Pathways involving five or greater links were discarded because they are longer than the average shortest connection between any two elements in the global network, and thus are less likely to be biologically relevant (Steffen *et al*, 2002). These networks represent specific, testable hypotheses of influence from the causal perturbation to the expression of affected genes.

From the genotype matrix, **G**, in Equation (1), we inferred quantitative cross-influences between seed genes that genetically interact (Figure 4A; Supplementary information). These correspond to influences on inferred *regulatory activity*, rather than seed gene expression. For instance, we inferred from *SKN7–CUP9* genetic interactions present throughout the expression data that *SKN7* has a positive influence on the regulatory activity of *CUP9* during the filamentous growth response (Figures 2A and 4A). We found candidate molecular paths to transmit these influences following the method described above (Figure 2C and Supplementary Figure S2). This allowed us to make predictions for perturbations of these path genes, in which the modified influences alter genetic interactions.

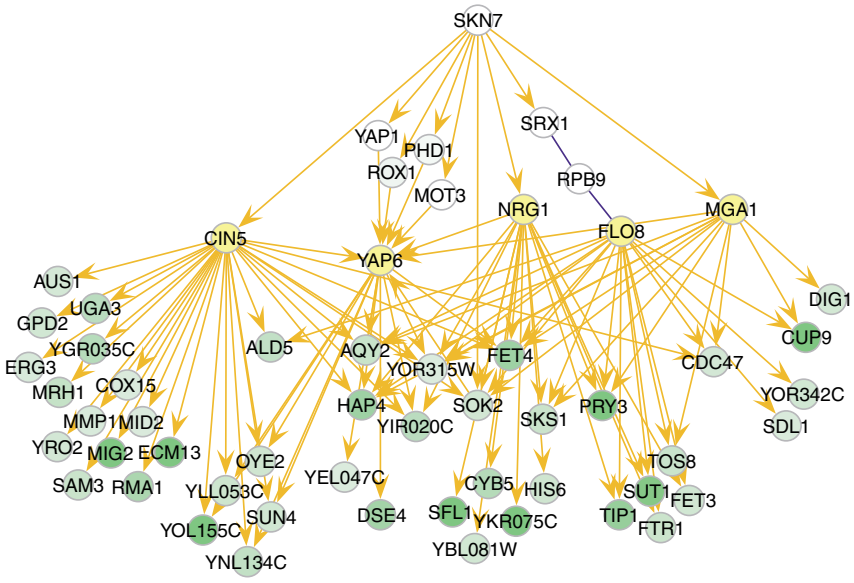

Figure 3 *SKN7*-positive influence network. An influence network is shown that maps putative molecular paths of influence from *SKN7* to target genes, which are shaded green in proportion to their influence coefficient. White nodes are proteins that fall on the shortest directionally consistent putative paths of influence from *SKN7* to the enriched transcription factors shown in yellow. Interactions are colored as: protein–protein in blue and protein–DNA in orange.

Predictions of gene expression

We made quantitative predictions of gene expression for additional perturbations. Because we could not associate the background expression influences (Equation 2, x_0, y_0, etc.) with specific molecular pathways, it was not clear what the total effects of a novel single-mutant deletion would be. However, we could predict precise expression levels for double perturbations once the single perturbations were known. Thus for each new double mutant, we predicted the quantity $X^{A\Delta B\Delta}$–$X^{A\Delta}$–$X^{B\Delta}$ for each gene. To make predictions for each new double mutant, we removed a genetic influence (quantitative contribution to gene expression) whenever the molecular path from influencer to influenced was broken by deletion of a gene on the path. If an alternative path of the same length exists, the influence was not removed. This gave more accurate results than removing the influence when an alternative path exists (data not shown). This is consistent with studies that show regulatory information often flows via parallel pathways (Kelley and Ideker, 2005). For broken paths, both expression influences in matrix $\boldsymbol{X}$ and activity influences in matrix $\boldsymbol{G}$ were removed. The matrix $\boldsymbol{X.G}$ was recomputed (Equation (1)) to obtain the prediction for mutant gene expression.

We compared our predictions to observed quantities. We collected microarray gene expression data for four additional single- and double-deletion strains: *yap6*Δ, *cup9*Δ *yap6*Δ, *sfl1*Δ *yap6*Δ, and *sok2*Δ *yap6*Δ. *YAP6*, which is not a seed gene, was chosen for its central role mediating influences from *SFL1*, *CUP9*, and *SOK2* to other genes (Supplementary Figure S2 and Supplementary Table S3). For these *YAP6* deletion strains, we identified genes that receive expression influences putatively transmitted by Yap6 (e.g. *DDR48* in Figure 2; Supplementary Table S2) and set those influence coefficients to zero. Yap6 is also a candidate for transmitting influences among the seed genes, specifically those from *SFL1* and *CUP9* (Figure 4A and Supplementary Figure S2). We initially removed all putative activity influences following the above procedure. These included activity influences from *SFL1* and *CUP9* to *TEC1*, *SOK2*, and each other, and from *CUP9* to *SKN7*. Removal of all of these resulted in poor predictions (data not shown). We refined the model with minimal alteration, and after testing the effects of the removal of each individual activity influence, we found that prediction inaccuracy was almost entirely due to the removal of the *CUP9* influence on *SKN7*. This suggests a parallel molecular path of influence from *CUP9* to *SKN7* that was not mapped. Its absence is possibly due to one or more interactions missing from current interaction databases. As a refined hypothesis, we left this influence intact and removed the others (Supplementary information). The deletion of *YAP6* in combination with other deletions embodies direct tests of the genetic interactions inferred explicitly in our model, and predicted effects of varying degrees on all genes.

These gene expression predictions were evaluated. Table I lists results for expression of all 1863 genes in our data set for the additional double-mutant strains. We also show results for a subset of genes determined to be filamentation-phenotype-correlated (Mode-2 genes; see below). To assess the accuracy of the model, we performed χ^2 tests over all data and determined the likelihood of the result from a χ^2 distribution (Supplementary information). We repeated the predictions using a linearly additive control model that lacks influences between seed genes and hence does not generate genetic interactions (Supplementary information). We computed the relative probability of the χ^2 fits of the genetic-interaction model and the control model to determine the likelihood that the genetic-interaction model performed significantly better. Relative to the control, our model provided an improvement in fit across almost all genes rather than large fit improvements for a small subset of genes. The genome-wide improvement is highly significant (Table I), and provides direct evidence for

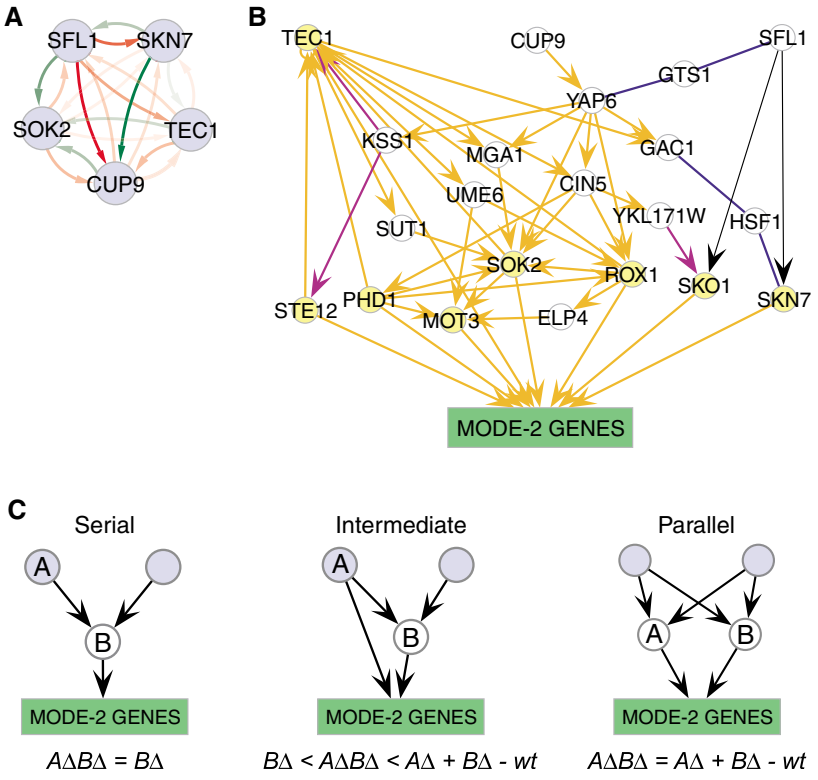

Figure 4 Seed gene influence network, filamentation-specific molecular network, and topological motifs. (**A**) Genetic influences between the seed genes inferred from genetic interactions. Green (red) arrows represent positive (negative) influence on regulatory activity. Color intensity is proportional to influence magnitude. (**B**) Mode-2 molecular network. The green box represents the Mode-2 genes with positive expression influences from the three topmost seed genes (*TEC1*, *CUP9*, and *SFL1*). Seed genes influence each other as in (A). Yellow nodes are transcription factors with enriched binding targets among the influenced genes. White nodes are proteins that fall on the shortest directionally consistent putative paths of influence from each seed gene to yellow transcription factors. Interactions are colored as: protein–protein in blue, protein–DNA in orange, and protein phosphorylation in violet. Black arrows denote inferred influences for which no molecular path with fewer than five interactions was found. Genes *TEC1*, *CUP9*, and *MGA1* are themselves members of the Mode-2 gene set. (**C**) Network topologies. The three network motifs and corresponding phenotype predictions for novel double-mutant pairs in (B). Labeled nodes denote deleted genes and inequalities represent phenotype predictions. Gray nodes represent genes that influence the Mode-2 gene set (green box) and white nodes represent candidates for transmission of the influences. Edges represent paths of influence involving any number of nodes and physical interactions.

Table I Summary statistics for gene expression predictions

Genes	N	χ^2	χ^{2C}	$P(\chi^2)/P(\chi^{2C})$
All genes	5589	0.16	0.20	1.7×10^{-208}
Mode-2	855	0.25	0.33	9.1×10^{-40}

Statistics for double mutants *cup9*Δ *yap6*Δ, *sfl1*Δ *yap6*Δ, and *sok2*Δ *yap6*Δ. N is the number of predictions, and the subscript c refers to the additive control model (see text). χ^2 values are reduced. Relative probabilities are computed with the χ^2 distribution.

both the biological importance of genetic interactions and the accuracy of our modeling technique.

Predictions of filamentation phenotype

We next sought to predict the filamentous-growth phenotype for novel double-mutant strains by integrating filamentation phenotype data and gene expression data. To find a connection between these two data types, we performed SVD (Alter *et al*, 2000) on the gene expression data matrix for our 1863 genes and compared the results with filamentation measurements. SVD is an unsupervised algebraic method that mathematically separates a data matrix into a set of 'modes' determined by quantitative patterns within the data. Each mode is manifest in the data as a global expression-pattern component that contributes to the expression of each gene to a degree varying from negligible to predominant. We examined the expression patterns of the SVD modes for correlation with filamentation data (Supplementary Table S1) for all 16 strains. We found that SVD Mode 2, quantitatively the second-greatest expression component (Supplementary Figure S3), was best correlated with the phenotype data (Supplementary Figure S4; Supplementary information). This implies that the 285 genes (Supplementary Table S4) that strongly exhibit the Mode-2 expression component are a quantitative proxy for the filamentation phenotype, even though this component is not the most dominant pattern in the data. Supporting this conclusion, 'cell wall' is the most significantly enriched (data not shown) Gene Ontology (Ashburner *et al*, 2000) annotation among the Mode-2 genes, which include the prototypical filamentation gene *FLO11* encoding a cell-wall protein, as well as many other known filamentation genes. The results raise the possibility that other SVD expression components and cognate gene sets might correlate with other phenotypes affected by our perturbations, such as cell adhesion

Table II Prediction of double-mutant phenotype

Deletions		Network motif	AΔBΔ phenotype		Phenotype inequality	
*A*Δ	*B*Δ		Predicted	Observed	Predicted	Observed
*sfl1*Δ	*yap6*Δ	Serial	−	−	$A\Delta B\Delta = B\Delta < wt < A\Delta$	$\boldsymbol{A\Delta B\Delta = B\Delta < wt < A\Delta}$
*skn7*Δ	*mot3*Δ		+	+	$A\Delta < wt < B\Delta = A\Delta B\Delta$	$\boldsymbol{A\Delta < wt < B\Delta = A\Delta B\Delta}$
*cup9*Δ	*yap6*Δ		−	−	$A\Delta B\Delta = B\Delta < wt < A\Delta$	$\boldsymbol{A\Delta B\Delta = B\Delta < wt < A\Delta}$
*sok2*Δ	*rox1*Δ		+	+	$wt < B\Delta = A\Delta B\Delta < A\Delta$	$wt < B\Delta < A\Delta B\Delta = A\Delta$
*sfl1*Δ	*cin5*Δ	Intermediate	+	+	$wt = B\Delta < A\Delta B\Delta < A\Delta$	$wt = B\Delta < A\Delta B\Delta = A\Delta$
*tec1*Δ	*sko1*Δ		−	−	$A\Delta B\Delta < A\Delta < B\Delta < wt$	$A\Delta B\Delta = A\Delta < B\Delta < wt$
*cup9*Δ	*phd1*Δ		+	+	$B\Delta = wt < A\Delta B\Delta < A\Delta$	$B\Delta = wt < A\Delta B\Delta = A\Delta$
*cup9*Δ	*cin5*Δ		+	+	$B\Delta = wt < A\Delta B\Delta < A\Delta$	$B\Delta = wt < A\Delta < A\Delta B\Delta$
*sok2*Δ	*yap6*Δ	Parallel	wt	wt	$\boldsymbol{B\Delta < A\Delta B\Delta = wt < A\Delta}$	$\boldsymbol{B\Delta < A\Delta B\Delta = wt < A\Delta}$
*sok2*Δ	*cin5*Δ		+	+	$B\Delta = wt < A\Delta = A\Delta B\Delta$	$B\Delta = wt < A\Delta < A\Delta B\Delta$
*sok2*Δ	*phd1*Δ		+	+	$B\Delta = wt < A\Delta B\Delta = A\Delta$	$B\Delta = wt < A\Delta B\Delta < A\Delta$
*skn7*Δ	*cin5*Δ		−	−	$A\Delta = A\Delta B\Delta < wt = B\Delta$	$\boldsymbol{A\Delta = A\Delta B\Delta < wt = B\Delta}$
*skn7*Δ	*sko1*Δ		−	−	$A\Delta B\Delta < B\Delta < A\Delta < wt$	$\boldsymbol{A\Delta B\Delta < B\Delta < A\Delta < wt}$

All 13 double gene deletion (*A*Δ*B*Δ) phenotypes were predicted correctly (P=0.0002) and six (in bold type) of 13 phenotype inequalities were predicted correctly (P=0.01) based on the Mode-2 network (Figure 3). Network motifs are illustrated in Figure 4c.

(Robertson and Fink, 1998). More generally, SVD may be a quantitative unbiased approach to associate distinct expression patterns with specific phenotypes obtained from other assays.

The Mode-2 genes are significantly enriched with genes bound by eight transcription factors (Supplementary Table S5), of which six (*TEC1*, *STE12*, *PHD1*, *SOK2*, *ROX1*, and *SKN7*) were known to have filamentous-growth-related phenotypes. Subsequently, we found that deletions of the other two (*MOT3* and *SKO1*) have filamentation phenotypes (see below).

Three (*TEC1*, *CUP9*, and *SFL1*) of the five transcription factor seed genes had significant influences on the Mode-2 expression component (Supplementary information). These influences were positive, and represent the genetically 'direct' effects of these seed genes on the expression of the Mode-2 genes. We found the shortest paths of molecular interactions to connect these three seed genes to the enriched transcription factors, although in some cases involving *SFL1*, no paths shorter than five interactions could be found. This procedure generated the Mode-2 network (Figure 4B), mapping specific molecular paths of putative influence. Through inclusion in the Mode-2 network, this process implicated three additional transcription factor genes: *YAP6*, *CIN5*, and *UME6*.

To initially probe the predictive value of the Mode-2 network, we constructed deletions of the newly implicated transcription factor genes *YAP6*, *CIN5*, *UME6*, *MOT3*, and *SKO1* and assayed the filamentation phenotype. *UME6* deletion was lethal in the filamentation-competent $\sum$1278b yeast background. All other deletions of transcription factor genes implicated by the Mode-2 network showed a filamentation phenotype: *yap6*Δ and *sko1*Δ mutants had filamentation defects, the *mot3*Δ mutant was strongly hyper-filamentous, and the *cin5*Δ mutant was marginally hypo-filamentous.

We then tested the capacity of the model to predict specific phenotypes for 13 novel combinatorial deletions. Predictions were based on three topological motifs present in the Mode-2 network and corresponding quantitative expectations (Figure 4C; Supplementary information). The model rendered predictions of both the phenotype of the double mutant (hyper, hypo, or wild-type filamentation) and the exact inequalities that order the observed phenotypes of the wild type, the two single mutants, and the double mutant on a scale of filamentation. For example, if a deletion of gene *A* is hyper-filamentous and epistatic to a hypo-filamentous deletion of gene *B*, the inequality would be $B\Delta < wt < A\Delta = A\Delta B\Delta$. There are 75 possible inequalities (Drees *et al*, 2005); we predicted one of these for each of the 13 novel double mutants.

Table II lists the predictions and experimental observations of the double-mutant phenotypes and phenotype inequalities. We assessed the accuracy of the model predictions by comparison with results generated from a training set of 1809 genetic interactions for invasive growth (Drees *et al*, 2005), a closely related phenotype (Supplementary information). The model correctly predicted all 13 double-mutant phenotypes, which was a very unlikely outcome using the training set (P=0.0002). Six of the 13 phenotype inequalities proved correct, which is also a significant improvement over the training set (P=0.009) due to the much larger number of possible outcomes. Note also that all of the incorrect phenotype-inequality predictions differed minimally from the observed phenotype inequalities.

Discussion

The model of filamentous growth control based on the Mode-2 network (Figure 4B) contains many genes known to be involved in filamentation. For example, the MAP-kinase Kss1 is correctly implicated as passing a positive influence on gene expression by derepressing the transcription factors Tec1 and Ste12 (Madhani *et al*, 1997). The network also implicated new regulators of filamentation (Yap6, Mot3, and Sko1) that were verified experimentally, and proposes new information flows where molecular support is currently sparse, such as a positive influence from *SFL1* on binding targets of *SKO1*.

In addition to implicating genes, our approach was often able to correctly infer functional relationships between genes that control filamentous growth. This is evident in the broad success in predicting double-mutant expression profiles (Table I) and phenotypes (Table II). In particular, all

predictions involving *YAP6* deletions proved accurate for both phenotype (P=0.02) and phenotype inequality (P=0.006), suggesting that its regulatory role in the network was mapped correctly. Predictions based on parallel and serial network topologies (Figure 4C) were also superior to the training set (phenotypes P=0.03 for both; inequalities P=0.04 and P=0.006, respectively). Furthermore, the model was able to accurately predict both filamentation phenotypes and the phenotype inequalities of all four double mutants in which the single mutants had opposite phenotypes, which would have been highly unlikely using the training set (P=0.009 and P=0.0002, respectively). The model is further supported by successful gene expression predictions for novel double-mutant strains. Although the control model was able to recover basic trends in the data, the predictions from our genetic-interactions model were significantly more accurate (Table I).

Notwithstanding the successful performance of our approach, its linear approximation may over-simplify many functional relationships and may miss complicated regulatory effects that are not as relevant for modeling genome-wide transcript levels. Dynamic modeling of the seed gene network could encompass nonlinear, post-transcriptional influences and feedback loops that often lead to more complex effects. Potential transcriptional feedback loops are apparent for *CUP9* in the Mode-2 network (Figure 4B), as Tec1, Ste12, Phd1, Sok2, and Mot3 bind its promoter (Harbison *et al*, 2004; Borneman *et al*, 2006). Dynamic modeling of these small networks might explain, for example, how the seed genes interact to generate their diverse single-mutant phenotypes.

Our methods are designed for application to any system in which multiple interacting genes are linked to phenotypes. The genetic influences decomposition can be used to dissect genetic-interaction effects between any number of seed genes, and a greater number can be expected to result in inference of a more comprehensive network of interactions. Combined with molecular data integration, this suggests an iterative approach in which a gene implicated in the system (such as *YAP6* in the filamentation network) is taken as an additional seed gene in a subsequent round of experimentation and analysis. Furthermore, although we have exclusively used null alleles in this study, the method can incorporate hypomorphic and hypermorphic alleles by fitting the genotype matrix elements to appropriate activity values relative to wild type. Possible methods to estimate these values include assays of protein levels (or phosphorylated protein levels for phospho-activated regulators) and using results fit with a cognate null mutant to constrain all parameters other than the activity levels of the non-null mutant allele. The method is extensible and can also predict the effects of higher-order combinatorial genotypes, such as triple gene deletions, through removal of the influence coefficients associated with every perturbed gene and the paths in which they form a critical link. Finally, the genetic influences decomposition is formulated to be directly applicable to all quantitative phenotypes, not only gene expression, with the requirement that the number of phenotypes assayed for each strain be equal to or greater than the number of seed genes plus one (Supplementary information).

With the abundance of molecular interactions, there are often numerous possible paths of influence among gene products. Likewise, genetic interactions often have multiple possible molecular interpretations. By emphasizing the complementarity of these data types, our integration of genetic influences decomposition and molecular interaction data greatly constrained these possibilities and assigned specific functional significance to molecular interactions in a network model of the transcriptional control of filamentous growth. This model generated predictions that relied on both the accuracy of our genetic influence decomposition and our data integration strategy. The integration strategy exploited the availability of accurate, genome-scale molecular interaction data sets, and identified instances in which functionally important molecular data are missing. With the increasing availability of human interaction data (Stelzl *et al*, 2005) and further modeling developments to address allelic variation in outbred populations, similar quantitative and integrative techniques may ultimately be applied to disease-related models.

Materials and methods

Genetic influences decomposition

The genetic influences decomposition method can be illustrated with the simplified case of two seed genes A and B that influence the expression of two genes X and Y. For a strain genotype labeled with superscript S, we write a linear pair of equations for gene expression:

$$X^S = x_0 + x_A g_A^S + x_B g_B^S$$

$$Y^S = y_0 + y_A g_A^S + y_B g_B^S \quad (2)$$

The parameters x_A, x_B, and x_0 represent contributions to the expression of X from the gene A, gene B, and the remainder of the genetic background, respectively (similarly for gene Y). These parameters are independent of the strain genotype. The coefficients g_A^S and g_B^S are the inferred activity levels of the seed genes A and B in the strain background S, and are independent of the transcript being measured. Gene knockout strains are modeled by setting the activity of the deleted gene to zero, such as $g_A^{A\Delta}$=0 and $g_A^{A\Delta B\Delta}$=0 for strains with gene A deleted. Influences between seed genes (observed as genetic interactions) can be systematically and quantitatively inferred from changes in activity levels of one gene when the other gene is perturbed, and *vice versa* (Supplementary information). For example, $g_A^{B\Delta} < g_A^{wt}$ would evince a positive influence from gene B on the activity of gene A. Note that these activity changes are relative to wild type (all g^{wt}=1) and are calculated parameters. Rather than substituting transcript level data for these activities (as in many regression methods), these model-derived parameters conceptually include all levels of gene control from initiation of transcription to protein localization, modification, and degradation.

The system of equations in Equation (2) can be expanded to model an arbitrary number of gene expression measurements and perturbed seed genes by systematically adding parameters (Supplementary information). The equations can be recast in matrix form.

In the decomposition of our genomic expression data, the number of measurements in Equation (1) far exceeds the number of model parameters. We found the least-squares best-fit solution (Supplementary information).

Directed data integration

To identify functionally important expression influences, we first determined coefficients in the influence matrix, **X**, that were significantly different from zero (Supplementary information). For example, the influences of *TEC1*, *CUP9*, and *SKN7* on *DDR48* are mapped in Figure 2A. For each seed gene, we then identified a set of

genes that received positive expression influences from the seed gene, and another set of genes that received negative expression influences, for a total of 10 overlapping gene sets (Supplementary Table S2). We queried each gene set for enrichment of genes bound by each known transcription factor (Lee *et al*, 2002; Zeitlinger *et al*, 2003; Harbison *et al*, 2004; Borneman *et al*, 2006), thus defining candidate transcription factors for each gene set (Supplementary Table S3). We next searched public data for protein–protein (Bader *et al*, 2001; Xenarios *et al*, 2001), protein phosphorylation (Ptacek *et al*, 2005), and protein–DNA (Lee *et al*, 2002; Zeitlinger *et al*, 2003; Harbison *et al*, 2004; Borneman *et al*, 2006) interactions and constructed networks including only the shortest directionally consistent paths connecting each seed gene with the transcriptional regulators and hence the genes it influenced (Figure 2; Supplementary information) (Carter *et al*, 2006). We were able to connect seed genes to about half of their influence targets on average (Supplementary Table S3).

Genomic expression data have been deposited in the Gene Expression Omnibus, accession GSE5938.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Pamela Troisch for assistance with microarray preparation and Christine Aldridge, David Galas, Ilya Shmulevich, and R James Taylor for their contributions. This work was supported by grant P50 GM076547 from NIH. GWC and TG were supported in part by grant FIBR EF-0527023 from NSF. TG is a recipient of a Burroughs Wellcome Fund Career Award in the Biomedical Sciences.

References

Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* **97:** 10101–10106

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29

Avery L, Wasserman S (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet* **8:** 312–316

Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW (2001) BIND—the Biomolecular Interaction Network Database. *Nucleic Acids Res* **29:** 242–245

Barton NH, Keightley PD (2002) Understanding quantitative genetic variation. *Nat Rev Genet* **3:** 11–21

Borneman AR, Leigh-Bell JA, Yu H, Bertone P, Gerstein M, Snyder M (2006) Target hub proteins serve as master regulators of development in yeast. *Genes Dev* **20:** 435–448

Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* **5:** 618–625

Carter GW, Rupp S, Fink GR, Galitski T (2006) Disentangling information flow in the Ras-cAMP signaling network. *Genome Res* **16:** 520–526

Chou S, Lane S, Liu H (2006) Regulation of mating and filamentation genes by two distinct Ste12 complexes in *Saccharomyces cerevisiae*. *Mol Cell Biol* **26:** 4794–4805

Drees BL, Thorsson V, Carter GW, Rives AW, Raymond MZ, Avila-Campillo I, Shannon P, Galitski T (2005) Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol* **6:** R38

Dudley AM, Janse DM, Tanay A, Shamir R, Church GM (2005) A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol Syst Biol* **1:** 1

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391:** 806–811

Galitski T (2004) Molecular networks in model systems. *Annu Rev Genomics Hum Genet* **5:** 177–187

Gavrias V, Andrianopoulos A, Gimeno CJ, Timberlake WE (1996) *Saccharomyces cerevisiae* TEC1 is required for pseudohyphal growth. *Mol Microbiol* **19:** 1255–1263

Gimeno CJ, Ljungdahl PO, Styles CA, Fink GR (1992) Unipolar cell divisions in the yeast *S. cerevisiae* lead to filamentous growth: regulation by starvation and RAS. *Cell* **68:** 1077–1090

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99–104

Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2:** 343–372

Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* **23:** 561–566

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799–804

Lengeler KB, Davidson RC, D'Souza C, Harashima T, Shen WC, Wang P, Pan X, Waugh M, Heitman J (2000) Signal transduction cascades regulating fungal development and virulence. *Microbiol Mol Biol Rev* **64:** 746–785

Lorenz MC, Heitman J (1998) Regulators of pseudohyphal differentiation in *Saccharomyces cerevisiae* identified through multicopy suppressor analysis in ammonium permease mutant strains. *Genetics* **150:** 1443–1457

Madhani HD, Styles CA, Fink GR (1997) MAP kinases with distinct inhibitory functions impart signaling specificity during yeast differentiation. *Cell* **91:** 673–684

Prinz S, Avila-Campillo I, Aldridge C, Srinivasan A, Dimitrov K, Siegel AF, Galitski T (2004) Control of yeast filamentous-form growth by modules in an integrated molecular network. *Genome Res* **14:** 380–390

Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJ, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M (2005) Global analysis of protein phosphorylation in yeast. *Nature* **438:** 679–684

Robertson LS, Fink GR (1998) The three yeast A kinases have specific signaling functions in pseudohyphal growth. *Proc Natl Acad Sci USA* **95:** 13783–13787

Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308:** 523–529

Segre D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. *Nat Genet* **37:** 77–83

Shook DR, Johnson TE (1999) Quantitative trait loci affecting survival and fertility-related traits in *Caenorhabditis elegans* show genotype–environment interactions, pleiotropy and epistasis. *Genetics* **153:** 1233–1243

Sinha H, Nicholson BP, Steinmetz LM, McCusker JH (2006) Complex genetic interactions in a quantitative trait locus. *PLoS Genet* **2:** e13

Steffen M, Petti A, Aach J, D'Haeseleer P, Church G (2002) Automated modelling of signal transduction networks. *BMC Bioinformatics* **3:** 34

Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, McCusker JH, Davis RW (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416:** 326–330

Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell* **122:** 957–968

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C (2004) Global mapping of the yeast genetic interaction network. *Science* **303:** 808–813

Tonouchi A, Fujita A, Kuhara S (1994) Molecular cloning of the gene encoding a highly expressed protein in SFL1 gene-disrupted flocculating yeast. *J Biochem (Tokyo)* **115:** 683–688

Van Driessche N, Demsar J, Booth EO, Hill P, Juvan P, Zupan B, Kuspa A, Shaulsky G (2005) Epistasis analysis with global transcriptional phenotypes. *Nat Genet* **37:** 471–477

Ward MP, Gimeno CJ, Fink GR, Garrett S (1995) SOK2 may regulate cyclic AMP-dependent protein kinase-stimulated growth and pseudohyphal development by repressing transcription. *Mol Cell Biol* **15:** 6854–6863

Winzeler EA, Liang H, Shoemaker DD, Davis RW (2000) Functional analysis of the yeast genome by precise deletion and parallel phenotypic characterization. *Novartis Found Symp* **229:** 105–109 discussion 109–111

Wong SL, Zhang LV, Tong AH, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey H, Boone C, Roth FP (2004) Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci USA* **101:** 15682–15687

Workman CT, Mak HC, McCuine S, Tagne JB, Agarwal M, Ozier O, Begley TJ, Samson LD, Ideker T (2006) A systems approach to mapping DNA damage response pathways. *Science* **312:** 1054–1059

Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, Marcotte EM, Eisenberg D (2001) DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res* **29:** 239–241

Yang YL, Suen J, Brynildsen MP, Galbraith SJ, Liao JC (2005) Inferring yeast cell cycle regulators and interactions using transcription factor activities. *BMC Genomics* **6:** 90

Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR, Young RA (2003) Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113:** 395–404

Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C, Roth FP (2005) Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *J Biol* **4:** 6

Zhong W, Sternberg PW (2006) Genome-wide prediction of *C. elegans* genetic interactions. *Science* **311:** 1481–1484