Molecular Systems Biology 5; Article number 310; doi:10.1038/msb.2009.69
Citation: ***Molecular Systems Biology*** **5:310**

www.molecularsystemsbiology.com

molecular systems biology

Harnessing gene expression to identify the genetic basis of drug resistance

Bo-Juen Chen[1,2,3], Helen C Causton[1], Denesy Mancenido[1], Noel L Goddard[4], Ethan O Perlstein[5] and Dana Pe'er[1,3,*]

[1] Department of Biological Sciences, Columbia University, New York, NY, USA, [2] Department of Biomedical Informatics, Columbia University, New York, NY, USA, [3] Center for Computational Biology and Bioinformatics, Columbia University, New York, NY, USA, [4] Department of Physics and Astronomy, Hunter College, 695 Park Avenue, 1225 Hunter North, New York, NY, USA and [5] Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA
* Corresponding author. Department of Biological Sciences, Columbia University, 2960 Broadway, 607D Fairchild Center, MC 2461, New York, NY 10027, USA. Tel.: +1 212 854 4397; Fax: +1 212 865 8246; E-mail: dpeer@biology.columbia.edu

Received 20.3.09; accepted 24.8.09

The advent of cost-effective genotyping and sequencing methods have recently made it possible to ask questions that address the genetic basis of phenotypic diversity and how natural variants interact with the environment. We developed Camelot (CAusal Modelling with Expression Linkage for cOmplex Traits), a statistical method that integrates genotype, gene expression and phenotype data to automatically build models that both predict complex quantitative phenotypes and identify genes that actively influence these traits. Camelot integrates genotype and gene expression data, both generated under a reference condition, to predict the response to entirely different conditions. We systematically applied our algorithm to data generated from a collection of yeast segregants, using genotype and gene expression data generated under drug-free conditions to predict the response to 94 drugs and experimentally confirmed 14 novel gene–drug interactions. Our approach is robust, applicable to other phenotypes and species, and has potential for applications in personalized medicine, for example, in predicting how an individual will respond to a previously unseen drug.
Molecular Systems Biology **5**:310; published online 13 October 2009; doi:10.1038/msb.2009.69
Subject Categories: functional genomics
Keywords: complex trait analysis; drug target/off-target discovery; genetical genomics

Introduction

Understanding how differences in genotype account for the wide range of phenotypic diversity between individuals is one of the fundamental challenges of biology. With the advent of high-throughput sequencing, the number of available genotypes is increasing at a staggering rate, and we are nearing the point where DNA sequence represents individuals rather than organisms, providing a toehold towards answering this question. Most traits are determined by multiple genes whose identities are largely unknown; therefore, the challenge of predicting an individual's phenome (i.e., spectrum of traits) from its genome requires both identification of the genes that influence the trait, and models that describe how they interact to determine the trait (Gabriel *et al*, 2002; Maller *et al*, 2006).

Our approach is to combine genotype and gene expression data to associate genetic factors with the downstream changes in phenotype. Our premise is that gene expression is useful because it integrates information from multiple loci that are individually too weak to detect but which, in combination, contribute significantly to the phenotype. Gene expression has proven a potent predictor of phenotype, most notably in cancer genomics, where gene expression is used to build classifiers that predict response to therapy (Alizadeh *et al*, 2000; van't Veer *et al*, 2002; Kutalik *et al*, 2008). While relatively accurate, these predictors typically consist of >100 genes and do not provide mechanistic insight regarding the genes responsible for this response. Ground breaking approaches in the genetics of gene expression (Brem *et al*, 2002; Cheung and Spielman, 2002; Dixon *et al*, 2007) have recently been used to show that gene expression can be used to associate genes with disease phenotypes (Mehrabian *et al*, 2005; Schadt *et al*, 2005; Chen *et al*, 2008; Emilsson *et al*, 2008); however, these methods only identify the genes involved and do not directly predict multi-gene traits from the genotype.

We developed Camelot (CAusal Modelling with Expression Linkage for cOmplex Traits) and applied it to genotype, gene expression and phenotype (growth in the presence of drug) data from segregants obtained from a cross between two diverse strains of *Saccharomyces cerevisiae* (Brem and Kruglyak, 2005;

Perlstein *et al*, 2007). The genotypic differences in these strains manifest in rich phenotypic diversity in the segregants. To our knowledge, Camelot is the first method that automatically builds a model based on both gene expression and genotype, selects genes that actively influence the phenotype and accurately predicts complex quantitative phenotypes. Having 'trained' a model, we can use it to accurately predict the growth of a new strain with an entirely different genotype. This is demonstrated by correctly predicting growth, in the presence of each of a panel of drugs, for segregants not used during training. Most importantly, the majority of genes used for predicting growth are causal factors. Thus, genetic manipulation of these genes (deletion or allele swap, that is, replacement of the causal gene with the same gene from the other parental strain) leads to a change in phenotype (e.g., drug resistance/sensitivity) matching our prediction.

An important distinguishing feature of Camelot is that it integrates genotype and gene expression data, generated under drug-free conditions, to detect causal genes and predicts the response to an entirely different condition, growth in the presence of a drug. Therefore, gene expression of an individual need only be assayed once. This single-gene expression profile can be harnessed to analyse the connection between genotype and phenotype for a large number of traits that manifest under many different conditions. Moreover, the response to a drug can be predicted before treatment, a critical feature for clinical application.

Our results demonstrate that Camelot can predict a strain's response to a drug, for 87/94 drugs. The inclusion of gene expression data measured under unrelated (drug-free) conditions significantly contributes to Camelot's accuracy in predicting drug response and in its ability to detect causal genes involved in this response. We experimentally confirmed 25/27 of Camelot's predictions regarding the influence of a specific gene in the response to a specific drug. Our data demonstrate that Camelot is able to identify genes involved in drug resistance robustly.

Results

We used a data set containing information from 104 segregants that arose from the mating of two genetically diverse strains, 'BY' and 'RM' (Brem and Kruglyak, 2005). The data include the growth yield from each segregant grown in the presence of one of 94 chemicals ('drugs') (Perlstein *et al*, 2007), 526 processed markers denoting genotype (Lee *et al*, 2006) and 6189 gene expression profiles, measured in rich media, for each segregant (Brem and Kruglyak, 2005).

The BY and RM strains used in this study are genetically distant, with 0.5% sequence diversity between them. This genetic diversity manifests in significant phenotypic diversity. Not only do the strains differ in their response to drugs; each drug has a different set of fast- and slow-growing segregants (Box 1A).

Gene expression measured in the absence of drug helps predict drug response

Our goal is to obtain baseline information about a strain, genotype and gene expression data measured from each segregant grown in the absence of drug, and use this to derive a quantitative prediction of the strain's phenome, its response to each drug in a panel of drugs. We seek to identify a small set of features, either genotypic markers or single genes (transcripts in the gene expression data) that influence growth in the presence of each drug, and to explain the observed differences between segregants. We use the term 'causal' to describe a feature that not only correlates with and predicts the phenotype, but which actively influences it. We define a feature as 'causal' if genetic manipulation of this feature, for example, by allele swap or gene deletion, changes the phenotype, as predicted by the model.

Identifying a predictive model defines a task of selecting a sparse set of features from a pool of markers and a precompiled list of transcripts that together predict growth in the presence of drug D. Although the true relationship may not be linear, we use linear models as these can be robustly inferred from the data (Hastie *et al*, 2001). Camelot selects a sparse set of features, markers $\{L\}$ and transcripts $\{E\}$ so that $D \sim \{L\} + \{E\}$ (Box 1B).

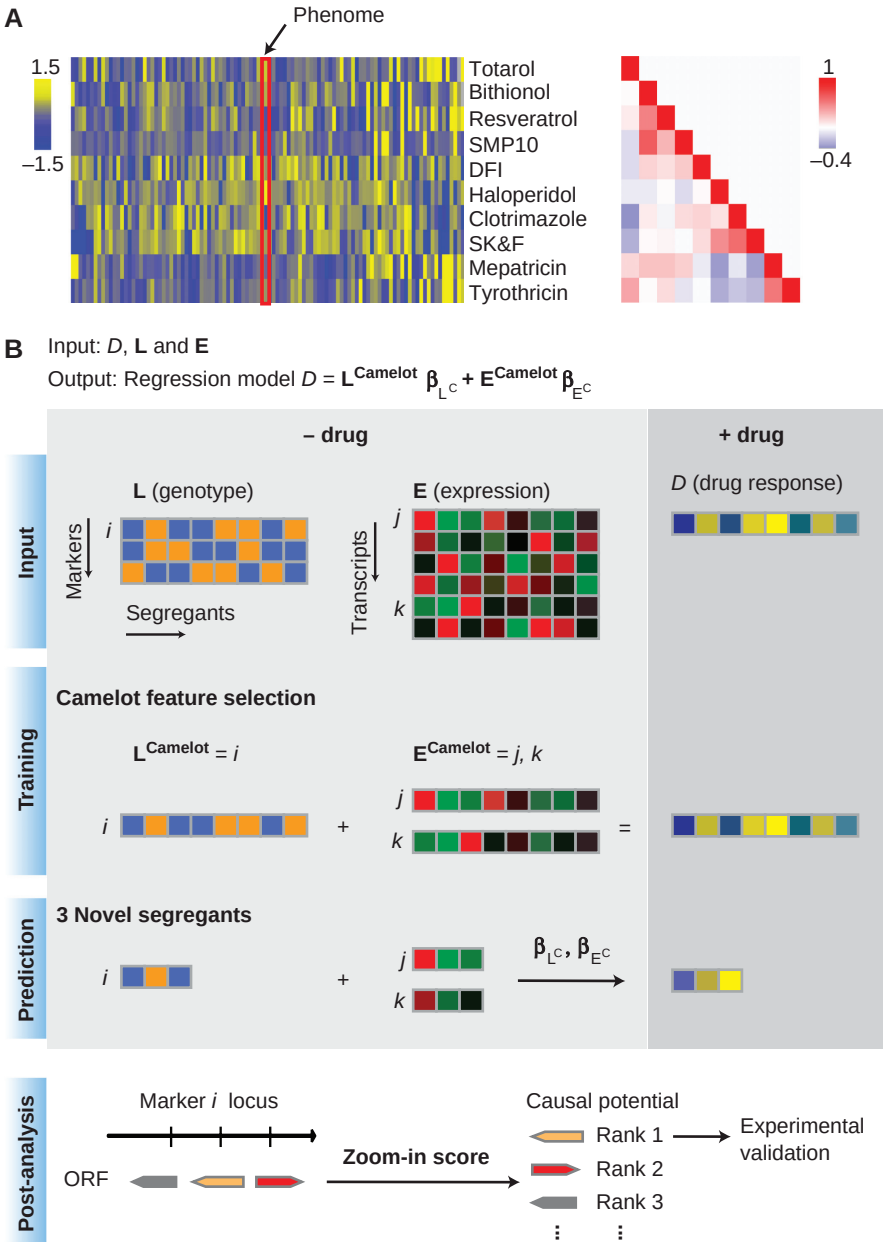
Identifying a small number of predictive features from thousands of candidates is a well studied problem of high dimensional feature selection (Hastie *et al*, 2001). To avoid identifying features that match the training data by chance, our algorithm uses a combination of statistical tools including elastic net regularized regression (Zou and Hastie, 2005), non-parametric bootstrap (Efron, 1979) and tests designed to further select only those that are most likely causal. The selected features are then used to optimise a linear prediction function (see section Materials and methods).

We evaluated the performance of our approach using 10-fold cross-validation; we randomly split the segregants (strains) into training and test sets and completely withheld any data relating to the test strains during model selection. Camelot uses gene expression, genotype and drug response data from strains in the training set to build a model that both predicts growth for each condition (+drug) and identifies the genes responsible for the differences in phenotype between strains. Camelot was subsequently used to predict the drug response for the withheld test strains using only genotype and gene expression data measured under drug-free conditions (see section Materials and methods). These test strains simulate a situation in which Camelot is used to predict the phenotype of new, previously unobserved, strains.

Camelot outperforms association and linkage analysis in providing a set of features that yield significantly more accurate prediction of drug response (see Figure 1A–D). We found that Camelot's predictions for growth in the test strains were more accurate for 88% of the conditions examined, compared with those obtained using standard linkage methods (Figure 1A), and in many cases led to dramatic improvement in the accuracy of prediction, for example, for clomiphene and haloperidol (Figure 1C and D).

While Camelot's statistically rigorous feature selection framework contributes to its success, so does the use of gene expression data, as evidenced when we compare our method with and without the use of expression data (Figure 1B–D). Note that the gene expression data were obtained from cells grown under nutrient-rich, non-perturbed conditions, whereas the growth data were measured in the presence of different

Box 1 Diversity of drug response and the outline of the Camelot algorithm



Box 1 (**A**) Growth in the presence of a subset of drugs is represented by the heat map on the left (blue corresponds to low growth yield and yellow to high growth yield). Each row represents the data for a single drug (SMP10 is 1,9-pyrazoloanthone, DFI is diphenyliodonium and SK&F is SK&F 96365) and each column represents a different strain/segregant. The red rectangle shows the response of a segregant to the set of drugs indicated, known as the 'phenome', of the strain. The heat map on the right represents the correlation between the responses of the segregants to the drugs (Pearson's correlation coefficient). The rows and columns are in the same order as the rows in the heat map on the left. The range in Pearson's correlation coefficient demonstrates that there is considerable diversity in the response of the segregants to these drugs; the correlation ranges from strong positive correlation (r=0.64) to strong anticorrelation (−0.40). The same scale is used for all the figures. (**B**) Overview of Camelot. The input data include matched genotype (L) and gene expression (E) data for each segregant measured under standard conditions (no drug) and growth yield/drug response (*D*) measured in the presence of a drug. Each column represents a strain/segregant and each row represents a marker feature in the genotype matrix or a transcript feature in the gene expression matrix. Camelot outputs a predictive regression model with a small set of markers and gene expression features. In the training phase, Camelot takes genotype, gene expression and drug response as input and uses feature selection methods (elastic net, bootstrap, the triangle test and model revision) to choose a small set of marker and gene expression features that best predict the drug response that are enriched for features likely to have a causal influence on the phenotype. Selected sets of features are denoted by $L^{Camelot}$ and $E^{Camelot}$, representing selected markers and transcripts, respectively. A linear regression model is then built on $L^{Camelot}$ and $E^{Camelot}$. In the prediction stage, Camelot uses the model built on the training data (regression coefficients β_{LC} and β_{EC}) and the genotype and expression data for the held-out segregants to predict growth in the presence of drug. Following model selection, Camelot takes each selected marker ($L^{Camelot}$) and uses the zoom-in score to prioritize the likelihood that each gene within the linked region is causal.

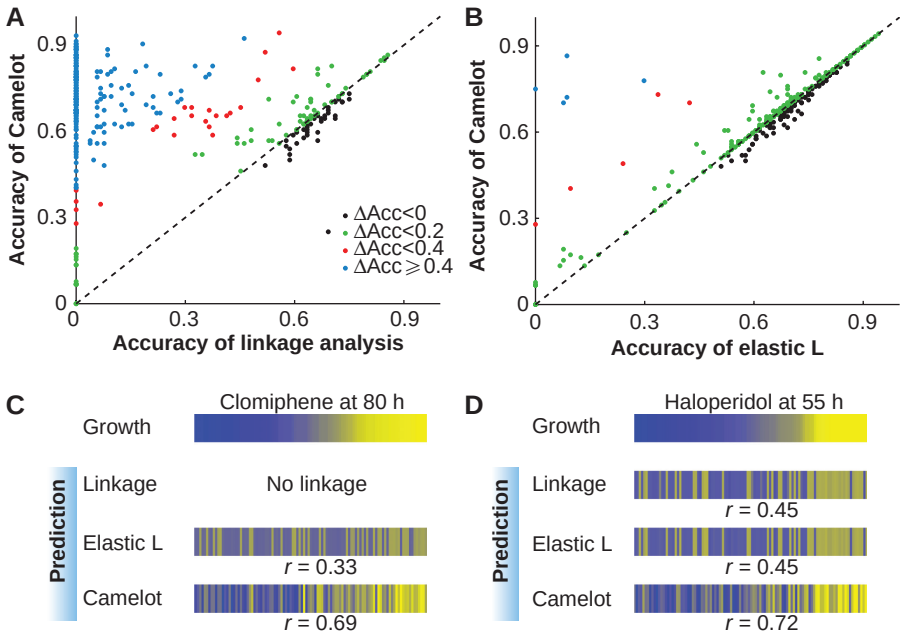


Figure 1 Camelot has superior predictive ability. Comparison of prediction methods on held out test data from different models. (**A**) Classification accuracy (see section Materials and methods): Camelot compared with linkage analysis. Each dot represents a condition (growth yield in the presence of a drug), showing the fraction correctly predicted by Camelot (*y*-axis) and linkage analysis (*x*-axis). Dots above the diagonal indicate the superior performance of Camelot and are colour coded to indicate the degree of improvement. (**B**) As in panel A, but the classification accuracy by Camelot is compared with that of the elastic-net L model lacking transcript features (see section Materials and methods). This demonstrates that for many conditions the inclusion of gene expression features improves Camelot's performance. (**C**) The top bar represents growth in the presence of clomiphene; each column is associated with a different segregant (matched horizontal positions within the panel) sorted by growth from low (blue) to high (yellow). The observed growth is compared with model prediction from linkage analysis, the elastic-net L model and Camelot. The bar marked elastic L represents predictions from bootstrapped elastic net regression using genotype alone, and the bottom bar represents prediction from Camelot. Prediction (on test data) improves from no detected linkage to most accurate for Camelot. The same scale is used for all the figures. (**D**) As in panel C, but for haloperidol.

drugs and that the expression features chosen differed between the drugs. Therefore, the features selected are unlikely to represent genes whose expression merely correlates with rapid growth (Airoldi *et al*, 2009).

The response of segregants to different conditions is heritable (Perlstein *et al*, 2006), so the boost in performance, over genotype alone, gained by using gene expression data (generated in the absence of drug) is counter-intuitive (Figure 1B). A factor that contributes to the accuracy is that transcript features chosen by Camelot typically correlate well with the measured growth yield in the presence of a drug. This success in prediction is similar to the success of gene-expression-based classifiers in predicting response to chemotherapy in cancer genomics (van't Veer *et al*, 2002). However, correlation does not necessarily imply causality.

Identifying features that actively influence the phenotype

Camelot aims to find a model that is not only predictive, but also identifies genes that are responsible for the phenotypic variation. Identification of these genes provides insight into the biological processes and stresses involved in response to a drug, and has practical implications for identifying alternative drug targets in resistant strains.

Care must be taken when attributing a causal interpretation to a correlated feature, even when the feature acts as a potent predictor (Pearl, 2000). When the feature correlated with growth is based on linkage to a DNA marker, the issue of causality is straightforward: the observed phenotype is likely influenced by genetic polymorphism within the linked region. However, when the feature is based on correlation between the abundance of a transcript and the phenotype, three possibilities exist: (1) the transcript and phenotype correlate due to a common cause resulting from DNA variation (Figure 2C), (2) DNA variation exerts its effect on the phenotype through the gene, and hence the expression level serves as an indicator of the causal effect of the genetic differences on the phenotype (Figure 2D) or (3) growth rate influences the abundance of the transcript. The last option is not considered in this experimental design, as gene expression was measured in the absence of drugs.

For example, there are 123 genes whose expression correlates with growth in hydrogen peroxide with an absolute coefficient of 0.35 or greater. Of these, Camelot only chose one transcript feature, *DHH1* (Figure 3A). We explain how Camelot goes beyond correlation to identify the most likely causal gene. First, Camelot limits the set of possible candidate transcript features to 854 transcripts that are not particular to any specific drug, yet are *a priori* more likely to be causal based on the functional classification of their cognate genes (see section Materials and methods). Camelot selects features using a bootstrap procedure on coefficients of regularized regression (see section Materials and methods). Systematic evaluation using synthetic data shows that bootstrapping of regression

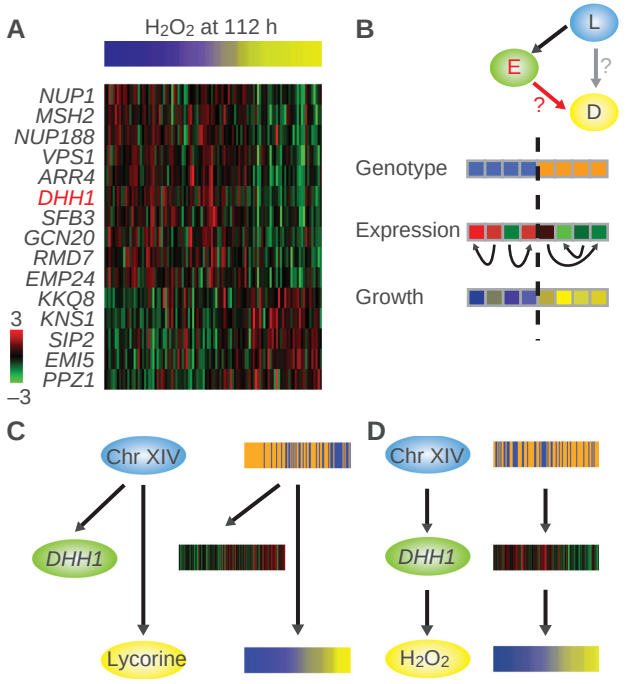


Figure 2 Correlation versus causality. (**A**) Growth yield in the presence of hydrogen peroxide and correlated expression profiles for genes in the candidate pool (absolute Pearson correlation coefficients $\geqslant 0.36$, $P < 2 \times 10^{-4}$), showing that the expression of multiple genes correlates with growth. Each column is associated with a different segregant (matched horizontal position across panel) sorted by growth yield (as in Figure 1) and gene expression on a red–green scale. (**B**) The triangle test evaluates the likelihood that each transcript feature causally explains the phenotype (red edge). It distinguishes between causal chain (left, red edge) and co-regulation structures (right, grey edge) using permutation testing to evaluate the contribution of gene expression controlled for the genotype of L. That is, expression is permuted under the allele of the linked genotype (see section Materials and methods). Orange represents RM and blue BY for the genotype at the locus. Some notation for all the figures: Blue ovals represent the genotype of a marker; yellow ovals, drug response; green ovals, gene expression; red arrows, the driving edge; black arrows, causal relationships and grey arrows, relationships tested in the test. Red letters indicate the type of the selected feature (expression in this case). (**C**) Sequence variation at Chromosome XIV locus is the cause of variation in both *DHH1* expression and the response to lycorine; however, in this situation *DHH1* expression does not causally influence the drug response. *DHH1* was a candidate feature for lycorine (correlation coefficient, r=0.44), but failed the triangle test, showing that high correlation does not necessarily reflect causality. The heat maps show segregants ordered based on response to lycorine. (**D**) Example of a causal chain where polymorphisms at a Chromosome XIV locus lead to change in *DHH1* expression that results in differences in cell growth in the presence of H_2O_2. *DHH1* was chosen as a candidate feature for H_2O_2 (r=−0.44) and passed the triangle test with P-value 1.6×10^{-5}.

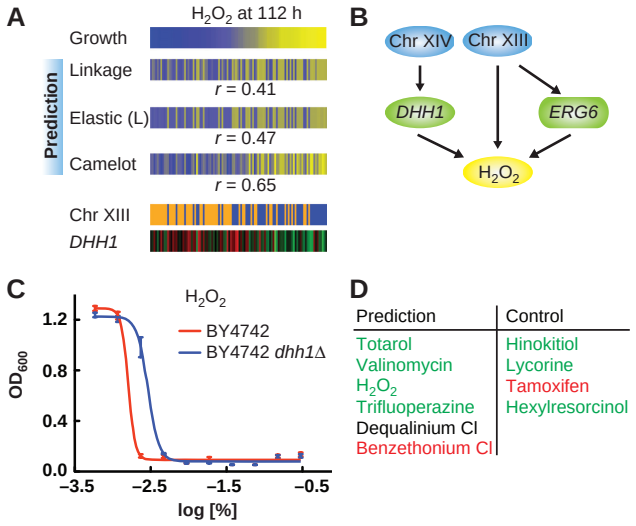


Figure 3 Causal role of *DHH1*. (**A**) Growth yield in the presence of H_2O_2 compared with model prediction from linkage analysis, elastic-net L model and Camelot, represented as in Figure 1C, demonstrating superior prediction by Camelot. Camelot chose a Chromosome XIII locus (227 254–243 624) and expression of *DHH1* as features to predict the drug response; the values for each segregant are represented in the same order within the panel. (**B**) The full prediction function obtained from Camelot for response to H_2O_2. *DHH1* is selected as a feature and confirmed by the triangle test; the Chromosome XIII marker is selected as a feature and the zoom-in score identifies *ERG6* as the causal gene within the region, fitting with reports that overexpression of *ERG6* leads to decreased resistance to hydrogen peroxide (Khoury *et al*, 2008). The Chromosome XIV locus is at position 449 639–486 861. Some notation for all the figures: Green rectangles (such as *ERG6*) represent expression of a gene within a linked region. (**C**) Averaged OD_{600} absorbance growth measurements of BY (red) and BY *dhh1*Δ mutant (blue) plotted against twofold dilution series of H_2O_2. The error bars represent the standard error of the mean for all growth yield data. These data confirm the causal effect of *DHH1*. (**D**) *DHH1* is a hub passing the triangle test for six drugs (left column). Five of these were tested; validated causal effects are in green, with one false positive listed in red. To assess the drug specificity of *DHH1*-mediated effects, four negative controls were tested (right column); confirmed negative predictions are listed in green and one false negative in red. See Supplementary Figure 2 for drug response curves for each of the drugs tested, as represented in Figure 3C.

dramatically increases Camelot's precision in correctly pinpointing the causal features that generate the phenotype, both compared with elastic net regression alone and the feature selection methods used by Schadt *et al* (2005) and Chen *et al* (2008) (Supplementary Figure 1). For the hydrogen peroxide response, GO-based filtering reduced the list of 123 candidate transcripts to the 15 genes shown in Figure 2A. Bootstrapping further reduced the list of expression features to a single gene, *DHH1*, that was subsequently experimentally validated (Figure 3C).

In the next stage, Camelot explicitly tests for causality. We apply a causality test to all transcript features chosen with significant confidence after bootstrapping. The permutation-based triangle test asks, 'Is gene expression significantly predictive of the growth beyond the contribution of the linked genotype?' (Figure 2B and section Materials and methods). We assume that the linked DNA marker is causative *a priori* and require that the transcript feature remains significantly predictive of growth even after the influence of the marker is controlled for. While this test does not guarantee that the transcript feature is indeed causal, it identifies transcript features that are more likely causal and enriches the final selection with causal features. For example, the abundance of the *DHH1* transcript was selected by our bootstrap procedure as a feature that predicted the response to 10 different drugs. After administering the triangle test, *DHH1* passed as causal for only six of these drugs. These were subsequently validated experimentally (Figure 3D). The variability in *DHH1* expression across segregants arises because of polymorphism in *MKT1* (chromosome XIV) (Lee *et al*, 2009), although it is likely that other genetic factors also affect *DHH1* expression. We believe that *DHH1* expression is influenced by multiple genetic factors, that are individually too weak to detect, and that this

explains why gene expression is so potent in improving prediction accuracy.

From prediction to mechanism

The true value of gene expression comes to light when one focuses not on how resistant a strain is, but rather why it is so. Rather than being a black box predictor, transcript features can help shed light on the mechanisms underlying resistance. *DHH1* was chosen as a feature for a large number of drugs, so we tested Camelot's prediction that *DHH1* plays a causal role in mediating resistance to these drugs. *DHH1* expression is negatively correlated with growth in the presence of hydrogen peroxide (correlation coefficient $r=-0.44$), and we tested the prediction that *DHH1* influences drug response by measuring the growth yield of wild-type and *dhh1*Δ strains in hydrogen peroxide (Figure 3C and section Materials and methods). The *dhh1*Δ strain grew better than the wild type, confirming that *DHH1* negatively influences the phenotype.

This result complements the finding that Dhh1 colocalizes with the sequence-specific RNA-binding protein Puf3 and regulates the abundance of 153 Puf3-bound mRNAs (Lee *et al*, 2009). Puf3 is a factor that binds select nuclear-encoded genes involved in mitochondrial biogenesis and likely regulates the transport/translation/stability of these messages (Garcia-Rodriguez *et al*, 2007; Saint-Georges *et al*, 2008). These Puf3-bound, mitochondrial-related genes are significantly upregulated in *dhh1*Δ strains (Lee *et al*, 2009). As *DHH1* is expressed at a higher level in the BY parent, this strain might have a lower capacity for detoxification of the reactive oxygen species produced on hydrogen-peroxide treatment and a lesser ability to withstand this insult. Genes annotated for mitochondria are upregulated in the RM strain (Litvin *et al*, 2009) and this strain is predisposed towards respiratory growth (Smith and Kruglyak, 2008).

Testing the causal role of transcript features

We confirmed our predictions for the influence of *DHH1* on growth in totarol, valinomycin, hydrogen peroxide and trifluoperazine. Benzethoniumchloride was the only false positive among the drugs tested (Figure 3D and Supplementary Figure 2). We included lycorine, hinokitiol, hexylresorcinol and tamoxifen as negative controls to demonstrate that *DHH1* activity is drug specific. Only growth in tamoxifen was influenced by *DHH1*; indeed tamoxifen perturbs mitochondrial function (Tuquet *et al*, 2000; Cardoso *et al*, 2001). This demonstrates the stringency of our approach, which is designed to minimize false positives and does not detect all genes that influence drug responsiveness or all drugs influenced by a gene. In summary, we confirmed 4/5 of the positive predictions tested and 3/4 of the negative predictions for *DHH1*, demonstrating the drug specificity of our predictions. Although the drugs linked to *DHH1* are diverse and include an antibiotic (valinomycin) and an antipsychotic drug (trifluoperazine), they all affect mitochondrial function (Nicolson *et al*, 1999; Evans *et al*, 2000; Nulton-Persson and Szweda, 2001; Lee *et al*, 2005; Safiulina *et al*, 2006; Yip *et al*, 2006; Sancho *et al*, 2007; Lee *et al*, 2008). This suggests a possible application of Camelot in predicting the mechanism of action of novel drugs.

MGA2, a gene whose product is involved in fatty-acid metabolism (Chellappa *et al*, 2001; Jiang *et al*, 2002; Kandasamy *et al*, 2004), was identified as another transcript feature predictive of growth for six drugs. Three of these (cerulenin, ikarugamycin and tomatine) act by perturbing processes involved in fatty-acid and lipid synthesis and membrane permeability (Vance *et al*, 1972; Hasumi *et al*, 1992; Friedman, 2002). Unsaturated fatty acids (FA) are essential components of membranes and FA synthesis is effected by controlling the stability of *OLE1* mRNA. Ole1 is required for the formation of monounsaturated FA precursors (Martin *et al*, 2007). Mga2 acts to stabilize or destabilize the *OLE1* message depending on the conditions (Kandasamy *et al*, 2004). The gene expression data show that in the non-perturbed state *MGA2* expression is negatively correlated with *OLE1* expression ($r=-0.54$) and positively correlated with the drug response.

These examples illustrate the power of Camelot to identify genes that causally influence the response to multiple drugs, predict the mechanism of action of drugs and provide insight into the underlying biology.

Using gene expression to identify causal genes within a linked region

Transcript features relate to a single gene and hence directly identify the involved gene. DNA marker features are better founded in their causal nature, but typically involve large chromosomal regions containing tens of genes. For these features, Camelot uses gene expression to help pinpoint the causal gene within the linked locus. The zoom-in score uses gene expression to prioritize the likelihood that each gene within the linked region is causal. Like the triangle test, the zoom-in score is a measure of how well gene expression predicts the phenotype. Linkage implies that the marker is driving the causality; therefore, the zoom-in score includes an additional measure for *cis*-linkage, how well the marker predicts the gene expression. The zoom-in score incorporates both of these qualities, as well as conservation of the protein sequence to prioritize genes within a locus (see Figure 4A and section Materials and methods).

Camelot chose two features for hydrogen peroxide, the *DHH1* transcript and a region on chromosome XIII (locus 227 254–243 624) containing 88 genes (Figure 3B). The zoom-in score identified *ERG6* as the causal gene within this region, that is, polymorphism in the *ERG6* sequence between the BY and RM strains is likely responsible for the differences in the response to hydrogen peroxide between the parent strains. Overexpression of *ERG6* leads to decreased resistance to hydrogen peroxide (Khoury *et al*, 2008), matching Camelot's prediction (Figure 3B). These results demonstrate how the triangle test and zoom-in score combine to provide a better understanding of the cellular response to each drug.

Similar to linkage analysis (Perlstein *et al*, 2007), Camelot identified the two largest marker hotspots, a region on chromosome XIII (locus 27 644–33 681), linked to 25 drugs, and a region on chromosome XIV (linked to 12 drugs). While

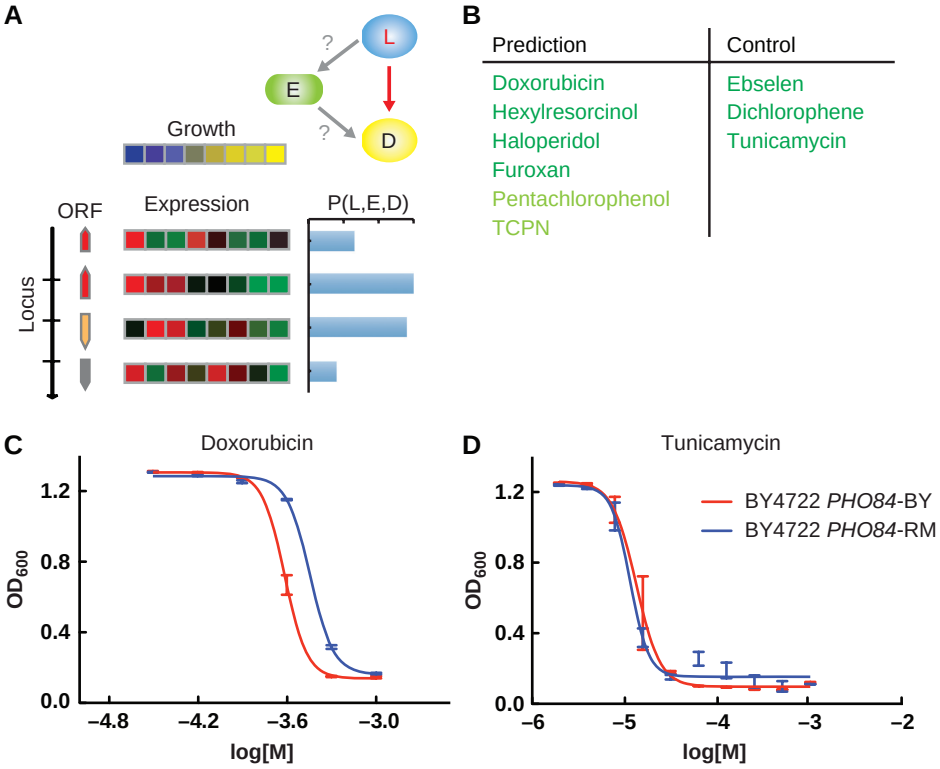


Figure 4 Causal role of *PHO84*. (**A**) The zoom-in score for each locus to drug association (red arrow) evaluates the likelihood, *P*(*L*, *E*, *D*), that each gene in the locus causally influences the growth of the strain in response to drug, based on its expression (each row represents a gene at the locus and each column a segregant; see section Materials and methods for details). (**B**) Validation of the ability of the zoom-in score to distinguish which drugs are influenced by *PHO84*. To the left are high-scoring drugs predicted to be influenced by *PHO84*; all six drugs were validated including two previously associated with *PHO84* (Perlstein *et al*, 2007) (light green) and four new drugs associated with *PHO84* (dark green). To the right are low-scoring drugs, not expected to be influenced by *PHO84*; *PHO84* had no effect on the response for any of the three, validating the ability of the zoom-in score to make positive and negative predictions. (**C**) Averaged OD_{600} absorbance growth measurements of BY (red) and BY with an allele swap for *PHO84*-RM (blue) plotted against concentration of doxorubicin. See Supplementary Figure 3 for the response to each drug represented as in panel C. (**D**) As panel C, but for tunicamycin, showing that variation in the DNA sequence of *PHO84* has little effect.

linkage alone only detects large multi-gene loci in this data set, the zoom-in score further identified *PHO84* (chromosome XIII), as the top-ranked causal variant for multiple drugs. Two of these drugs, tetrachloroisophthalonitrile and pentachlorophenol, were manually identified and verified previously (Perlstein *et al*, 2007). *PHO84* was top scored for a number of additional drugs linking to the chromosome XIII hotspot, but scored poorly for other drug phenotypes linking to this hotspot. We used the zoom-in score to distinguish which drugs are causally influenced by *PHO84* and validated these predictions by growing wild-type BY and allele-swapped (AS) strains (BY strain containing *PHO84* with one amino-acid substitution, L259P, from the RM strain) individually in the presence of one of nine drugs. We included drugs with both positive and negative predictions. Camelot correctly predicted both positive and negative responses 9/9 times, demonstrating that the zoom-in score can be used to identify which of the Chromosome XIII-linked drug phenotypes are causally influenced by the *PHO84* allele (Figure 4B–D and Supplementary Figure 3). We performed a similar analysis for the drugs linking to the Chromosome XIV region and identified three drugs likely to respond to *MKT1* and three linked drugs that are unlikely to be affected by *MKT1*. Again Camelot correctly predicted the response to the drugs in an AS (BY *MKT1*-RM) strain 6/6 times (Supplementary Figure 4). These data validate our approach and demonstrate that Camelot is also able to capture factors accounting for phenotypic variation, using markers as features, for a number of causal genes.

GPB2 a new causal gene for multiple drugs

Both *PHO84* and *MKT1* have previously been shown to influence phenotypic differences between BY and RM, although Camelot successfully linked four new phenotypes (response to drug) to *PHO84* and three new phenotypes to *MKT1*. To further test Camelot, we assessed whether it could identify new genes, not previously implicated in the differences between BY and RM. One of the strongest signals from our zoom-in analysis comes from the locus of Chromosome I: 1–55 329, which links to growth under a number of drugs including haloperidol, E6 berbamine and gliotoxin. Segregants bearing the RM allele are highly sensitive to these drugs. *GPB2* is consistently the top-scored gene at this locus for all these drugs. Sequence alignment showed that *GPB2* differs by 10 amino-acid substitutions between BY and RM and that one of them is highly conserved across fungal species (P269L, BY-*GPB2* encodes proline and RM-*GPB2* encodes leucine). We engineered an AS strain (BY *GPB2*-RM) in which the entire

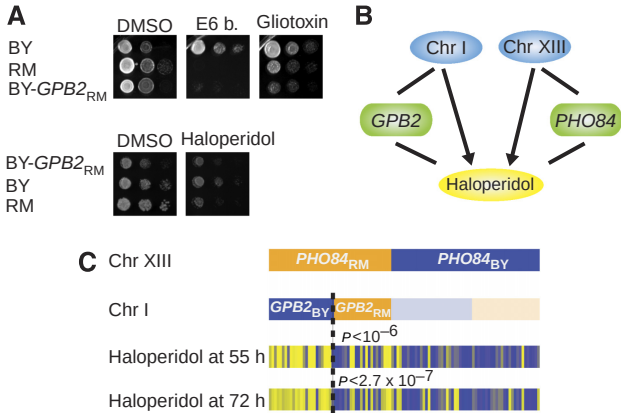


Figure 5 Causal role of *GPB2* in response to drugs. (**A**) Strains were grown overnight in YPD medium, diluted to $OD_{600} \sim 0.2$ and plated with 10-fold dilution on YPD + drug media (see section Materials and methods). The top three panels are photos of YPD plates containing DMSO (control), E6 berbamine or gliotoxin. The bottom panels are photos of YPD plates containing DMSO or haloperidol. The results show a large difference in drug sensitivity between BY and RM. The AS strain (BY *GPB2*-RM) grows at a rate similar to the RM strain. (**B**) Camelot identifies two loci (Chromosome I: 1–55 329 and Chromosome XIII: 27 644–33 681) and causal genes encoded within these loci, *GPB2* and *PHO84*, that are responsible for the response to haloperidol. (**C**) Analysis shows that *GPB2* and *PHO84* interact with each other to influence growth in the presence of haloperidol. Shown are the genotypes for *PHO84* and *GPB2*, and growth in the presence of haloperidol. Segregants with both the *PHO84*-RM and *GPB2*-BY alleles have significantly better resistance (*P*-value from Wilcoxon rank-sum test) to haloperidol compared with other segregants.

BY *GPB2* coding region was replaced with that from the RM strain (see section Materials and methods) and experimentally validated Camelot's prediction that *GPB2* plays a causal role in response to these drugs by showing that the BY *GPB2*-RM strain is more sensitive to the presence of E6 berbamine, gliotoxin and haloperidol than the BY strain. Indeed, the AS strain is highly similar to the RM strain on E6 berbamine and haloperidol, suggesting that variation in the *GPB2* sequence accounts for much of the difference in the response to these drugs (Figure 5A).

Gpb2 is an effector of Gα protein Gpa2 and inhibits PKA downstream of Gpa2, which increases dependence on cAMP (Harashima *et al*, 2006; Peeters *et al*, 2006). Both gliotoxin and haloperidol affect the cAMP/PKA pathway. Gliotoxin is a fungicide that increases cAMP/PKA activity (Waring *et al*, 1997), whereas haloperidol, a clinical antidepressant (dopamine D2 receptor antagonist), increases cAMP/PKA activity in striatum (Kaneko *et al*, 1992; Turalba *et al*, 2004). These results support our finding that polymorphism in *GPB2* has an effect on the response to these drugs, and suggest that the mechanism of action involves G-protein signalling. Our findings suggest that E6 berbamine, whose pharmacological effect remains unknown, may also have similar effect on the cAMP/PKA pathway.

The response to haloperidol is highly variable among segregants. Although the mechanism of action could not be established based on linkage to a large region alone, Camelot provided clues by zooming in on *GPB2* and *PHO84*. These genes were subsequently validated as causal for the drug response phenotype (Figure 5B). We assessed the combined influence of both genes for growth under haloperidol statistically, using data from the segregants. Strains carrying both RM-*PHO84* and BY-*GPB2* grow better than strains with other combinations of alleles (Figure 5C), indicating that *PHO84* and *GPB2* may function through a common pathway. The involvement of Pho84 as a sensor and signalling molecule for phosphate-based activation of PKA (Giots *et al*, 2003) further implicates PKA function in the response to haloperidol.

PHO84 gene expression and feedback

In total, we validated 18/18 predictions made using the zoom-in score, including 9/9 for *PHO84*. Although *PHO84* has two SNPs between BY and RM in its coding region, there is no genetic variation in regulatory regions such as the promoter or 3′UTR. Moreover, the AS strain, containing only one amino-acid substitution (L259P) in the coding region of *PHO84* in the BY background, recapitulated the growth rate of RM for many of the positive drugs tested. So it is surprising that expression of *PHO84*, generated in the absence of any drugs, could accurately distinguish between drugs that are affected by *PHO84* and those that are not. To better understand why this information is encoded in the expression data, we carried out RT–PCR using strains grown in YPD media (no drug) to monitor *PHO84* expression in the BY, RM and the AS (BY *PHO84-RM*) strains.

Although the AS strain contains BY *cis*- and *trans*-regulatory factors, the presence of the RM coding region alone (with one amino-acid substitution L259P) brought the expression of *PHO84* in the AS strain down to that of the RM strain (Figure 6A). The difference in expression results from negative feedback that acts through the Pho84 protein under high-phosphate conditions (Figure 6B; Wykoff *et al*, 2007). To quantify the degree of negative feedback between strains, we used RT–PCR to measure *PHO84* expression under both low and high-phosphate conditions. As expected, *PHO84* expression is significantly downregulated under high-phosphate, relative to low phosphate, conditions in all three strains (Figure 6C). Nevertheless, the negative feedback is stronger in the RM and AS strains (817- and 170-fold change, respectively) relative to the BY strain that only goes down 11-fold. Under low-phosphate conditions, the gene expression for all three strains is similar, suggesting that the loop is not active (Supplementary Figure 8). This implies that the relative strength of the negative feedback differs between strains under the high-phosphate conditions that activate this loop.

We used arsenate, a toxic non-metabolizable phosphate analogue, as an indicator of the relative affinity of Pho84 for phosphate. The RM and AS strains are significantly more sensitive to arsenate, suggesting that the RM version of *PHO84* is a more efficient transporter of phosphate than the BY strain (Supplementary Figure 9). This effect is mediated by Pho84 as addition of methylphosphonate (an inhibitor of Pho84) reverses this phenotype (data not shown). It is likely that the differences in Pho84 function between BY and RM are responsible for the observed differences in drug sensitivity. Variation in gene expression serves as an indicator of the variation in protein function, which acts through a feedback mechanism; the expression level itself is unlikely to be causal directly.

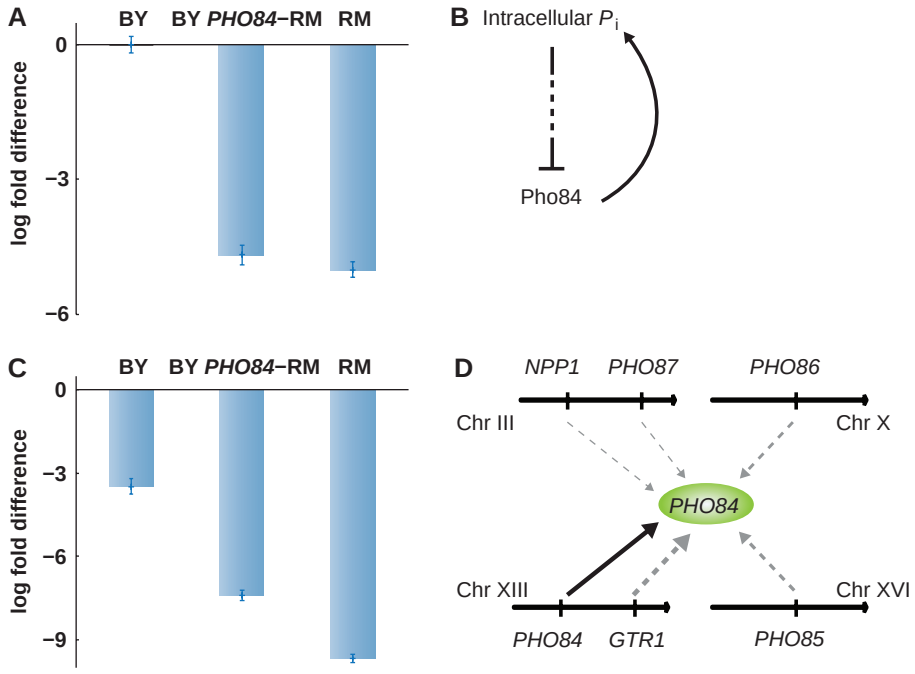


Figure 6 Feedback in *PHO84* expression. (**A**) Expression levels of *PHO84* in the BY, RM and the *PHO84*-RM AS strains. The expression of *PHO84* in the AS and RM strains is similar and significantly lower than in BY. The fold difference is calculated relative to the BY strain. Since the AS strain only differs from the BY strain in the *PHO84* coding sequence, feedback regulation must act through the *PHO84* gene itself. The error bars represent the s.d. of three replicates. All RT–PCR experiments were conducted independently at least three times. (**B**) A negative feedback loop regulates expression of Pho84 in response to the concentration of intracellular phosphate. When phosphate levels are low, Pho84 is expressed and transports inorganic phosphate into the cell; Pho84 is repressed as intracellular phosphate levels rise. (**C**) Expression of *PHO84* in SC + high-phosphate media compared with that in SC + low-phosphate media for each strain. The cells were harvested 90 min after the addition of phosphate. The AS and RM strains are repressed to a greater extent than BY in response to the addition of phosphate. (**D**) Weak eQTL that influence the expression of *PHO84*. These loci are enriched in genes involved in phosphate metabolism and phosphate transport. *PHO84* expression links to regions that contain *GTR1*, *NPP1*, *PHO84*, *PHO85*, *PHO86* and *PHO87*. The width of arrows corresponds to the significance of linkage (*P*-value for each linkage <0.01; see section Materials and methods).

The only region that links to *PHO84* expression is its own. *PHO84* has strong *cis*-linkage with *P*-value 6.3×10^{-5}. We therefore asked why *PHO84* expression might provide information beyond that of the presence of the *PHO84* allele. Removing the genome-wide correction for multiple testing in eQTL, we detect additional regions, each with very weak linkage (Supplementary Table V). These regions contain *GTR1*, *NPP1*, *PHO85*, *PHO86* and *PHO87*, each involved in phosphate metabolism/transport, that contain multiple non-synonymous coding SNPs. This suggests that many genes associated with phosphate metabolism/transport (enrichment *P*-value 7.4×10^{-6}; see section Materials and methods) weakly influence *PHO84* and the expression data represent the combined influence of these factors (Figure 6D).

Discussion

We systematically applied Camelot to predict growth of 104 yeast strains in the presence of one of a panel of 94 diverse drugs. Camelot consistently performed well and successfully built robust predictive models for 87/94 drugs. It is intriguing that a single gene expression profile measured in the absence of any drugs empowered the prediction of traits under novel conditions (+drugs) that are dramatically different from the perturbation-free conditions used for expression profiling.

The models constructed by Camelot are not 'black box' predictors, but explain the variation in phenotype between the segregants by identifying the genes that influence the phenotype. We use gene expression data to pinpoint causal variants within large linked regions and to identify genes, outside linked regions, whose change in expression mediates the drug response. For each feature type (transcript and marker) we took the two largest hubs (i.e., a gene associated with many drugs) and systematically validated Camelot's predictions. We also identified a new causal gene *GPB2* and linked it to a number of drugs including the antidepressant haloperidol. Twenty-five out of 27 predictions of causal factors associated with response to a drug were confirmed, demonstrating that our method is robust. By incorporating signal from gene expression, Camelot not only identifies the causal genes driving the phenotype, but also provides insights into changes in the underlying regulatory network and the mechanisms involved. For example, the results from Camelot suggest a role for mitochondria in response to a number of drugs.

Identification of a transcript feature does not necessarily mean that the amount of transcript is responsible for the difference in phenotype between the strains. In the case of *DHH1* (whose coding sequence is identical in BY and RM), it is likely that a difference in *DHH1* expression accounts for variation in the regulation of mitochondrial biogenesis genes

between individual segregants, and that this influences the drug response. However, for *PHO84*, it is likely variation in Pho84 function that accounts for the differences in drug sensitivity and that gene expression varies through a feedback mechanism that 'reports' the difference in protein function. We note that a large number of the detected linkages between BY and RM involve feedback loops, including *AMN1*, *HAP1*, *HAP4* and *ZAP1* (Ronald *et al*, 2005). This could explain why expression frequently helps in the identification of differences in protein function, including in human genetics, where strong *cis*-eQTLs have been identified for genes whose cognate proteins harbour functional variation associated with human disease, for example, *SORT1* associated with lipid metabolism (Willer *et al*, 2008) and multiple genes associated with metabolic traits (Emilsson *et al*, 2008).

Others have demonstrated the predictive value of gene expression towards classifying phenotype (Golub *et al*, 1999; Alizadeh *et al*, 2000; van't Veer *et al*, 2002; Huang *et al*, 2007; Kutalik *et al*, 2008) and how integrating genotype and gene expression data could be used to better understand the relationship between genotype and phenotype in populations (Mehrabian *et al*, 2005; Schadt *et al*, 2005; Chen *et al*, 2008; Emilsson *et al*, 2008). However, to our knowledge, Camelot is the first approach to both quantitatively predict phenotype and identify genes that causally affect the phenotype. Central to Camelot is the interplay between causality and predictability; causal genes are better predictors and good predictors are more likely to be causal. Optimization of Camelot for both goals concurrently results in the model's exceptionally robust performance across an unprecedented number of complex traits.

Camelot uses gene expression data generated under control conditions to predict the phenotype under a new condition. The additional power gained from gene expression is remarkable given that the gene expression and genotype data used here were generated in the absence of drugs, two years before the generation of the growth (drug) data in another laboratory (Brem and Kruglyak, 2005; Perlstein *et al*, 2007). This shows that our results are based on a robust phenomenon and represent an inherent characteristic of the segregants. They are compatible with our work demonstrating that genetic variation alters cell state and predisposes the segregants towards different cellular responses (Litvin *et al*, 2009). We propose that gene expression is useful because it integrates information from multiple loci that are individually too weak to detect, but which, in combination, contribute significantly to the phenotype (Figure 6D). In this way, the combined influence of a large number of weak linkages (many of which are undetectable) can explain a large part of the heritable variation and as a consequence, gene expression data, generated under reference conditions, helps in predicting the response of segregants to new drugs. Three explanations are likely; the gene expression data might reflect (i) whether the cell is 'prepared' to tolerate a particular type of insult (Tagkopoulos *et al*, 2008), (ii) genetic variation in the regulatory network and the manner in which it is perturbed in response to the conditions or (iii) genetic variation in protein function via feedback loops. We expect that one or more of these explanations describe the situation for distinct phenotypes, genes and conditions.

Camelot's integration of genotype and gene expression not only enhances its ability to pinpoint causal genes, but it can also potentially identify the mechanism of action and the biological processes involved, thereby expanding the number of drug targets, for example, by identifying a connection between Dhh1 and mitochondria. Our method, therefore, has immediate application for identifying alternative or novel drug targets, for example, in drug-resistant pathogens. Our approach is highly robust and is applicable to other phenotypes and species, including humans. For example, genotype and gene expression data generated from each patient in the non-perturbed (non-diseased or non-drugged) state prior to the onset of disease could be used to predict outcomes (positive or negative responses to a drug or adverse reactions) in response to the therapeutic interventions under consideration. A critical feature is that appropriate drugs/interventions could be predicted for the healthy individual before a drug is administered. While the statistical and algorithmic improvement required to accommodate a genome of greater scale and complexity carries a heavy statistical burden, Camelot provides another step towards the realization of personalized medicine, as well as highlighting the power to be gained by exploiting gene expression data for this application.

Materials and methods

Data and pre-processing

The strain, genotype and gene expression measurements used are those of Brem and Kruglyak (2005). Growth yields in the presence of chemicals ('drug'), consisting of 313 growth conditions (different concentrations of chemicals and time points) and 94 different chemical molecules, were from Perlstein *et al* (2007). These include genotype, gene expression and drug response data for 104 strains. We merged adjacent, highly-correlated markers, to obtain a total of 526 markers (Lee *et al*, 2006). For our analysis, we normalized all data to have a mean of 0 and variance of 1. We compiled a list of candidate gene expression features based on two sources. One contained genes with potential regulatory effects, including transcription factors, signalling molecules, chromatin factors and RNA factors, as described by Lee *et al* (2006). The other list included genes involved in vacuolar transport, endosome, endosome transport and vesicle-mediated transport, since these functions, or cellular compartments, are enriched for multi-drug resistance genes (Hillenmeyer *et al*, 2008). We combined these two lists and filtered out genes with s.d.$\leqslant$0.2 in expression level, obtaining 854 expression profiles, which were used as candidate features for all our models. GO categories from http://www.yeastgenome.org/ were used to associate genes with each category.

Generation of the *GPB2*$_{RM}$ AS strain

The mating type of BY4724 was first switched to generate HCY413 using a plasmid that expresses *HO* from a *GAL* promoter. BY strains harbouring the *GPB2* coding sequence from RM11-1a were generated using the *Delitto Perfetto* method of Storici and Resnick (Storici *et al*, 2003; Storici and Resnick, 2006). The *GPB2* coding sequence and 5′UTR were sequenced to confirm that the coding sequence of the AS strain matched that of RM, whereas the upstream region remained that of BY. Primers used in this study are listed in Supplementary Table II.

Validation growth experiments

Strains used in this study are listed in Supplementary Table I. The *MKT1*-SK1 (D55N) and *PHO84* (L259P) AS strains are as described (Deutschbauer and Davis, 2005; Perlstein *et al*, 2007). The *dhh1*Δ

strain was a gift from Liz Miller (Columbia University). *MKT1*, *PHO84* and *dhh1*Δ growth yield experiments were performed in multi-well (96- or 384-well) plates as described by Perlstein *et al* (2007). Serial dilutions were carried out in replicate and the resulting growth yield and IC_{50} values generated using GraphPad Prism (v. 4.01).

For plate assays, overnight cultures of cells were grown in YPD medium at 30°C, diluted to $OD_{600} \sim 0.1$ and plated at 10-fold dilutions on YPD plates containing DMSO or DMSO + drug. Plates were incubated at 30°C or room temperature for 1–2 days for *GPB2* or 5 days for the arsenate assay. Final concentrations of drugs were as described by Perlstein *et al* (2007). For *GPB2*: gliotoxin (15.3 μM), E6-berbamine (16.5 μM) and haloperidol (66.6 μM); for *PHO84*: arsenate was used at a final concentration of 2 mM and methylphosphonate at 10 mM as described by Mouillon and Persson (2005).

Outline of the Camelot algorithm

An outline of the Camelot algorithm, including the triangle test and zoom-in score, is provided here; see Supplementary information for technical details of the statistical procedures and computational steps.

As input, Camelot is given a matrix **X** of features by segregants for two types: (1) Genotypes of genomic markers, **L**, derived from SNP microarrays and (2) gene expression data, **E**, obtained using microarrays under standard conditions (no drug). Additionally, Camelot is given a target matrix **Y** of drugs by segregants; each row represents D, the response of 104 segregants in the presence of a drug at a particular dose and time point.

For each drug response D, Camelot selects a linear regression model, $D \sim \{L\} + \{E\}$, involving a small number of selected features ($\{L\}$ and $\{E\}$). The objective is to select a model that is accurate in its prediction of D and whose features are likely to have a causal influence on D, that is, experimentally altering these features (by allele swap, deletion or overexpression) influences the drug response D. The Camelot algorithm progresses in three steps: feature selection, causality testing and model refinement.

A biologically plausible model should have a small number of causal factors with a non-zero weight. To achieve this goal, we use the elastic net (Zou and Hastie, 2005) regression method to select only the most significant features. Briefly, we solve the optimization problem

$$\hat{\beta} = \arg\min_{\beta} |D - \mathbf{X}\beta|^2,$$

$$\text{subject to } (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \leqslant t \text{ for some } t,$$

where D represents growth, **X** is the feature matrix (both D and **X** are standardized), β ($\hat{\beta}$, the solution) is the vector of coefficients and α and t are regularization parameters chosen using a 10-fold cross-validation procedure. The regularization terms reduce over-fitting the data. The constraint enforced by the l_1 norm assures sparseness of selected features and l_2-norm prevents arbitrary choice of only one out of several highly correlated features. The latter is especially important in the gene expression domain, which is abundant in large groups of highly correlated features. To compute the coefficients $\hat{\beta}$, we used least angle regression (LARS) (Efron *et al*, 2004).

However, the elastic net target function optimizes only for prediction error, which is a proxy for the goal of identifying underlying causal features. Not all predictive features are necessarily causal and indeed elastic net regression alone yields models with too many features (Supplementary Figure 5). We further reduce the number of selected features using non-parametric bootstrap (Efron, 1979). Indeed, our performance on synthetic data demonstrates that wrapping elastic net with bootstrapping enhances the precision with which we identify causal factors (Supplementary Figure 1).

The elastic net and bootstrap procedures are used to generate an initial small set of high-quality candidate features. For each selected transcript feature ($\{E\}$), we use the triangle test for causality (see below) to refine our set of selected features. To improve the likelihood that the final feature set contains causal genes, transcript features that pass the triangle test are kept in the regression and their associated (genotype) markers are removed, whereas transcript features that fail the test are removed and replaced with their associated genotype markers. Once a final set of features is selected, regression coefficients predicting D (response to drug) are re-optimized.

Triangle test

The triangle test is applied to every transcript feature selected and is used to evaluate the likelihood that the gene is significantly predictive of the response to a drug, beyond the contribution of the linked genotype. Assuming a transcript feature E is selected for phenotype D, we use permutation testing to evaluate the significance of causal edge $E \rightarrow D$. This is carried out for all genetic loci L that link to E and is controlled for the influence of $L \rightarrow D$. More specifically, we assess the significance of association between E and D by permuting E fixed under the genotype L. If gene expression remains significantly predictive (even when permuted while keeping the allele at L fixed), we determine that E holds additional information beyond that encoded in the marker L and is likely a causal factor.

Zoom-in score

The zoom-in score is a Bayesian prioritization score that ranks all genes within a linked region, evaluating the likelihood that each gene is causal. It is used to pinpoint the causal gene variant responsible for creating the linkage signal, and is applied to each of the marker features selected. The method integrates three cues: 'Is the gene expression level a good predictor of drug resistance', that is, does the gene expression correlate with the drug resistance? 'Is the gene '*cis*-linked', that is, is the gene's expression linked to its own locus?' and 'How well is the sequence of the gene conserved across 19 yeast species (Wapinski *et al*, 2007)?', consistent with our intuition that deviations from the conserved sequence are more likely to have a causal influence. This allows us to prioritize genes within each linked genomic region for their potential effect on the phenotype D.

Let gene g reside in genotype L_g and have an expression profile E_g; we can decompose the joint probability $P(D, E_g, L_g)$ as follows:

$$P(D, E_g, L_g) = P(D|L_g, E_g)P(E_g|L_g)P(L_g)$$

We calculate both $P(D|L_g, E_g)$ and $P(E_g|L_g)$ using least-square fitting regression and $P(L_g)$ based on the conservation of the coding sequence (see Supplementary information for more details). The decomposed probability consists of three parts. The first term $P(D|L_g, E_g)$ explains the phenotype with both genotype and expression profile of gene g, suggesting g has a causal effect. The second and third terms act as prior probabilities that the gene has a causal role, independent of the specific phenotype.

Statistical analysis

Camelot, the elastic net L model and linkage analysis are evaluated with 10-fold cross-validation (n=93–94). Elastic net L models are derived in the same way as Camelot models, except that only genotype features were allowed in regression. Linkage analysis is performed with Wilcoxon rank-sum test to scan the 526 merged markers for genome-wide significant linkages (FDR=2%, $P < 5.6 \times 10^{-5}$) (Perlstein *et al*, 2007). Linear regression models are built on significant linkages using robust regression (robustfit function in Matlab). Classification accuracy is used to evaluate predictions of models. Growth data are discretized into three classes according to their normalized values: resistant to the drug, no significant response and sensitive to the drug. Predictions of responses to drugs were made based on the predicted values from regression models. Classification accuracy (Acc) is defined as the number of correct classifications divided by the number of test data.

The significance of the interaction between *PHO84* and *GPB2* alleles is assessed by Wilcoxon rank-sum test, where segregants with both the *PHO84*-RM and *GPB2*-BY alleles are treated as one sample and the other segregants as another independent sample. Enrichment of phosphate metabolism/transport-related genes (GO annotation) in the linked regions shown in Figure 6D was calculated using the hypergeometric distribution. Each linked marker was expanded to 40 kb for the purposes of enrichment analysis. The linkages for *PHO84* gene expression were obtained with Wilcoxon rank-sum test ($P < 0.01$).

RT–PCR of *PHO84*

RT–PCR experiments were carried out to quantify the abundance of the *PHO84* transcript. Total RNA was prepared using the Ambion RiboPure-Yeast kit according to the manufacturer's instructions, with the exception that a 10 μg sample was digested twice each for 1 h at 37°C with 2 U DNase I. cDNA was made using the Stratagene AffinityScript kit and random primers. RT–PCR was performed with a Chromo4 machine (BioRad) using iQ SYBR Green Supermix (BioRad) and primers listed in Supplementary Table II. Data were scaled to *ERV25* (Pfaffl, 2001).

For the low and high-phosphate conditions, overnight cultures were washed twice with sterile distilled water and used to inoculate SC medium containing low phosphate (250 μM). After at least two doublings, cultures were split in two and phosphate was added at 15 mM final concentration ('high phosphate') to one flask. The cultures were grown for a further 80 min before harvesting. Phosphate media was made using YNB–potassium phosphate (Sunrise Science) supplemented with amino acids, glucose, ammonium sulphate and potassium phosphate. Potassium chloride (10 mM) was added to low-phosphate media.

Supplementary information

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

Acknowledgements

This research was supported by the National Institutes of Health Roadmap Initiative, NIH Director's New Innovator Award Program, through Grant number 1-DP2-OD002414-01 and National Centers for Biomedical Computing Grant 1U54CA121852-01A1. DP holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. NLG is supported by NIH G12 RR003037-24-2245476. We thank Ron Davis for the kind gift of YAD350 and Fred Winston for FY1333. We also wish to thank Oren Litvin, Itsik Pe'er, Aviv Regev, Eran Segal, Olga Troyanskaya, Lyle Ungar and Dennis Wykoff for valuable comments. *Author contributions:* BJC, HCC, NLG and DP designed research; BJC and DP designed the Camelot method; BJC implemented the Camelot method; BJC, HCC and DP analysed the data; EOP performed the drug validation for *DHH1*, *PHO84* and *MKT1*; DM and HCC constructed the *GPB2* allele swap; BJC and HCC performed all experiments related to *PHO84* feedback and carried out the drug validation for *GPB2*; and BJC, HCC and DP wrote the paper.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Airoldi EM, Huttenhower C, Gresham D, Lu C, Caudy AA, Dunham MJ, Broach JR, Botstein D, Troyanskaya OG (2009) Predicting cellular growth from gene expression signatures. *PLoS Comput Biol* **5:** e1000257

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson Jr J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC *et al* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403:** 503–511

Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci USA* **102:** 1572–1577

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* **296:** 752–755

Cardoso CM, Custodio JB, Almeida LM, Moreno AJ (2001) Mechanisms of the deleterious effects of tamoxifen on mitochondrial respiration rate and phosphorylation efficiency. *Toxicol Appl Pharmacol* **176:** 145–152

Chellappa R, Kandasamy P, Oh CS, Jiang Y, Vemula M, Martin CE (2001) The membrane proteins, Spt23p and Mga2p, play distinct roles in the activation of *Saccharomyces cerevisiae* OLE1 gene expression. Fatty acid-mediated regulation of Mga2p activity is independent of its proteolytic processing into a soluble transcription activator. *J Biol Chem* **276:** 43548–43556

Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA *et al* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature* **452:** 429–435

Cheung VG, Spielman RS (2002) The genetics of variation in gene expression. *Nat Genet* **32:** 522–525

Deutschbauer AM, Davis RW (2005) Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat Genet* **37:** 1333–1340

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO (2007) A genome-wide association study of global gene expression. *Nat Genet* **39:** 1202–1207

Efron B (1979) Bootstrap methods: another look at the jackknife. *Ann Stat* **7:** 1–26

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. *Ann Stat* **32:** 407–451

Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S *et al* (2008) Genetics of gene expression and its effect on disease. *Nature* **452:** 423–428

Evans GB, Furneaux RH, Gainsford GJ, Murphy MP (2000) The synthesis and antibacterial activity of totarol derivatives. Part 3: modification of ring-B. *Bioorg Med Chem* **8:** 1663–1675

Friedman M (2002) Tomato glycoalkaloids: role in the plant and in the diet. *J Agric Food Chem* **50:** 5751–5780

Gabriel SB, Salomon R, Pelet A, Angrist M, Amiel J, Fornage M, Attie-Bitach T, Olson JM, Hofstra R, Buys C, Steffann J, Munnich A, Lyonnet S, Chakravarti A (2002) Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nat Genet* **31:** 89–93

Garcia-Rodriguez LJ, Gay AC, Pon LA (2007) Puf3p, a Pumilio family RNA binding protein, localizes to mitochondria and regulates mitochondrial biogenesis and motility in budding yeast. *J Cell Biol* **176:** 197–207

Giots F, Donaton MC, Thevelein JM (2003) Inorganic phosphate is sensed by specific phosphate carriers and acts in concert with glucose as a nutrient signal for activation of the protein kinase A pathway in the yeast *Saccharomyces cerevisiae*. *Mol Microbiol* **47:** 1163–1181

Golub TR, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286:** 531–537

Harashima T, Anderson S, Yates III JR, Heitman J (2006) The kelch proteins Gpb1 and Gpb2 inhibit Ras activity via association with the yeast RasGAP neurofibromin homologs Ira1 and Ira2. *Mol Cell* **22:** 819–830

Hastie T, Tibshirani R, Friedman JH (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: with 200 Full-Color Illustrations*, p xvi, 533; ISBN: 0387952845 (alk. paper). New York: Springer

Hasumi K, Shinohara C, Naganuma S, Endo A (1992) Inhibition of the uptake of oxidized low-density lipoprotein in macrophage

J774 by the antibiotic ikarugamycin. *Eur J Biochem* **205:** 841–846

Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D, Altman RB, Davis RW, Nislow C, Giaever G (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320:** 362–365

Huang RS, Duan S, Bleibel WK, Kistner EO, Zhang W, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME (2007) A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc Natl Acad Sci USA* **104:** 9758–9763

Jiang Y, Vasconcelles MJ, Wretzel S, Light A, Gilooly L, McDaid K, Oh CS, Martin CE, Goldberg MA (2002) Mga2p processing by hypoxia and unsaturated fatty acids in *Saccharomyces cerevisiae*: impact on LORE-dependent gene expression. *Eukaryot Cell* **1:** 481–490

Kandasamy P, Vemula M, Oh CS, Chellappa R, Martin CE (2004) Regulation of unsaturated fatty acid biosynthesis in *Saccharomyces*: the endoplasmic reticulum membrane protein, Mga2p, a transcription activator of the OLE1 gene, regulates the stability of the OLE1 mRNA through exosome-mediated mechanisms. *J Biol Chem* **279:** 36586–36592

Kaneko M, Sato K, Horikoshi R, Yaginuma M, Yaginuma N, Shiragata M, Kumashiro H (1992) Effect of haloperidol on cyclic AMP and inositol trisphosphate in rat striatum *in vivo*. *Prostaglandins Leukot Essent Fatty Acids* **46:** 53–57

Khoury CM, Yang Z, Li XY, Vignali M, Fields S, Greenwood MT (2008) A TSC22-like motif defines a novel antiapoptotic protein family. *FEMS Yeast Res* **8:** 540–563

Kutalik Z, Beckmann JS, Bergmann S (2008) A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat Biotechnol* **26:** 531–539

Lee CS, Park SY, Ko HH, Song JH, Shin YK, Han ES (2005) Inhibition of MPP + -induced mitochondrial damage and cell death by trifluoperazine and W-7 in PC12 cells. *Neurochem Int* **46:** 169–178

Lee IH, Kim HY, Kim M, Hahn JS, Paik SR (2008) Dequalinium-induced cell death of yeast expressing alpha-synuclein–GFP fusion protein. *Neurochem Res* **33:** 1393–1400

Lee S-I, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, Koller D (2009) Learning a prior on regulatory potential from eQTL data. *PLoS Genet* **5:** e1000358

Lee SI, Pe'er D, Dudley AM, Church GM, Koller D (2006) Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci USA* **103:** 14062–14067

Litvin O, Causton HC, Chen B-J, Pe'er D (2009) Modularity and interactions in the genetics of gene expression. *Proc Natl Acad Sci USA* **106:** 6441–6446

Maller J, George S, Purcell S, Fagerness J, Altshuler D, Daly MJ, Seddon JM (2006) Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat Genet* **38:** 1055–1059

Martin CE, Oh C-S, Jiang Y (2007) Regulation of long chain fatty acid synthesis in yeast. *Biochim Biophys Acta* **1771:** 271–285

Mehrabian M, Allayee H, Stockton J, Lum PY, Drake TA, Castellani LW, Suh M, Armour C, Edwards S, Lamb J, Lusis AJ, Schadt EE (2005) Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat Genet* **37:** 1224–1233

Mouillon JM, Persson BL (2005) Inhibition of the protein kinase A alters the degradation of the high-affinity phosphate transporter Pho84 in *Saccharomyces cerevisiae*. *Curr Genet* **48:** 226–234

Nicolson K, Evans G, O'Toole PW (1999) Potentiation of methicillin activity against methicillin-resistant *Staphylococcus aureus* by diterpenes. *FEMS Microbiol Lett* **179:** 233–239

Nulton-Persson AC, Szweda LI (2001) Modulation of mitochondrial function by hydrogen peroxide. *J Biol Chem* **276:** 23357–23361

Pearl J (2000) *Causality: Models, Reasoning, and Inference*, p xvi, 384; ISBN: 0521773628. Cambridge: Cambridge University Press

Peeters T, Louwet W, Gelade R, Nauwelaers D, Thevelein JM, Versele M (2006) Kelch-repeat proteins interacting with the Galpha protein Gpa2 bypass adenylate cyclase for direct regulation of protein kinase A in yeast. *Proc Natl Acad Sci USA* **103:** 13034–13039

Perlstein EO, Ruderfer DM, Ramachandran G, Haggarty SJ, Kruglyak L, Schreiber SL (2006) Revealing complex traits with small molecules and naturally recombinant yeast strains. *Chem Biol* **13:** 319–327

Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L (2007) Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nat Genet* **39:** 496–502

Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Res* **29:** e45

Ronald J, Brem RB, Whittle J, Kruglyak L (2005) Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet* **1:** e25

Safiulina D, Veksler V, Zharkovsky A, Kaasik A (2006) Loss of mitochondrial membrane potential is associated with increase in mitochondrial volume: physiological role in neurones. *J Cell Physiol* **206:** 347–353

Saint-Georges Y, Garcia M, Delaveau T, Jourdren L, Le Crom S, Lemoine S, Tanty V, Devaux F, Jacq C (2008) Yeast mitochondrial biogenesis: a role for the PUF RNA-binding protein Puf3p in mRNA localization. *PLoS ONE* **3:** e2293

Sancho P, Galeano E, Nieto E, Delgado MD, Garcia-Perez AI (2007) Dequalinium induces cell death in human leukemia cells by early mitochondrial alterations which enhance ROS production. *Leuk Res* **31:** 969–978

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A *et al* (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37:** 710–717

Smith EN, Kruglyak L (2008) Gene–environment interaction in yeast gene expression. *PLoS Biol* **6:** e83

Storici F, Durham CL, Gordenin DA, Resnick MA (2003) Chromosomal site-specific double-strand breaks are efficiently targeted for repair by oligonucleotides in yeast. *Proc Natl Acad Sci USA* **100:** 14994–14999

Storici F, Resnick MA (2006) The *delitto perfetto* approach to *in vivo* site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast. *Methods Enzymol* **409:** 329–345

Tagkopoulos I, Liu YC, Tavazoie S (2008) Predictive behavior within microbial genetic networks. *Science* **320:** 1313–1317

Tuquet C, Dupont J, Mesneau A, Roussaux J (2000) Effects of tamoxifen on the electron transport chain of isolated rat liver mitochondria. *Cell Biol Toxicol* **16:** 207–219

Turalba AV, Leite-Morris KA, Kaplan GB (2004) Antipsychotics regulate cyclic AMP-dependent protein kinase and phosphorylated cyclic AMP response element-binding protein in striatal and cortical brain regions in mice. *Neurosci Lett* **357:** 53–57

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van de Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415:** 530

Vance D, Goldberg I, Mitsuhashi O, Bloch K (1972) Inhibition of fatty acid synthetases by the antibiotic cerulenin. *Biochem Biophys Res Commun* **48:** 649–656

Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449:** 54–61

Waring P, Khan T, Sjaarda A (1997) Apoptosis induced by gliotoxin is preceded by phosphorylation of histone H3 and enhanced sensitivity of chromatin to nuclease digestion. *J Biol Chem* **272:** 17929–17936

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D *et al* (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40:** 161–169

Wykoff DD, Rizvi AH, Raser JM, Margolin B, O'Shea EK (2007) Positive feedback regulates switching of phosphate transporters in *S. cerevisiae*. *Mol Cell* **27:** 1005–1013

Yip KW, Mao X, Au PY, Hedley DW, Chow S, Dalili S, Mocanu JD, Bastianutto C, Schimmer A, Liu FF (2006) Benzethonium chloride: a novel anticancer agent identified by using a cell-based small-molecule screen. *Clin Cancer Res* **12:** 5557–5569

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Statist Soc* **B:** 301–320