PREDICTING THE EFFECTS OF COPY-NUMBER VARIATION IN DOUBLE AND TRIPLE MUTANT COMBINATIONS*

GREGORY W. CARTER†
The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA
Email: greg.carter@jax.org

MICHELLE HAYS
Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA
Email: mhays@systemsbiology.org

SONG LI
Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA
Email: soli@systemsbiology.org

TIMOTHY GALITSKI
Millipore Corporation, 290 Concord Road, Billerica, MA 01821, USA, and
Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA
Email: timothy.galitski@merckgroup.com

The study of genetic interactions is a powerful tool in inferring structure and function of biological networks. To date, genetic interaction studies have been dominated by pair-wise gene deletion screens. However, classical genetic analysis and natural genetic variation involve diverse gene forms ranging from null alleles to copy number variants. Moreover, genetic variation is typically multifactorial. Addressing multiple combinatorial genetic variations ranging in gene activity is therefore of critical value. We approach this problem using genetic network modeling that quantitatively encodes how genes influence the activity of one another and phenotype outcomes. A network model was initially inferred from linear decomposition of gene expression data. We used this network to predict the effects of combining multi-copy and deletion mutations of specific gene pairs and a gene triplet. Predicted expression patterns across hundreds of genes were experimentally validated. Prediction success was critically dependent on how a multi-copy gene interacted with other genes in the network model. This strategy provides a template for the inference, prediction, and testing of genetically complex hypotheses involving diverse genetic variation.

* This work was supported by grants P50 GM076547 from NIGMS and by the ISB-University of Luxembourg Program. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIMGS or the NIH.

† Work partially supported by grant K25 GM079404 from NIGMS.

1. Introduction

Understanding how genetic variation affects phenotype is a central challenge of modern genetics. The hundreds of disease-related genetic variants identified in genome-wide association studies[1,2] typically account for a fraction of the heritable phenotype variation.[3] One likely contributor to this "missing heritability" is that undetected genetic interactions amplify the effects of genetic variations. In natural populations these interacting variants arise from mutations that affect genetic activity in a wide range of hypomorphic and hypermorphic allele forms. Advances in high-throughput sequencing technologies now allow greater experimental detection of structural variation and other types of genetic diversity beyond single-nucleotide polymorphisms.[4] These genetic variants appear in arbitrary combinations within a population. Therefore analytical methods designed to analyze and predict how diverse allele forms combine to influence phenotypes will be a valuable tool in the study of systems genetics.

Studies in model organisms provide evidence that genetic interactions are both prevalent and diverse in their effects. Systems-genetic studies allow the inference of multiple genetic factors[5,6] and their interactions.[7-10] To date, large-scale genetic interaction studies have been dominated by pair-wise gene deletion screens in model organisms.[11-15] However, genetic variation typically results in partial losses or gains of function, and population mixing combines these alleles in combinations with more than pair-wise complexity.[16] Interactions between these variations can result in a rich spectrum of phenotypes[7] including health outcomes.[17] For example, multi-copy suppression (or rescue) is the reversal of the effect of one mutant gene by the overexpression of a second gene. Predicting such effects requires techniques that encompass continuous variation in gene activity rather than simplified binary (on/off) gene states and multi-gene models that allow assessment of arbitrary combinations of two or more perturbations. These polygenic effects are often best accounted for by network models of interacting system elements rather than models of genes that affect phenotypes independently.[18]

To address this problem we extend a previously-developed method of analyzing genetic interactions to construct quantitative network models.[18] We approach genetic interactions as quantitative influences, defined as positive or negative numbers of varying magnitude that account for the fraction of a measurable phenotype (e.g. the expression level of a gene) inferred to be caused by a system element (e.g. a gene product). The measured phenotype is modeled by multiple influences acting throughout the inferred network. Our method 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.[19] Genetically "direct" (not necessarily molecularly direct) effects from regulator genes on the phenotype are separated from the genetically "indirect" effects that involve genetic interactions between regulator genes. The model correctly implicated novel genes in a cell differentiation process and proved accurate in predicting phenotypes for novel combinations of genetic knockouts.

A key aspect of the model is that it infers the effective biological activity of each regulator, rather than assuming a correspondence between the regulator's gene and protein expression level.

This modeling strategy was based on the fact that the net contribution of a gene to a phenotype is the result of many steps including transcription, translation, and various post-translational modifications like phosphorylation. In our initial study, a wide range of activity levels were inferred for a set of regulators in response to variations in genetic background.[18] These inferred activities were indirectly validated when the model was used to successfully predict the outcome of novel pair-wise combinations of gene knockouts. This flexibility in inferred gene activity makes this modeling approach amenable to studying more diverse allele forms, such as multiple copies of a gene.

In this work, we assess the capacity of our modeling approach to predict the effects of a multi-copy genetic perturbation, and how that perturbation interacts with deletions of other network genes. Furthermore, we show that our models can predict the effects of triple-mutant combinations. Following our previous methods, we first inferred a model from gene expression data for single and pair-wise gene deletions. We then extended our methods to predict how multiple copies of one of the network genes modify the activity of all regulator genes in the network (including its own) and, in turn, the downstream effects on genome-wide gene expression patterns. These quantitative predictions were directly tested in the laboratory by constructing the required strains and collecting additional microarray data. The model accurately predicted complex patterns of gene expression.

2. Network Model Inference

A central result of our previous work was the mapping and verification of a molecular network that regulates the filamentous growth response in yeast.[18] Many of the genes and pathways in this network were previously implicated in filamentous growth, and most of the newly identified genes exhibited strong knockout phenotypes. We chose four genes from that network for further study: *TEC1*, *CUP9*, *CIN5*, and *YAP6*. *TEC1* and *CUP9* were two of the five transcription factor genes chosen in our previous study. The transcription factors Yap6 and Cin5 were implicated through that study as occupying central positions in the regulatory network and candidate transmitters of genetic influences from *TEC1* and *CUP9*. We chose *CIN5* as our gain-of-function gene to further explore how alternate mutations interact with other network genes. Furthermore, there was evidence that *YAP6* and *CIN5* were involved in feedback in the network, since *TEC1* and *CUP9* were mapped in the network as both upstream regulators and downstream targets of transcript regulation. Therefore the subnetwork comprising the genes *TEC1*, *CUP9*, *YAP6*, and *CIN5* is a suitable test case for modeling the effects of genetic copy-number perturbation in a highly interactive genetic network.

2.1.1. *Yeast Gene Expression Profiling*

We collected expression profiles in triplicates[20] of wild type and mutant strains grown under filamentous-form conditions for 10 hours, as previously described.[21] Target labeling with the Affymetrix GeneChip® One-Cycle Target Labeling kit and hybridization to Yeast Genome 2.0

Arrays was done according to the manufacturer's protocols. Microarray data were collected for yeast diploid strains of 15 genotypes including wild type, deletion mutants for each of the four transcription factors, all six double-deletion combinations, and five strains transformed with a plasmid bearing *CIN5* and its native promoter (Supplemental Material and Table S2). Microarray data were normalized using robust multi-array averaging (RMA)[22] as implemented in the BioConductor software package.[23] Each gene expression data point was taken as the mean of the three corresponding biological replicates. We verified that the mean variation between biological replicates was less than the mean variation between different strains. Expression intensities for each gene were transformed into Log2 ratios relative to yeast-form wild-type expression. We restricted subsequent analysis to a set of 1267 differentially expressed genes, defined as having a factor of two difference between their lowest and highest expression intensities.

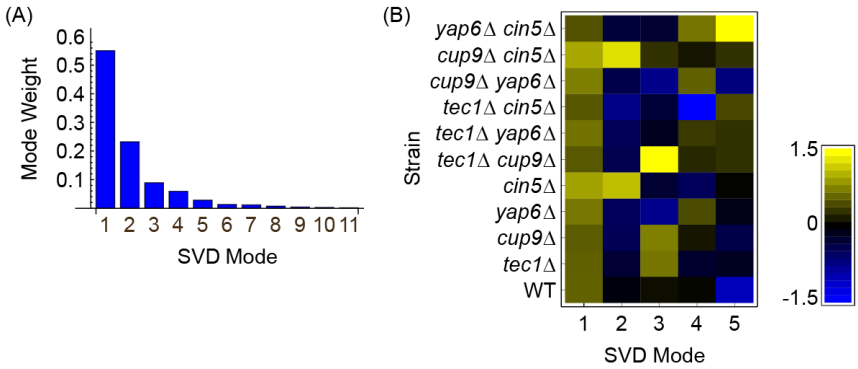

Figure 1. SVD eigenvalues and eigengenes matrix. (A) Bar chart of eigenvalues and (B) raster plot of eigengenes matrix are shown for the first five SVD modes.

2.1.2. *Singular Value Decomposition Analysis*

To identify the global expression patterns we performed singular value decomposition (SVD) on the matrix of differentially regulated genes (Figure 1).[24] 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, or eigengene, that contributes to the expression of each gene to a degree varying from negligible to predominant. We performed SVD on data for wild-type, single-deletion, and double-deletion strains (11 in total). Of the 11 resulting SVD modes, the first five modes account for 96% of the information in the data set. This provides support for the suitability of a linear model, because it is consistent with the dimensional reduction of 11 experimental conditions to five linearly independent modes (the four seed genes plus the collective remainder of the genome), plus a small noise component. We followed the procedures in previous publications[18,25] to identify gene sets positively and negatively associated with each mode and to query those sets for enriched Gene Ontology annotations and transcription factor targets.

2.1.3. *Genetic Influences Decomposition*

We derived a network model of genetic influences using genetic influences decomposition, a technique designed to dissect the complexities of genetic interactions.[18] We considered the data set of all single and double deletions of genes *CIN5, CUP9, TEC1,* and *YAP6*. We refer to these four genes hereafter as "seed genes" in the analysis. This procedure reformulates an expression data matrix D as the product of 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. 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. Note that other allele types can be accommodated readily with a measured or inferred 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 in terms of X and G. The form of matrix G guarantees the existence a unique best-fit solution due to the strict arrangement of ones and zeros required by the genotypes (i.e., the rows of matrix G are linearly independent). We used SVD to aid in finding the best-fit solution. Genetic-influences decomposition was performed on the first five SVD modes. The full influences matrix X was determined by multiplying the results by the SVD eigenarrays. Finding the best-fit solution was a tractable problem using commercial software (Mathematica) on a PC. In matrix notation, this procedure is summarized:

$$D = u.v.w^T \cong u.v.x.G = X.G \qquad (1)$$

where the symbol $\cong$ denotes a best-fit solution. The matrices u, v, and w^T are the singular value matrices for the first six modes. The eigengene matrix w^T is further decomposed into $x.G$ with a least-squares best fit. The 5 x 5 square matrix x encodes the expression influences for the first five eigengenes. The 1267 x 5 matrix X contains the expression influences for each gene.

To model how the seed genes influence one another, we inferred and quantified the influences the seed genes exert on each others' activity level. We stress that these parameters correspond to influences on inferred regulatory activity, rather than mRNA levels. Inferring interactions between seed genes is a further dimensional reduction of G into a matrix of gene-gene interactions, M, and a matrix of basal activities, G_0, with diagonal elements { G_0^{TEC1}, G_0^{CUP9}, G_0^{YAP6}, G_0^{CIN5}} and all off-diagonal elements equal to zero. Diagonal elements of M were set to zero because self-interactions cannot be mathematically distinguished from basal activity. Gene deletions are computed by taking the limit of the deleted genes basal activity to zero, which effectively removes that gene from the network. The equations used to calculate these quantities are written for each strain background (i.e. each column in G omitting the first row value of 1):

$$
\begin{aligned}
&\textit{Wild Type}: && (G)^{WT} = [(G_0)^{-1} - M]^{-1}.1 \\
&\textit{Single Deletion}: && (G)^{A\Delta} = lim_{G_0^A \to 0}[(G_0)^{-1} - M]^{-1}.1 \\
&\textit{Double Deletion}: && (G)^{A\Delta B\Delta} = lim_{G_0^B \to 0}\, lim_{G_0^A \to 0}[(G_0)^{-1} - M]^{-1}.1
\end{aligned}
\tag{2}
$$

The vector 1 is defined as {1, 1, …, 1} with length equal to the number of seed genes (four in this case). For the derivation of these equations see our previous work.[18] With the fit for G found above, we found least-squares best fit solutions for $G_0$and M as shown in Figure 2. We found strong negative influences from *CIN5* on the activities of *TEC1*, *YAP6*, and *CUP9*, with positive influences from *TEC1* and *YAP6* back to *CIN5*. The gene pairs *TEC1* and *YAP6* and *TEC1* and *CUP9* are both mutually influential with positive values. A major consequence of this network complexity is the possibility of non-linear behavior generated by the sum of linear influences.

To assess the goodness of fit for the genetic influences decomposition, we compared fitted double-mutant expression values with those predicted by an additive control model, in which the expression of gene X in a double-mutant background is:

$$X^{A\Delta B\Delta} - X^{WT} = (X^{A\Delta} - X^{WT}) - (X^{B\Delta} - X^{WT}) \tag{3}$$

This estimates the expression of every double mutant as the sum of effects for the two single mutants. To identify the most significant influences, we performed a series of bootstrap cross-validations by repeatedly re-analyzing half of the 1267 genes.[18] We obtained approximately 18000 solutions and defined significant influences as those with a mean more than four standard deviations from zero, which correspond to an empirical significance of $p = 6 \times 10^{-6}$. These influences are shown in Figure 2. Although these were the dominant influences, all influences (Figure S2 and Table S3) were used in calculations.

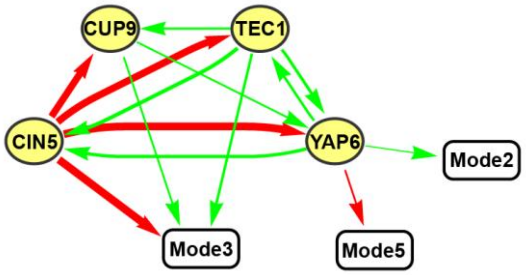

Figure 2. Network of significant positive (green) and negative (red) gene-to-gene and gene-to-expression influences. Yellow nodes are regulator genes, white nodes are expression patterns shared by multiple genes. Each edge corresponds to a parameter inferred in the genetic model with width proportional to influence magnitude.

We compared our model fit with the additive control (Eq. 3). For the expression of each gene, we computed the Pearson correlation between the experimental data and (i) the model fit and (ii) the fit to the additive control. The median correlation for the model fit was 0.88, compared to 0.56 for the additive control. Since the additive control is designed to fit the single mutants, the

difference in fit quality between model and control is much greater for the six double-mutants. For this subset of strains, the median correlations were 0.96 for the model and 0.36 for the additive control. We also assessed the fit for the expression of each gene in a given double-knockout strain relative to its expression in the wild-type strain, and summary statistics in terms of expression fold-change are shown in Figure 3. These results demonstrate that the genetic interactions account for much of the goodness of fit.

We identified the genes that significantly exhibit each SVD pattern of gene expression, allowing us to associate a positive-valued and negative-valued set of genes with each mode. Each set of genes was then queried for statistical enrichment of Gene Ontology annotations[26] and transcription-factor binding targets[27-30] (Supplementary Table 1). This allowed us to characterize the gene sets by the enriched annotations and regulators. Of particular interest, we identified the set of 197 genes that exhibit the Mode 3 expression pattern, defined as having eigenarray coefficients greater than one standard deviation above the mean[25]. This set of genes is enriched in carbohydrate metabolism (p = 5.5 x 10^{-7}) and targets of the following nine transcription factors: Phd1, Sok2, Sut1, Msn2, Rox1, Flo8, Mga1, Ume6, and Skn7 (Set 3-Positive, Table S1). No other set had this many enriched regulators. Four of these transcription factors (Phd1, Sok2, Rox1, and Skn7) were found to have enriched targets in filamentation-related genes in our earlier study, and three (Phd1, Sok2, and Rox1) were previously mapped downstream of the four seed genes.[18] Mga1 and Ume6 were implicated in the larger network that controls filamentation in that study. The remaining three (Sut1, Msn2, and Flo8) are known regulators of filamentation.[7,31] Given the number and functional coherence of transcription factors that regulate the Mode 3 gene set, we identified this as the most functionally relevant gene set for our filamentous growth study. While this is not the quantitatively largest expression pattern in the data, the greater expression patterns are signatures of more general processes such as cell growth (Mode 2) which generate more generic expression changes. Among the four seed genes, the Mode 3 expression pattern is positively influenced by *TEC1* and *CUP9* and negatively influenced by *CIN5* (Figure 2). *YAP6* indirectly influences this set via its influences on the other seed genes.

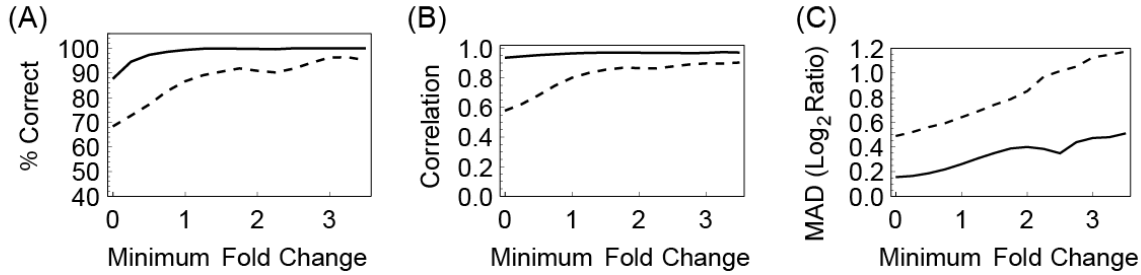

Figure 3. Fit accuracy for the six double-knockout strains (data combined). Solid lines are results for the interactive model; dashed lines are results for the additive control. Plots show (A) percentage of correct up or down-expression, (B) overall correlation of fit and data, and (C) median absolute deviation (MAD) for fit versus measured expression. All quantities are plotted as a function of fold-change, where the x-axis denotes the subset of expression ratios with the x-value or greater fold change.

2.2. *Predictions and Validation for a Multicopy Perturbation*

2.2.1. *Prediction for Multi-Copy Strains*

In previous work we limited our predictions to gene deletion strains. Here, we extend the method to address a multicopy allele. To predict the effects of a *CIN5* gain-of-function (*CIN5**) we replaced the basal level of *CIN5* activity, which is represented by the parameter G_0^{CIN5}. Network interactions adjust this value to produce the *CIN5* activities in the G matrix (Eq. 1). The precise increase in the value of G_0^{CIN5*} is difficult to fix for many reasons, including: the number of extra copies of *CIN5* borne by a plasmid with the 2μ replication origin is variable; microarray assays are performed on large cell populations and thus represent population-averaged signals; and the relationship between number of gene copies and regulatory activity is uncertain in our model. We find, however, that the predictions from our model are insensitive to increases in G_0^{CIN5*} for all values above 5 times the original value G_0^{CIN5} (which is equal to 0.58). This is demonstrated for an example gene in Figure 4. Therefore to predict global expression for *CIN5* multi-copy strains we artificially increased the basal activity value for *CIN5* as $G_0^{CIN5} \rightarrow G_0^{CIN5*} = 50G_0^{CIN5}$ with the factor 50 chosen as an asymptotic value. We computed new columns for the G matrix corresponding to these strains, following Eq. 2 with zero limit replaced by the multi-copy limit. Thus multiple copies of Gene A with a deletion of Gene B forms a column with 1 followed by:

$$Deletion\ and\ Multicopy: \qquad (G)^{A\Delta B*} = lim_{G_0^B \rightarrow 50G_0^B}\, lim_{G_0^A \rightarrow 0}[(G_0)^{-1} - M]^{-1}.1 \qquad (4)$$

For instance, the net activity level predicted for *CIN5*, g_{CIN5}^{CIN5*}, is 0.40. We then multiplied this new column in G by the previously-derived matrix X to produce a column of 1267 predictions for the expression of each gene for each of the four new strains.

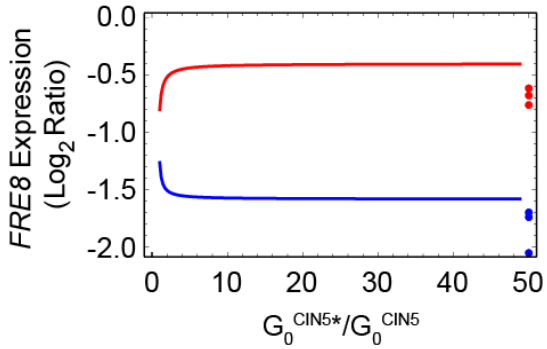

Figure 4. Predicted expression of an example gene (*FRE8*) is insensitive to basal *CIN5* activity beyond small values. The red line is predicted expression for *CIN5** strain background, and the blue line is for *yap6*Δ *CIN5** double mutant. Both lines begin at best-fit values for wild-type *CIN5* (multiplier of 1). Red and blue points are experimental results (in triplicate) for the corresponding strains.

That gene expression predictions rapidly approach asymptotic values that are stable for arbitrarily high values of G_0^{CIN5*} is a direct result of the self-regulating interactions between the

seed genes. Although we have not derived a dynamic model, the steady-state solution from our system of equations is sufficient to limit the effect of an arbitrarily high activity for the *CIN5* gene.

Although the model is composed of linear influences, the network of interactions predicts a highly nonlinear effect on gene expression for *CIN5**. The network model derived above (Figure 2) predicts that an attempt to exogenously increase the activity of *CIN5* will suppress the activities of *TEC1* and *YAP6*, which will in turn repress an increased activity of *CIN5*. Moreover, rather than predicting an increase in *CIN5* activity, the model predicts an effective reduction relative to the wild type. Our model predicts an activity level for *CIN5* in the *CIN5** strain to be 0.40 relative to the wild-type value fixed at 1. This behavior is the result of the genetic interactions inferred between the seed genes (Figure 2).

2.2.2. *Experimental Test of Predictions*

To test our predictions, we constructed a set of *CIN5* gain-of-function (*CIN5**) strains by transforming the wild-type, *tec1*Δ, *yap6*Δ, *cup9*Δ, and *tec1*Δ *cup9*Δ strains with a 2μ plasmid bearing the *CIN5* gene. The *CIN5* gene was cloned with its native promoter to amplify *CIN5* expression without modifying possible genetic interactions mediated by transcription. We collected microarray data as specified in Section 2.1.1. Our prediction of nonlinear behavior in *CIN5* activity was verified by gene expression for the *CIN5** strains, as our predictions matched measured levels very well. These predictions are summarized in Figure 5, which shows statistics for the expression of each gene in a given multi-copy strain relative to its expression in the wild-type strain (formatted as in Figure 3). Prediction accuracy increased with the magnitude of predicted effects (Figure 5), suggesting that predictive power was greatest above experimental noise. For the 777 expression predictions of over one-fold change relative to the wild-type expression, we found a Pearson correlation between experiment and data of 0.60 and agreement between up/down differential expression 81% of the time. For the 147 predictions of two-fold or greater change, the Pearson correlation was 0.71 and 90% of the predictions were directionally correct. These predictions can be contrasted with the additive model, which fares increasingly poorly as the fold-change increases in terms of correlation and directional accuracy (Figure 5). Most substantially, the additive model does not include regularizing genetic interactions and thus the numerical predictions are off by a factor proportional to the estimated number of extra gene copies (here set to 50) (Figure 5C). The model correctly predicted genome-wide expression *CIN5* gain-of-function effects and how these effects interact with deletions of *TEC1*, *CUP9*, and *YAP6*. In the context of our network model, the explanation for this effect is the inferred negative feedback from Cin5 to the three other seed genes.

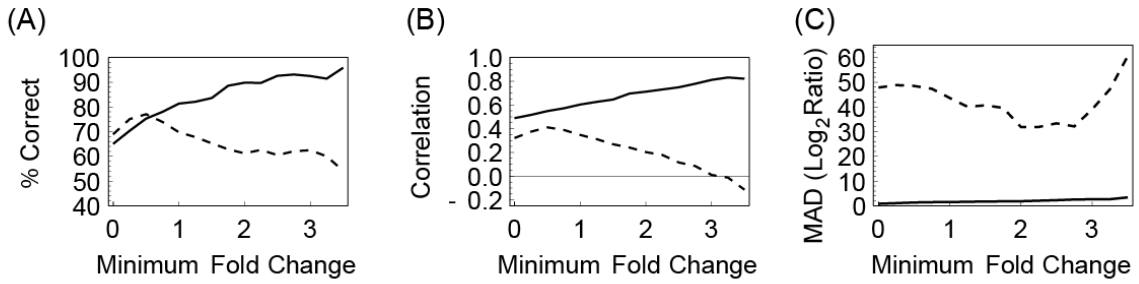

Figure 5. Prediction accuracy for *CIN5* gain-of-function in wild-type, *tec1Δ*, *cup9Δ*, *yap6Δ*, and *tec1Δcup9Δ* backgrounds (data combined). Solid lines are results for the interactive model; dashed lines are results for the additive control. Plots show (A) percentage of correct up or down-expression, (B) overall correlation of predictions and data, and (C) median absolute deviation (MAD) for predicted versus measured expression. All quantities are plotted as a function of fold-change, where the x-axis denotes the subset of expression ratios with the x-value or greater fold change. Predictive power generally increases with differential expression.

3. Discussion and Conclusions

The model we derived agrees well with the template of molecular interactions we mapped in our previous study.[18] In that study, we mapped genetic influences between *TEC1*, *CUP9*, and three other genes. *YAP6* and *CIN5* were implicated through the integration of molecular interaction data as candidate intermediaries of influences that traveled from *TEC1* and *CUP9* to downstream transcriptional targets. A number of the implications of that model were borne out in this study. For example, we mapped *CUP9* influence passing through Yap6 and Cin5 to regulate a set of filamentation-related genes. Indeed, our model shows *CUP9* influencing *CIN5* via *YAP6*. The seed gene with the greatest direct influence on the filamentation-related genes (Mode 3) in this study is *CIN5*, consistent with our earlier model that placed Cin5 downstream of the other three seed genes. Furthermore, our previous network exhibited strong hints of regulatory feedback involving the four genes studied in this work. The network derived in this study captures this type of feedback and it was experimentally validated by the feedback-dependent predictions.

The importance of an interactive network model was underscored by the fact that fit and prediction accuracy increased with the number of genetic perturbations involved. For the knockout data, the model fit generally predicted double-deletion expression values better than those of single deletions. This was also true for predictions of multi-copy *CIN5*. Our least successful predictions were for the *CIN5** strain, particularly for genes with minor expression changes from wild-type values. Overall, the effects of *CIN5** were small compared to *cin5*Δ effects. Therefore a possible reason for the relatively poor predictions is that the model was derived from the very strong effects of the *CIN5* deletion, which might have over-predicted the effects of multiple *CIN5* copies by assigning large values in the *X* matrix. However, when combined with one or two additional perturbations, the model was able to predict expression patterns with high accuracy. The *tec1*Δ *cup9*Δ *CIN5** triple-perturbation predictions were the most accurate and correctly captured the moderation of the double-knockout effects by the *CIN5* gain-of-function (Supplementary Figure S1). For genes with one or greater fold change, 98% of predictions were

correct compared to 81% for all strains (Figure 5A) and the correlation between prediction and experiment was 0.88 compared to 0.60 for all strains (Figure 5B). Predicting such complex interactions necessarily requires a network model.

Although our focus in this study was forward prediction, the modeling methods could be used to infer the relative activity levels of uncharacterized alleles in novel data. This would be a straightforward application of the influences decomposition defined in Eq. 1. For example, if we treat our validation data as an independent data set and directly fit the activity level for *CIN5* in a *CIN5** strain (*i.e.*, the parameter G_0^{CIN5*}) the result is 0.75. This is an inferred reduction from the wild-type baseline of 1, and similar to our forward predicted value of 0.40. Thus for alleles with uncertain activity, our approach has the flexibility to make either reverse inferences on experimental data or forward predictions based on training data (knockout data in this instance). Reverse inference may prove particularly useful when addressing populations with natural genetic variation, because in many cases the trait-associated genes will harbor partial loss or gain of function alleles rather than complete gene deletions or multiplications.

We note that while our linear modeling technique accurately predicted general trends in gene expression, it did not always accurately predict precise effects. Inaccuracies in prediction were primarily due to errors in magnitude rather than misprediction of overall trends of up and down-regulation. The model generally over-predicted the effects of novel genetic perturbations, suggesting that there are additional genes which attenuate the perturbations we introduced. This is not surprising given that we have only analyzed four genes embedded in a highly complex system. However it is also certain that the linear assumption of our influence decomposition over-simplifies the complex biochemical relationships between gene products. Thus we view the resulting genetic influences model (Figure 2) as a template for more detailed biochemical models that require additional parameters (*e.g.* kinetic rate constants) but provide more precise predictions.

Predicting the effects of multifactorial genetic variation will require understanding how the variant genes operate in a network of genetic interactions. In this work we were able to demonstrate the power of a network inferred from gene knockouts to predict combinations of both hypermorphic and hypomorphic genetic variants. Prediction accuracy depended on genetic interactions being included in the model. This predictive power was a direct result of the feedback interactions in the inferred model. As the potential activity of the *CIN5* gene grows, interactions with the other network genes quickly stabilize its actual activity and, therefore, downstream phenotypes like global gene expression. We stress that this behavior was predicted in a model inferred from genetic deletions of *CIN5*, and the model was capable of predicting the outcomes of a genetic perturbation with a priori unknown effects on the levels of activity of the gene product. This is potentially an important capability when addressing genetically diverse populations with uncharacterized alleles. Indeed, genetic studies are increasingly revealing the role of diverse genetic variants, such as copy-number variation and promoter polymorphisms, in human disease. The capability to model and predict how these variants interact with null mutations will be of critical value in genomic medicine.

4. Supplementary Material

Supplementary Material located at http://www.jax.org/research/faculty/carter/supplement.html contains yeast strain information, supplementary tables, and a supplementary figure.

5. Acknowledgments

We are grateful to Susanne Prinz for helpful advice and comments on the manuscript.

References

1. M. M. Iles, *PLoS Genet* **4**, e33, (2008).
2. P. M. Visscher & G. W. Montgomery, *JAMA* **302**, 2028, (2009).
3. T. A. Manolio, F. S. Collins, N. J. Cox *et al.*, *Nature* **461**, 747, (2009).
4. R. E. Mills, K. Walter, C. Stewart *et al.*, *Nature* **470**, 59, (2011).
5. A. M. Dudley, D. M. Janse, A. Tanay, R. Shamir & G. M. Church, *Mol Syst Biol* **1**, 2005 0001, (2005).
6. J. Zhu, B. Zhang, E. N. Smith *et al.*, *Nat Genet* **40**, 854, (2008).
7. B. L. Drees, V. Thorsson, G. W. Carter *et al.*, *Genome Biol* **6**, R38, (2005).
8. R. P. St Onge, R. Mani, J. Oh *et al.*, *Nat Genet* **39**, 199, (2007).
9. A. H. Tong, G. Lesage, G. D. Bader *et al.*, *Science* **303**, 808, (2004).
10. N. Van Driessche, J. Demsar, E. O. Booth *et al.*, *Nat Genet* **37**, 471, (2005).
11. M. Costanzo, A. Baryshnikova, J. Bellay *et al.*, *Science* **327**, 425, (2010).
12. B. Lehner, *J Exp Biol* **210**, 1559, (2007).
13. S. L. Ooi, X. Pan, B. D. Peyser *et al.*, *Trends Genet* **22**, 56, (2006).
14. X. Pan, P. Ye, D. S. Yuan *et al.*, *Cell* **124**, 1069, (2006).
15. M. Schuldiner, S. R. Collins, N. J. Thompson *et al.*, *Cell* **123**, 507, (2005).
16. R. B. Brem & L. Kruglyak, *Proc Natl Acad Sci U S A* **102**, 1572, (2005).
17. S. Prabhakar, A. Visel, J. A. Akiyama *et al.*, *Science* **321**, 1346, (2008).
18. G. W. Carter, S. Prinz, C. Neou *et al.*, *Molecular systems biology* **3**, 96, (2007).
19. L. Avery & S. Wasserman, *Trends Genet* **8**, 312, (1992).
20. K. L. Thompson, B. A. Rosenzweig, R. Honchel *et al.*, *Mol Carcinog* **32**, 176, (2001).
21. S. Prinz, I. Avila-Campillo, C. Aldridge *et al.*, *Genome research* **14**, 380, (2004).
22. R. A. Irizarry, B. Hobbs, F. Collin *et al.*, *Biostatistics* **4**, 249, (2003).
23. R. C. Gentleman, V. J. Carey, D. M. Bates *et al.*, *Genome Biol* **5**, R80, (2004).
24. O. Alter, P. O. Brown & D. Botstein, *Proc Natl Acad Sci U S A* **97**, 10101, (2000).
25. G. W. Carter, S. Rupp, G. R. Fink & T. Galitski, *Genome Res* **16**, 520, (2006).
26. M. Ashburner, C. A. Ball, J. A. Blake *et al.*, *Nat Genet* **25**, 25, (2000).
27. A. R. Borneman, J. A. Leigh-Bell, H. Yu *et al.*, *Genes Dev* **20**, 435, (2006).
28. C. T. Harbison, D. B. Gordon, T. I. Lee *et al.*, *Nature* **431**, 99, (2004).
29. K. D. MacIsaac, T. Wang, D. B. Gordon *et al.*, *BMC Bioinformatics* **7**, 113, (2006).
30. J. Zeitlinger, I. Simon, C. T. Harbison *et al.*, *Cell* **113**, 395, (2003).
31. H. Liu, C. A. Styles & G. R. Fink, *Genetics* **144**, 967, (1996).