

Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks.

Supplemental Information

Diego di Bernardo^{*†}, Michael J. Thompson^{*‡}, Timothy S. Gardner^{*‡},
Sarah E. Chobot[§], Erin L. Eastwood^{§¶}, Andrew P. Wojtovich[§],
Sean J. Elliott[§], Scott E. Schaus^{§¶} & James J. Collins[‡]

*These authors contributed equally.

[†]Telethon Institute for Genetics and Medicine, Naples, Italy

[‡]Center for BioDynamics and Department of Biomedical Engineering

[§]Department of Chemistry

[¶]Center for Chemical Methodology and Library Development
Boston University, Boston, Massachusetts, USA

1 Model structure

To predict the mode of action of compounds, the MNI algorithm first infers a model of regulatory influences in a cell. We assume that only measurements of transcript concentrations in a cell are available. Thus, our model relates changes in gene transcript concentrations to each other. In particular, we use the following ordinary differential equation model to represent the rate of synthesis of a transcript as a function of the concentrations of every other transcript in a cell:

$$\dot{y}_i = f_i(y_1, \dots, y_N, u_i), \quad (1)$$

where $f_i(\dots)$ is a nonlinear influence function for transcript i , y_i is the concentration of transcript i , N is the number of transcripts measured, and u_i is the net external influence on the rate of synthesis of transcript i . An external influence is any effect on the rate of transcription of a gene that cannot be represented as a function of changes in the concentrations of the N transcripts. Examples of external influences include protein activity or metabolite concentration changes due to the action of a compound, and environmental stress on a cell.

The influence functions f_i are not generally known and must be inferred from the experimental measurements of the transcript concentrations. Determination of the exact functional form of f_i for each gene in a cell would require measurement of transcript concentrations under an infeasibly large number of experimental conditions. To make this inference problem tractable, we choose the following functional form for f_i :

$$f_i(y_1, \dots, y_N, u_i) = u_i \prod_j y_j^{n_{ij}} - d_i y_i, \quad (2)$$

where n_{ij} is a parameter describing the influence of transcript j on transcript i , and d_i is the rate of degradation of transcript i . This model structure can be viewed as a simplification of Hill-type transcription kinetics [1]. Although it is an imperfect representation of the influence functions f_i , the model is capable of capturing rough functional relationships between transcripts including nonlinear relationships common in gene regulation, such as combinatorial integration of regulatory influences by a promoter. Moreover, the model structure has relatively few parameters, each of which can be efficiently estimated (as described below). Thus, the model minimizes the number and complexity of experiments required to estimate model parameters.

We assume that all measurements of RNA are obtained under steady-state experimental conditions. Thus, Equations 1 and 2 become:

$$\dot{y}_i = 0 = u_i \prod_j y_j^{n_{ij}} - d_i y_i. \quad (3)$$

Because DNA measurement technology currently allows only the measurement of concentrations relative to a baseline (e.g., expression ratios), we make the following transformations of the model:

$$\frac{y_i}{y_{ib}} = \left(\frac{u_j}{u_{jb}} \right) \prod_j \left(\frac{y_j}{y_{jb}} \right)^{n_{ij}}. \quad (4)$$

Taking the logarithm of both sides yields:

$$\log_{10} \left(\frac{y_i}{y_{ib}} \right) = \log_{10} \left(\frac{u_j}{u_{jb}} \right) + \sum_j n_{ij} \log_{10} \left(\frac{y_j}{y_{jb}} \right). \quad (5)$$

By substituting variables and parameters, we obtain the following linear network model:

$$\sum_j a_{ij} x_j = -p_i, \text{ where} \quad (6)$$

$$\begin{aligned} a_{ij} &= n_{ij}, & j &\neq i, \\ a_{ij} &= n_{ij} - 1, & j &= i, \\ x_j &= \log_{10} \left(\frac{y_j}{y_{jb}} \right), & \text{and} \\ p_i &= \log_{10} \left(\frac{u_j}{u_{jb}} \right). \end{aligned}$$

The model coefficients, a_{ij} , of this model represent the influence of the concentration of transcript j on the rate of synthesis of transcript i . The variables x_j are the log-transformed expression-change ratios of each transcript, and the variable p_i is the change ratio of the net external influences on the synthesis of transcript i . Any value of p_i that is significantly different from 1.0 means there is an influence acting on transcript i that cannot be described by the concentration changes in the RNA transcripts. The prediction of these external influences, p_i , using the model coefficients, a_{ij} , and the expression data, x_j , is the objective of the MNI algorithm.

Due to the simplicity of the model of Equations 1 and 2, it cannot capture some basic characteristics of the influence functions, f_i . For example, the rate of transcription of a gene must saturate at some maximum value, but this model does not exhibit saturation. Consequently, the model may only provide accurate predictions for experimental conditions near those used to acquire the training data set. Nevertheless, as we show in this work, the model is sufficiently accurate to correctly identify the mode of action of compounds using a technically and economically feasible experimental set-up. An additional limitation of the model is that many dynamic variables in a cell, such as protein concentration, protein activity and metabolite concentration, are not observed experimentally, or included in the model (hidden variables). However, the effect on the network of changes in the activity of hidden variables may be captured in the model by the external influence variables (u_i). In addition, this information may be used, in conjunction with additional data or experiments, to identify the proteins, metabolites or other factors responsible for the external influence.

2 Model inference

The first task of the MNI algorithm is to learn the network model coefficients, a_{ij} , using a training data set of transcript expression data, x_{jl} . We assume that the expression data

consists of transcript measurements in which cells have been treated with compounds or environmental stresses in a series of M experiments. Thus for each transcript in the network, we can write M equations:

$$\sum_j a_{ij}x_{jl} = -p_{il}, \quad l = 1, \dots, M, \quad (7)$$

where l is the index for each experiment. Provided a sufficient number of independent experiments and a set of measurements of the expression data, x_{jl} , and the external influences, p_{il} , it is possible to calculate the model coefficients, a_{ij} , using multiple regression [2]. However, we assume that the training data are obtained using perturbations for which the external influences are not known or measured. Thus, Eq. 7 is ill-posed and cannot be solved directly.

To enable solution of the system without measurements or knowledge of p_{il} , we use the following strategy. First, we assume that any given treatment will directly influence a small fraction of the thousands of transcripts in a genome. (A large fraction of genes may nevertheless show a transcription response in each experiment due to propagation of the external influence through the network.) Thus, most p_{ij} will be equal to zero. If, for a given gene i , we can identify all experiments in which the gene has not been externally influenced, we can write the following equation:

$$\sum_j a_{ij}x_{jh} = 0, \quad (8)$$

where h includes all experiments in which gene i has not been perturbed. This reduced set of points represents a level set of the solution to Eq. 7. To determine a non-trivial solution to Eq. 8, we note that $a_{ii} = n_{ii} - 1$ must be nonzero, even when the direct self-feedback, n_{ii} , is zero. Thus we can set a_{ii} to an arbitrary constant and solve for the remaining coefficients using a regression strategy such as that presented in Appendix 1. The solution will be correct within an undetermined scaling factor, and will represent the relative influences of each transcript on the rate of synthesis of gene i .

It remains to determine all experiments in which gene i has not been externally influenced. To this end, we use the following recursive scheme. In step (1), we make an initial guess, \hat{a}_{ij} , of the model coefficients. Typically, we choose $\hat{a}_{ij} = -1$ for $i = j$, and $\hat{a}_{ij} = 0$ otherwise. This guess represents gene i as not being regulated by any gene.

In step (2), we use \hat{a}_{ij} to calculate an estimate, \hat{p}_{il} , of the external influences, p_{il} , from Eq. 7. An external influence is considered significant if it satisfies:

$$\hat{p}_{il} \geq \theta \cdot \max_{1 \leq l \leq M} (|\hat{p}_{il}|), \quad (9)$$

where θ is the significance threshold. We choose $\theta = 0.25$, i.e., an estimate of the external influence on gene i is considered significant if it is greater than 25% of the maximum absolute value of p_{il} in all experiments $l = 1, \dots, M$. There is some flexibility in choosing the threshold. Decreasing the threshold leads to more false positive predictions of significant

external influences. Through experimentation with simulated data sets and the yeast data set, we found that false negative predictions (i.e., missed detections) of external influence were more detrimental to algorithm performance than false positives. Thus, the threshold was chosen to err towards identifying more false positive predictions. (Note, the prediction of external influences may be further improved by using confidence intervals on p_{il} estimated from the regression statistics, though this was not explored in the present study.)

In step (3) of the recursion, we remove from the training data set all experiments in which the estimate of the external influence is significant. We then obtain a new estimate, \hat{a}_{ij} , of the model coefficients using Eq. 8 with the remaining experiments. We then return to step (1) of the recursion using the newly estimated coefficients and iterate until the estimates \hat{a}_{ij} and \hat{p}_{il} converge.

3 Prediction of mode of action of compounds

Once we have estimated the network model using the recursive procedure, we can predict the targets of any compound. In what follows we will use the subscript \mathbf{c} to indicate quantities that refer to a compound, i.e., $x_{j\mathbf{c}}$ is the expression of gene j in response to treatment with compound \mathbf{c} . With the expression profile for the compound, $x_{j\mathbf{c}}$, $i = 1, \dots, N$, and the model coefficients \hat{a}_{ij} , we use Eq. 7 to obtain an estimate of the external influences on each gene i :

$$\sum_j a_{ij} x_{j\mathbf{c}} = -\hat{p}_{i\mathbf{c}} \quad (10)$$

The targets of a compound are those genes calculated to have the most significant external influences. The significance of $\hat{p}_{i\mathbf{c}}$ is determined by computing a z-score for each gene i as:

$$z_{i\mathbf{c}} = \frac{\hat{p}_{i\mathbf{c}}}{\sigma_{i\mathbf{c}}}, \quad (11)$$

where $\sigma_{i\mathbf{c}}$ is the standard deviation on $\hat{p}_{i\mathbf{c}}$ computed by applying the propagation of error to Equation 7:

$$\sigma_{i\mathbf{c}}^2 = \sum_j \sum_k \sigma_{jk}^i x_{j\mathbf{c}} x_{k\mathbf{c}} + \sum_j \hat{a}_{ij}^2 \text{var}(x_{j\mathbf{c}}). \quad (12)$$

and σ_{jk}^i are the elements of the covariance matrix, Σ_a , for the parameters \hat{a}_{ij} in Equation 10, calculated as described in Section 7. Thus, σ_{jk}^i is the coefficient in the j^{th} row and k^{th} column of the covariance matrix, Σ_a , for gene i . The genes are then ranked according to the magnitude of their z-score.

Note, the same calculation can be applied to the expression profiles used to form the training data. Thus, the algorithm can identify the most likely targets of compounds used as training data. Conversely, any expression profile obtained for a new compound may be included in the training data to obtain an improved estimate of the network model.

4 Dimensional reduction

Typically the number of independent experiments in the training data, M , is much less than the number of genes in a cell, N . For example, in the present study we analyzed a data set of 515 experiments in which more than 6000 transcripts were measured. This means that Eq. 8 is underdetermined. Additional constraints must be applied to find a unique solution. It has been noted that regulatory networks have a sparse structure [3–5]; thus, one strategy is to use subset regression to identify a small set of non-zero model coefficients for each gene i . Another strategy is to make use of the fact that many genes are similarly regulated and share highly correlated expression profiles. Thus, genes may be associated with a reduced set of “characteristic” expression profiles. For example, expression profiles for the genes may be clustered, and the average profile for the cluster used in subsequent processing.

Here we use a strategy for dimensional reduction based on singular value decomposition (SVD). Like clustering, the approach makes use of the fact that the expression profiles for the N genes may be approximated by a smaller set of characteristic expression profiles. Using SVD, we first identify the principal components of the expression profiles for the N genes. Writing the training data set as a matrix, $X = x_{il}$, $i = 1, \dots, N$, $l = 1, \dots, M$, we can use singular value decomposition to obtain:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \quad (13)$$

where \mathbf{U} is an $N \times M$ matrix, and \mathbf{S} is a diagonal matrix of dimension $M \times M$ containing the singular values of \mathbf{X} , and \mathbf{V} is an $M \times M$ matrix containing the principal components of the gene expression profiles in columns. We then choose Q principal components ($Q < M$) associated with the largest singular values. These Q profiles serve as the characteristic expression profiles for the N genes and together describe most of the expression variation represented among the N genes. The characteristic profiles (“metagene” profiles) can be used to approximate \mathbf{X} as follows:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{U}_Q \mathbf{S}_Q \mathbf{V}^T, \quad (14)$$

where \mathbf{U}_Q is an $N \times Q$ matrix and contains only the first Q columns of \mathbf{U} , and \mathbf{S}_Q is a $Q \times M$ diagonal matrix of the top Q singular values. Defining matrix $\mathbf{Z} = \mathbf{S}_Q \mathbf{V}^T$, we can rewrite Eq. 14 as:

$$\hat{\mathbf{X}} = \mathbf{U}_Q \mathbf{Z}, \quad (15)$$

or as:

$$\mathbf{U}_Q^+ \mathbf{X} = \mathbf{Z}, \quad (16)$$

where \mathbf{U}_Q^+ is the pseudo-inverse of \mathbf{U}_Q . Matrix \mathbf{Z} , which is of reduced dimension $Q \times M$, can be interpreted as the expression profiles of the Q metagenes. We use \mathbf{U}_Q^+ to project the N dimensional expression data into the lower dimensional space of the metagenes (Eq. 16). We then apply the recursive algorithm described in Sections 2 and 3 to \mathbf{Z} to identify a

network model for the metagenes, and estimate the external influences on the metagenes. Finally, we use \mathbf{U}_Q to project the estimated external influences on the metagenes back into N -dimensional gene space (Eq. 15). We also calculate standard deviations on the estimates and transform them between gene and metagene spaces using the propagation of error formula [6]. To determine the number of significant singular values (metagenes), we used an approach by Everitt and Dunn [7] of identifying those singular values that have a relative variance above some threshold value. The relative variance of a singular value is calculated as the square of that singular value divided by the sum of the squares of all of the singular values. Each singular value was considered significant if its relative variance was greater than the threshold of $0.7/n$, where $n = 515$ is the number of experiments. Based on this approach, we used $Q = 117$ singular values (metagenes).

5 Improving the specificity of target prediction

Many genes in the training data set, \mathbf{X} , will share very similar expression profiles over the M experiments. These genes will be difficult to distinguish. This problem is amplified by the use of dimensional reduction techniques which tend to average out or discard the few differences that do exist between similarly expressed genes. Thus, when estimating external influences on these genes (i.e., predicting compound targets), many genes will be identified with similar significance. Hence, the algorithm would identify many false positive targets in addition to the correct targets.

To improve the specificity of the algorithm, we adopt a ‘‘tournament’’ approach. For a given expression profile obtained following a treatment with a test compound, we apply the algorithm of Sections 2–4, in three successive iterations. In the first iteration, we rank all genes measured in the test expression profile. We then select the 1/3 of the genes ranked highest (which is approximately 2000 genes out of approximately 6000 in the yeast data set). We then reapply the algorithm to the selected genes and rank them. Once again, we apply the algorithm to the remaining genes and select the 1/3 highest-ranked genes (which is approximately 600 genes from the yeast data set). We then apply the algorithm one more time to obtain a final ranking.

The advantage of this approach is that the dimensional reduction via singular value decomposition preserves more of the differences between genes as the number of genes processed (N) becomes closer to the number of experiments (M). Thus with each successive application of the algorithm, differences between similarly regulated genes are more clearly identified and the specificity of target prediction is improved.

We also added a convergence check that stops the iterations if the *SVD-error* increases compared to the previous iteration. The *SVD-error* is simply $\|\hat{\mathbf{X}} - \mathbf{X}\|^2$. After each iteration the error should decrease since N becomes closer to M . If this does not happen, it means that the algorithm did not choose the proper genes; therefore the recursion is stopped and the final ranking is computed on the last ‘good’ iteration.

This approach has one limitation in that occasionally genes that are the true targets

of a compound are eliminated between successive applications of the algorithm. In other words, the genes are improperly eliminated in the early iterations of the algorithm when the specificity is still low. To overcome this problem, we modified the z-score used to rank genes prior to selection of the highest-ranked 1/3 of the genes. The modified z-score, z_{ic}^m , is:

$$z_{ic}^m = z_{ic} + \frac{x_{ic}}{\sigma_{x_{ic}}}, \quad (17)$$

where $\sigma_{x_{ic}}$ is the standard deviation on the expression ratio, x_{ic} . This score was designed to boost the likelihood of including genes with significant changes in the test expression profile. From simulations, we observed that using the modified MNI score to rank the genes sometimes improved the prediction of targets compared with using the standard z-score for the selection of genes.

To use the modified z-score, we ran the three iterations of ranking and gene selection using both scoring approaches. First we ran the three iterations using the standard z-score to select genes; then we ran the three iterations again using the modified z-score score to select genes. In the final iteration of both approaches, we ranked the genes using only the standard z-score. At the end, we are left with two ranked lists of genes ordered by their corresponding standard z-scores. Then we chose as the final ranking the list of genes with the highest mean standard z-score in the final iteration. The mean standard z-score for one list of gene is computed by taking the mean of the z-score of each gene in the list. For comparison, Tables S1 and S2 show the performance of the MNI algorithm with and without use of the modified z-score strategy.

6 Additional results: identifying target pathways and genes using association analysis and mRNA expression change

As mentioned in the main text, it is possible to use the gene rankings obtained by the MNI algorithm to identify the pathway in which a compound target operates. To identify the pathway, we used the GO Term Finder tool (www.yeastgenome.org) to identify GO ontologies that were significantly (i.e., $p \leq 0.01$) shared by the 50 highest-ranked genes according to the MNI algorithm.

For comparison, we also determined the pathways identified among the 50 highest-ranked genes according to the significance of their RNA expression changes. Of the nine compounds with known or probable targets and pathways, the MNI algorithm correctly identified the target pathway for seven compounds, while ranking by mRNA change (z-score) identifies pathways matching the known target pathways of five compounds (Table S3). However, while ranking based on mRNA change identified several target pathways, the target genes were often not highly ranked. For instance, the targets of only three compounds (hydroxyurea, cycloheximide, and dyclonine) were ranked in the top 50 by largest mRNA expression change (Table S2). Most notably, ranking based on mRNA expression change alone failed to identify either the discovered targeted pathway or the

discovered targeted genes of the novel compound, PTSB, examined in this study.

The association analysis approaches were also examined for their ability to rank the known gene and pathway targets of the compounds (Tables S2 and S4). Of the nine compounds with known or probable targets and pathways, the linear combination approach [8] correctly identified the target pathway for four compounds, while ranking with the correlation method [9, 10] identifies pathways matching the known target pathways of three compounds (Table S4). The linear combination approach ranked the known gene targets of five of the nine compounds in the top 50 (Table S2). The correlation method ranked the known gene targets of four of the nine compounds in the top 50. The targets of the other compounds were either not ranked highly or not themselves perturbed in the training set, and could therefore not be identified by these approaches. Most notably, ranking by both association analysis approaches failed to identify either the discovered targeted pathway or the discovered targeted genes of the novel compound, PTSB, examined in this study.

7 Multiple regression

Given a set of data, x_{jl} , in which $j = 1, \dots, N$ regressor variables (e.g., transcript concentrations), and a dependent variable, y_l (e.g., external influences or transcript concentrations), are measured in $l = 1, \dots, M$ experiments, we desire to calculate the coefficients, a_j , of the following linear regression model:

$$y_l = \sum_j x_{lj} a_j + \eta_l, \quad l = 1, \dots, M, \quad (18)$$

where η_l are stochastic normal variables with zero mean and variance representing the net measurement error in each experiment. Note, η_l includes contributions from measurement noise on the regressor variables and the dependent variable, i.e.,

$$\text{var}(\eta_l) = \sum_j a_j^2 \text{var}(\gamma_{lj}) - \text{var}(\epsilon_l), \quad (19)$$

where ϵ_{il} and γ_{lj} are uncorrelated, zero-mean, normal variables representing measurement noise on the regressor and dependent variables, respectively. We can write this equation in vector form as:

$$\underline{y} = \mathbf{X}\underline{a} + \underline{\eta}, \quad (20)$$

where \underline{y} is an $M \times 1$ vector with elements y_l , \mathbf{X} is an $M \times N$ matrix with elements x_{lj} , \underline{a} is an $N \times 1$ vector with elements a_j , and $\underline{\eta}$ is an $M \times 1$ vector with elements η_l . We desire to find an estimate, $\hat{\underline{a}}$, of the model coefficients that minimizes the L2 norm of the model error:

$$\hat{\underline{a}} = \underset{a_j}{\text{argmin}} \left(\sum_l (y_l - \sum_j x_{lj} a_j)^2 \right). \quad (21)$$

The solution to this minimization problem is:

$$\hat{\underline{a}} = \mathbf{X}^+ \underline{y}, \quad (22)$$

where \mathbf{X}^+ is the pseudo-inverse of \mathbf{X} .

To calculate the error on the estimated model coefficients, we first calculate the covariance matrix, Σ_a . If, in each experiment, the noise, $\underline{\eta}$, is uncorrelated and Gaussian with zero mean and known variance, then [11]:

$$\Sigma_a = \text{cov}(\hat{\underline{a}}) = \mathbf{X}^+ \Sigma_\eta \mathbf{X}^{+T}, \quad (23)$$

where Σ_η is an $M \times M$ diagonal matrix with the elements $\text{var}(\eta_l)$ on the diagonal.

Note that $\hat{\underline{a}}$ calculated using Eq. 21 is not the maximum likelihood estimate of the coefficients \underline{a} when the regressors, \mathbf{X} , are noisy (as is the case here). Nevertheless the pseudo-inverse estimate is reasonable in this situation. Alternately, if we desire the maximum likelihood estimate, we can solve a weighted version of Eq. 21:

$$\hat{\underline{a}} = \underset{a_j}{\text{argmin}} \left(\sum_l \frac{(y_l - \sum_j x_{lj} a_j)^2}{\text{var}(\eta_l)} \right). \quad (24)$$

However, Eq. 22 is not a solution to this equation. The minimizing solution must be estimated using a numerical optimization scheme.

8 Preprocessing of data

For the Hughes et al. [9] data set, each element of the gene expression data matrix x_{jl} has an associated standard deviation σ_{jl} (where j stands for the gene and l for the experiment). Before running the algorithm, the gene expression data are preprocessed as follows: a z-score and the corresponding p -value are computed for each value x_{jl} . If the $p \geq 0.5$ then x_{jl} is set to 0. Missing values in the gene expression data matrix \mathbf{X} are set to 0.

The gene expression data from Mnaimneh et al. [12] lacked associated standard deviations. Therefore, we computed it for each value by a k-nearest neighbor approach. For each value y_{jl} (i.e., the expression of a gene in a given experiment) in the Mnaimneh data set, we selected the eight closest values of the *same gene* across the ‘‘compendium’’ data set. We then took as the standard deviation for the value in the Mnaimneh data set, the mean of the standard deviations in the compendium data set associated to the eight values selected according to the nearest neighbor approach.

- [1] J C Liao, R Boscolo, Y L Yang, L M Tran, C Sabatti, and V P Roychowdhury. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA*, 100:15522–15527, 2003.

- [2] T S Gardner, D di Bernardo, D Lorenz, and J J Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301:102–105, 2003.
- [3] D Thieffry, A M Huerta, E Pérez-Rueda, and J Collado-Vides. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *BioEssays*, 20:433–40, 1998.
- [4] J Tegner, M K Yeung, J Hasty, and J J Collins. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. USA*, 100:5944–5949, 2003.
- [5] H Jeong, S P Mason, A-L Barabási, and Z N Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.
- [6] DC Montgomery, E A Peck, and G G Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, Inc., New York, 2001.
- [7] B S Everitt and G Dunn. *Applied multivariate data analysis*. Arnold, London, 2001.
- [8] Roland Stoughton and Stephen H Friend. Methods for identifying pathways of drug action. *US Patent No. 5,965,352*, 2003.
- [9] Timothy R Hughes, Matthew J Marton, Allan R Jones, Christopher J Roberts, Roland Stoughton, Chirstopher D Armour, et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.
- [10] M J Marton, J L DeRisi, H A Bennett, V R Iyer, M R Meyer, C J Roberts, R Stoughton, J Burchard, D Slade, H Dai, D E Jr Bassett, L H Hartwell, P O Brown, and S H Friend. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.*, 4(11):1293–1301, Nov 1998.
- [11] L Ljung. *System Identification: Theory for the User*. Prentice Hall, Upper Saddle River, NJ, 1999.
- [12] Sanie Mnaimneh, Armaity P. Davierwala, Jennifer Haynes, Jason Moffat, Wen-Tao Peng, Wen Zhang, Xueqi Yang, Jeff Pootoolal, Gardon Chua, Andres Lopez, Miles Trochesset, Darcy Morse, Nevan J. Krogan, Shawna L. Hiley, Zhijian Li, Quaid Morris, Jorg Grigull, Nicholas Mitsakakis, Christopher J. Roberts, Jack F. Greenblatt, Charles Boone, Chris A. Kaiser, Brenda J. Andrews, and Timothy R. Hughes. Exploration of essential gene functions via titratable promoter alleles. *Cell*, 118:31–44, 2004.
- [13] Vlasta Klobucnikova, Peter Kohut, Regina Leber, Sandra Fuchsbichler, Natascha Schweighofer, Friederike Turnowsky, and Ivan Hapala. Terbinafine resistance in a pleiotropic yeast mutant is caused by a single point mutation in the ERG1 gene. *Biochem. Biophys. Res. Commun.*, 309(3):666–671, Sep 2003.

- [14] J Rine, W Hansen, E Hardeman, and R W Davis. Targeted selection of recombinant clones through gene dosage effects. *Proc. Natl. Acad. Sci. USA*, 80(22):6750–6754, Nov 1983.
- [15] G Daum, N D Lees, M Bard, and R Dickson. Biochemistry, cell biology and molecular biology of lipids of *Saccharomyces cerevisiae*. *Yeast*, 14(16):1471–1510, Dec 1998.
- [16] D A Rittberg and J A Wright. Relationships between sensitivity to hydroxyurea and 4-methyl-5-amino-1-formylisoquinoline thiosemicarbazone (MAIO) and ribonucleotide reductase RNR2 mRNA levels in strains of *Saccharomyces cerevisiae*. *Biochem. Cell Biol.*, 67(7):352–357, Jul 1989.
- [17] W Stocklein and W Piepersberg. Binding of cycloheximide to ribosomes from wild-type and mutant strains of *Saccharomyces cerevisiae*. *Antimicrob. Agents Chemother.*, 18(6):863–867, Dec 1980.
- [18] G Barnes, W J Hansen, C L Holcomb, and J Rine. Asparagine-linked glycosylation in *Saccharomyces cerevisiae*: genetic analysis of an early step. *Mol. Cell Biol.*, 4(11):2381–2388, Nov 1984.
- [19] J P Gaughran, M H Lai, D R Kirsch, and S J Silverman. Nikkomycin Z is a specific inhibitor of *Saccharomyces cerevisiae* chitin synthase isozyme Chs3 in vitro and in vivo. *J. Bacteriol.*, 176(18):5857–5860, Sep 1994.
- [20] R M Anderson, M Latorre-Esteves, A R Neves, S Lavu, O Medvedik, C Taylor, K T Howitz, H Santos, and D A Sinclair. Yeast life-span extension by calorie restriction is independent of NAD fluctuation. *Science*, 302:2124–2126, 2003.
- [21] M Ueda, H Kinoshita, T Yoshida, N Kamasawa, M Osumi, and A Tanaka. Effect of catalase-specific inhibitor 3-amino-1,2,4-triazole on yeast peroxisomal catalase in vivo. *FEMS Microbiol. Lett.*, 219:93–98, 2003.
- [22] R Furumai, A Matsuyama, N Kobashi, K-H Lee, M Nishiyama, H Nakajima, Akito Tanaka, Y Komatsu, N Nishino, M Yoshida, and S Horinouchi. FK228 (depsipeptide) as a natural prodrug that inhibits class I histone deacetylases. *Cancer Research*, 62:4916–4921, 2002.

Table S1: TET-inducible experiments: comparison between normal and modified z-scores

TET-inducible allele	Target	rank MNI	rank MNI*	rank LC	rank C	rank logRatio ^a	rank logRatio/ σ ^a
tet-idi1	idi1	1	-	-	-	1	1
tet-rho1	rho1	4	-	-	-	1	1
tet-yef3	yef3	1	1	-	-	4	116
tet-aur1	aur1	1	1	-	-	10	14
tet-fks1	fks1	1	1	89	2	11	41
tet-kar2	kar2	1	67	-	-	78	64
tet-cdc42	cdc42	1	1	278	22	24	141
tet-hmg2	hmg2	1	2	-	-	2	19
tet-pma1	pma1	6	21	-	-	12	22
tet-erg11	erg11	42	42	-	-	2560	2820
tet-cmd1	cmd1	1	1	-	-	1	1

“*” indicates MNI without modified z-score

LC: linear combination, C: correlation

^a ranking by the log₁₀-normalized mRNA expression change upon perturbation (rank logRatio) and the z-score of expression change (rank logRatio/ σ). The sign of the change (up- vs. down-regulation) is ignored.

“-” indicates gene not ranked in top 50

Table S2: Treatment with drugs: comparison between normal and modified z-scores

Drug	Known Targets	rank MNI	rank MNI*	rank LC	rank C	rank logRatio ^a	rank logRatio/ σ ^a
Terbinafine [13]	ERG1	5	5	25	58	224	175
Lovastatin [14]	HMG2	30	30	415	2	52	78
	HMG1	98	98	1	33	162	1301
Itraconazole [15]	ERG11	2	2	3	1	61	192
Hydroxyurea [16]	RNR2	6	-	2 (RNR1)	1 (RNR1)	4	5
	RNR4	2	48	-	-	2	1
Cycloheximide [17]	RPL26b (ribosome)	32	-	2 (RPL6b)	2 (RPL8a)	70	22
	RPS29a	34	-	-	-	200	31
Tunicamycin [18]	ALG7	-	-	188	210	4749	4743
Nikkomycin [19]	CHS3	-	-	-	-	3569	2736
Drugs not in the original "Compendium" data set							
3-aminotriazole [20, 21]	CTA1	-	-	-	-	554	2276
	HIS3	-	-	-	-	197	1261
Dyclonine [9]	ERG2	4	-	23	3	81	6

“*” indicates MNI without modified z-score

LC: linear combination, C: correlation

^a ranking by the log₁₀-normalized mRNA expression change upon perturbation (rank logRatio) and the z-score of expression change (rank logRatio/ σ). The sign of the change (up- vs. down-regulation) is ignored.

“-” indicates gene not ranked in top 50

Table S3: Pathways involved in compound mode of action: MNI approach versus mRNA expression change

Drug	Significant GO ontology (MNI)	Known pathway	Significant GO ontology (RNA change)
Terbinafine	steroid metabolism ; lipid transport	ergosterol biosynthesis [13]	steroid metabolism ; sexual reproduction
Lovastatin	lipid metabolism	ergosterol biosynthesis [14]	lipid biosynthesis
Itraconazole	steroid metabolism ; sterol transport	ergosterol biosynthesis [15]	amino acid biosynthesis; steroid biosynthesis
Hydroxyurea	chromosome organization and biogenesis; DNA replication	DNA replication [16]	chromosome organization and biogenesis; leucine metabolism; pyridoxine metabolism; DNA replication
Cycloheximide	nuclear mRNA splicing, via spliceosome	protein biosynthesis [17]	cytoplasm organization and biogenesis
Tunicamycin	protein-ER targeting ; secretory pathway	N-linked glycosylation [18]	cell wall organization and biogenesis; sexual reproduction; aldehyde metabolism
Nikkomycin	protein amino acid alkylation; growth	cell wall chitin biosynthesis [19]	chromosome organization and biogenesis; protein amino acid alkylation
Drugs not in the original compendium data set			
3-aminotriazole	organic acid metabolism ; vitamin metabolism	organic acid metabolism[20]; oxygen and reactive oxygen species metabolism [21]	chromosome organization and biogenesis; polysaccharide metabolism
Dyclonine	sterol biosynthesis ; amino acid biosynthesis	ergosterol biosynthesis [9]	sterol biosynthesis ; amino acid biosynthesis
Drugs with unknown modes of action			
2-deoxy-D-glucose	amino acid metabolism	–	ubiquitin cycle
Calcofluor White	chromosome organization and biogenesis; transcription initiation	–	chromosome organization and biogenesis; protein amino acid methylation
Doxycycline	ergosterol biosynthesis; siderochrome transport	–	siderochrome transport; response to stimulus; cell redox homeostasis
FR901228 (FK228)	NAD biosynthesis; conjugation; glycogen biosynthesis	histone deacetylation; chromatin silencing[22]	conjugation; vitamin biosynthesis
Glucosamine	chromosome organization and biogenesis; macromolecule biosynthesis	–	chromosome organization and biogenesis; glutamine family amino acid metabolism
MMS	sterol biosynthesis; alcohol metabolism	DNA repair [9]	alcohol metabolism; sterol biosynthesis

bolded text indicates matches with previously reported pathways targeted by each compound.

“–” indicates that the target pathway is not known

Table S4: Pathways involved in compound mode of action: Association analysis approaches

Drug	Significant GO ontology (LC)	Known pathway	Significant GO ontology (C)
Terbinafine	sterol metabolism	ergosterol biosynthesis [13]	ergosterol biosynthesis
Lovastatin	ergosterol metabolism	ergosterol biosynthesis [14]	ergosterol metabolism
Itraconazole	ergosterol metabolism	ergosterol biosynthesis [15]	ergosterol biosynthesis
Hydroxyurea	cellular metabolism	DNA replication [16]	cell cycle
Cycloheximide	protein polyubiquitination	protein biosynthesis [17]	processing of 20S pre-rRNA
Tunicamycin	cell wall organization and biogenesis	N-linked glycosylation [18]	physiological process
Nikkomycin	small GTPase mediated signal transduction	cell wall chitin biosynthesis [19]	conjugation
Drugs not in the original compendium data set			
3-aminotriazole	replicative cell aging	oxygen and reactive oxygen species metabolism [21]	ergosterol metabolism
Dyclonine	ergosterol biosynthesis	ergosterol biosynthesis [9]	regulation of transcription
Drugs with unknown modes of action			
2-deoxy-D-glucose	protein modification	–	loss of chromatin silencing during replicative cell aging
Calcofluor White	regulation of transcription	–	mitotic cell cycle
Doxycycline	mitochondrial inter-membrane space protein import	–	protein transport
FR901228 (FK228)	G-protein signaling, adenylate cyclase activating pathway	histone deacetylation; chromatin silencing[22]	replicative cell aging
Glucosamine	protein-ER targeting	–	reproduction
MMS	cell wall organization and biogenesis	DNA repair [9]	mitosis

bolded text indicates matches with previously reported pathways targeted by each compound.

LC: linear combination, C: correlation

“–” indicates that the target pathway is not known