Harnessing gene expression to identify the genetic basis of drug resistance
Supplementary Information

Table of Contents

Camelot Algorithm

Camelot algorithm aims to provide a causative model for quantitative traits. There are two parts, development and testing of the predictive model and post-linkage analysis. The predictive model provides a final regression model containing predictive and likely causative features to explain a trait. More precisely, Camelot takes a set of training data consisting of phenotype, genotype and gene expression data as input, and selects a small set of features to build a linear regression model. Let **L** be the markers representing genotype data, **E** the expression of transcripts and *D* the response of segregants to a drug. Camelot obtains a predictive model by selecting a subset of features from **L** and **E** (Box 1B):

$$D = \mathbf{L}^{*}\beta_{\mathbf{L}^{*}} + \mathbf{E}^{*}\beta_{\mathbf{E}^{*}},$$

where $\mathbf{L}^{*}$ and $\mathbf{E}^{*}$ are subsets of markers and transcripts, respectively; and $\beta_{\mathbf{L}^{*}}$ and $\beta_{\mathbf{E}^{*}}$ are associated coefficients. A regression model such as this has two distinct merits: it provides a "program" to reveal the causal factors behind the drug response and it offers a quantitative model for predicting the drug response of unseen segregants or new strains. After Camelot establishes a predictive model, the zoom-in post-linkage analysis is applied to prioritize genes within a linked region (markers in $\mathbf{L}^{*}$) responsible for a trait.

We describe Camelot's procedure to build a predictive model for a drug response. Camelot progresses in three steps to build a model: initial feature selection, causality testing and model revision. Procedures in each step are described.

Feature Selection

Our goal is to identify marker and/or transcript features that predict quantitative phenotypes (in our case growth in the presence of drugs) using a linear regression model:

$$D = \mathbf{X}^{*}\boldsymbol{ß}^{*},$$

where *D* is the response of each of segregants to the presence of a drug, $\mathbf{X}^{*}$ is a matrix containing a *set of selected* features ($\mathbf{L}^{*}$ and/or $\mathbf{E}^{*}$) for each segregant and $\boldsymbol{ß}^{*}$ is the vector of coefficients associated with markers or transcripts in $\mathbf{X}^{*}$.

Camelot attempts to identify features ($\mathbf{X}^{*}$) that are not only predictive, but are causative for the phenotype. We consider a gene to be causal if perturbing it (*e.g.* allele swap, deletion or over-expression) actively results in a change to the phenotype. Our assumption is that predictive features are more likely to be the causal factors underlying phenotypic variation. While correlation does not necessarily imply causation, Camelot has a number of procedures that reduce the pool of candidate features towards those more likely to be causal.

A biologically plausible model should have a small number of causal factors with a non-zero weight. To achieve this goal, we used the elastic net regression method (Zou and

Hastie, 2005) to select only the most significant features $\mathbf{X}^*$. In brief, elastic net regression solves the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta} |y - \mathbf{X}\beta|^2, \quad \text{subject to } (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \le t \text{ for some } t,$$

where y represents response (D), $\mathbf{X}$ is the feature matrix (containing marker $\mathbf{L}$ and/or transcript $\mathbf{E}$). Both y and $\mathbf{X}$ are standardized to mean of 0 and variance of 1. $\boldsymbol{\beta}$ (from here we use $\boldsymbol{\beta}$ to denote $\hat{\beta}$ for simplicity) is the vector of regression coefficients and α and t are regularization parameters. The regularization terms reduce over-fitting of the data. The constraint enforced by the ℓ_1 norm controls sparseness of selected features and ℓ_2-norm prevents an arbitrary choice of only one out of several highly correlated features. The latter is especially important in the gene expression domain in which large groups of highly correlated features are abundant. To compute the coefficients $\boldsymbol{\beta}$, we used least angle regression (LARS) (Efron *et al*, 2004), an efficient algorithm for solving this type of regularized regression problem. All implementation is in MATLAB, including a modified implementation (Sjöstrand, 2005) of elastic net. Parameters of elastic net were chosen using standard cross-validation techniques.

The elastic net target function optimizes for fitting error, which is only a proxy for identifying the correct underlying causal features. Not all predictive features are necessarily causal and 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). We randomly sampled the segregants with replacement, including all genotype, expression and growth data associated with each strain, to obtain 200 bootstrap datasets. We applied the described elastic net regression to each bootstrap dataset to obtain a set of regression coefficients $\boldsymbol{\beta}^B$, $B = 1...200$, where B indicates the index of the bootstrap set. Those features with a non-zero coefficient in $\boldsymbol{\beta}^B$, define a sparse set of regression solutions for the bootstrap set B. In order to obtain statistically robust features, we define the *selection-frequency*, γ_i, for each feature i:

$$\gamma_i = \sum_{B=1}^{200} \delta(\beta_i^B \neq 0) \Big/ 200,$$

where β_i^B is the coefficient of feature i in bootstrap model B, and δ is an indicator function. We obtain set $\mathbf{X}^*$ according to γ_i:

$$\mathbf{X}^* = \{X_i \mid \gamma_i \ge 0.5, i : \text{index of feature}\}.$$

That is, for each condition, we choose the set $\mathbf{X}^*$ to include features that have non-zero coefficients for at least 50% of the bootstrap runs. Indeed, our performance on synthetic data demonstrates that wrapping elastic net with bootstrapping enhances the precision with which Camelot identifies causal factors (See section on *Statistical Evaluation*).

Since we have two types of features (markers $\mathbf{L}$ and transcripts $\mathbf{E}$), we consider including them in $\mathbf{X}$ in two ways: (1) $\mathbf{X}$ contains both $\mathbf{L}$ and $\mathbf{E}$ (denoted as $\mathcal{LE}$ model), and (2) $\mathbf{X}$

contains **L** only (denoted as $\mathcal{L}$ model). The causality testing phase will consider features chosen by both $\mathcal{LE}$ and $\mathcal{L}$ models and select from these a subset of features that most likely have causal influence on the phenotype *D* (see *Triangle Test* and *Model Revision*).

Here, we use 526 processed markers (Lee *et al*, 2006) for **L** and 854 candidate transcripts selected using GO categories (see main text) for **E**. More precisely, $\mathcal{LE}$ and $\mathcal{L}$ models are formalized as follows:

$\mathcal{LE}$ model:

$$D \sim \mathbf{L}^{\mathcal{LE}} + \mathbf{E}^{\mathcal{LE}}$$
$$\mathbf{L}^{\mathcal{LE}} = \{L_m \mid \gamma_m^{\mathcal{LE}} \geq 0.5, m : \text{index of marker}\}$$
$$\mathbf{E}^{\mathcal{LE}} = \{E_g \mid \gamma_g^{\mathcal{LE}} \geq 0.5, g : \text{index of transcript}\}$$
$$\gamma_m^{\mathcal{LE}} = \sum_B^{N_B} \delta(\beta_m^{B,\mathcal{LE}} \neq 0) \Big/ N_B \,, \gamma_g^{\mathcal{LE}} = \sum_B^{N_B} \delta(\beta_g^{B,\mathcal{LE}} \neq 0) \Big/ N_B \,,$$

where L_m represents the genotype of marker *m*, E_g the expression of transcript *g*, $\beta_m^{B,\mathcal{LE}}$ the coefficient for marker *m* in $\mathcal{LE}$ model for bootstrap *B*, $\beta_g^{B,\mathcal{LE}}$ the coefficient for transcript *g* in $\mathcal{LE}$ model for bootstrap *B*, and N_B the number of bootstrap runs (200).

$\mathcal{L}$ model (termed *elastic net L* model in the main text):

$$D \sim \mathbf{L}^{\mathcal{L}}$$
$$\mathbf{L}^{\mathcal{L}} = \{L_m \mid \gamma_m^{\mathcal{L}} \geq 0.5, m : \text{index of marker}\}$$
$$\gamma_m^{\mathcal{L}} = \sum_B^{N_B} \delta(\beta_m^{B,\mathcal{L}} \neq 0) \Big/ N_B \,,$$

where $\beta_m^{B,\mathcal{L}}$ is the coefficient for marker *m* in $\mathcal{L}$ model for bootstrap *B*. Note that the solution $\mathbf{L}^{\mathcal{LE}}$ may be different from $\mathbf{L}^{\mathcal{L}}$ because transcript features are often chosen instead of markers when expression data are considered.

$\mathcal{LE}$ and $\mathcal{L}$ models now provide small sets of makers and/or transcripts that are highly predictive, but not necessarily causative. To further test whether the selected features are causal for the trait, we administer the *triangle test* on the selected transcripts.

Triangle Test

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 transcript, and hence the expression level serves as a mediator of the causal effect from genetic variation on the phenotype (Figure 2D), or (3) growth rate influences the abundance of the transcript. The last case is not considered in this experimental design, as

gene expression was measured in absence of drug. Therefore, we developed the triangle test to distinguish between the first two cases.

The triangle test is applied to a triplet of marker, transcript and phenotype (L_m, E_g, D) and is used to evaluate whether E_g is significantly predictive of D, beyond the contribution of L_m. We use permutation testing to evaluate the significance of association between E_g and D by permuting E_g fixed under the marker L_m. If the transcript E_g remains significantly predictive (even when permuted while keeping the allele at L_m constant), we determine that E_g holds additional information beyond that encoded by the marker L_m and is therefore more likely to be a causal factor, rather than chosen simply due to its correlation to L_m. The association is tested with linear regression; 10^6 permutations are performed to obtain empirical null distributions to assess the significance. We use p-value < 0.002 (corresponding to FDR = 0.004) as a threshold to determine if E_g is significantly more predictive to D than L and therefore E_g is likely a causal factor for D.

We apply the triangle test to each transcript feature E_g selected for the quantitative phenotype D (that is, $E_g \in \mathbf{E}^{\mathcal{LE}}$). In order to collect triplets (L_m, E_g, D), we search for markers L_m that link to transcript E_g, using the same feature selection procedure (bootstrapped elastic net) but with E_g as the response variable (y) and markers (**L**) as features (**X**). More precisely, $L_m \in \mathbf{L}_g^{e\mathcal{L}}$, which is obtained from '$e\mathcal{L}$ model' defined as

$$
\begin{aligned}
& E_g \sim \mathbf{L}_g^{e\mathcal{L}} \\
& \mathbf{L}_g^{e\mathcal{L}} = \{L_m \mid \gamma_m^{e\mathcal{L}(g)} \geq 0.5, m : \text{index of marker}\} \\
& \gamma_m^{e\mathcal{L}(g)} = \sum_B^{N_B} \delta(\beta_m^{B,e\mathcal{L}(g)} \neq 0) \Big/ N_B ,
\end{aligned}
$$

where $\beta_m^{B,e\mathcal{L}(g)}$ is the coefficient for marker m in $e\mathcal{L}$ model for bootstrap B. This model can be viewed as an eQTL model, in which multiple linkages are detected through bootstrapped elastic net regression. We only consider (L_m, E_g, D) as a triplet when the two associations (L_m, E_g) and (L_m, D) are significant (p-value < 0.05; such pairs are not of high number, since the number of L_m and E_g have been controlled by bootstrapped elastic net). The significance of associations (L_m, E_g) and (L_m, D) is determined by 10^6 permutations with least square fitting linear regression, in which L_m is used as an independent variable and E_g or D is used as a dependent variable.

Moreover, when a transcript feature is tested against multiple markers, we need to test whether the transcript is more predictive than the markers in regions near to those under consideration. Because genotypes of markers in neighbouring regions are highly correlated, neighbouring markers are often chosen as linkages. If we do not correct for these linkages, we might conclude, incorrectly, that the transcript feature is causal. For example, assuming we test two triplets (L_1, E_g, D) and (L_2, E_g, D) where L_1 and L_2 are neighbouring markers, it is possible that E_g shows stronger association to D than L_1, but not L_2. In this case, L_2 is more likely to be the causal factor than E_g. In order to determine whether E_g passes the triangle test, we require E_g to be more predictive of the phenotype than any of the linked neighbouring markers.

Based on the test results, we can define the following sets:

$$\mathbf{E}^{Cause} = \{E_g \mid E_g \in \mathbf{E}^{\mathcal{AE}} \text{ and } E_g \text{ passes test for at least one set of linked neighbouring markers}\},$$
$$\mathbf{L}^{Indirect} = \{L_m \mid L_m \in \mathbf{L}_g^{e\mathcal{A}} \text{ and } E_g \text{ passes test for the triplet } (L_m, E_g, D)\},$$

where $\mathbf{E}^{Cause}$ is the set of transcripts that pass the triangle test, that is, those that are likely to have a causal effect on the phenotype; and $\mathbf{L}^{Indirect}$ is the set of genotype markers that link these transcripts. The latter likely have indirect effects on D that are mediated by E_g.

In contrast, $\mathbf{E}^{Cor}$ is defined as the set of transcripts that fail the test, and $\mathbf{L}^{Cause}$ as the set of markers that link to transcripts in $\mathbf{E}^{Cor}$:

$$\mathbf{E}^{Cor} = \{E_g \mid E_g \in \mathbf{E}^{\mathcal{AE}} \text{ and } E_g \text{ fails test for all involved triplets}\},$$
$$\mathbf{L}^{Cause} = \{L_m \mid L_m \in \mathbf{L}_g^{e\mathcal{A}} \text{ and } E_g \text{ fails test for triplet } (L_m, E_g, D)\}.$$

This definition implies transcripts in $\mathbf{E}^{Cor}$ are associated with D due to co-regulation that results from upstream sequence variation, and $\mathbf{L}^{Cause}$ is likely the sequence variation that influences both D and E_g .

We further classify transcripts that pass the triangle test into two categories: *weak* and *strong* factors. We classify as strong those transcripts that pass the triangle test for all linked markers. We classify as weak those transcripts that only pass the triangle test for some of their linked markers. In addition, we use all different combinations of features (for example, regulators only or regulators with genotype as the feature pool) to broadly access genes related to drug response with the triangle test. All transcripts that pass the triangle test are listed in Supplementary Table III.

While our triangle test evaluates the significance of the association between E_g and D conditioned on L_m (i.e. the causal relationship E_g -> $D \mid L_m$), we compared our test to another test, which assesses the significance of the association between L_m and D conditioned on E_g (i.e. the causal relationship L_m -> $D \mid E_g$) (see Supplementary Table III). Out of the 317 triplets that passed the triangle test, only 12 are significant (same FDR as the triangle test) for the test L_m -> $D \mid E_g$. The remaining 305 triplets are in accordance with the triangle test, E_g explains away the relationship between L_m and D and the influence from L_m to D is no longer significant conditioned on E_g. These 305 triplets include all cases specifically discussed in the main text. The 12 triplets could either be false positives, or more interestingly, suggest true complex cases, in which both L_m and E_g have a direct causal effect on D. In fact, two-third of these 12 triplets involve a region on chromosome XIV (*MKT1* locus), the response to rapamycin, and *ERG4* or *MLP1* transcripts. Previous work has shown the complex influence of *MKT1* (personal communication), *ERG4* (Xie *et al*, 2005) and *MLP1* (Hillenmeyer *et al*, 2008) on the

resistance to rapamycin, supporting the complex causal relationship that L_m exerts an influence on D both through E_g and an additional factor.

Model Revision

Using the results of our triangle test, we revise our final feature set. For a phenotype D, our goal is to obtain the following model

$$D = \mathbf{L}^{Camelot}\beta_{\mathbf{L}^C} + \mathbf{E}^{Camelot}\beta_{\mathbf{E}^C},$$

where $\mathbf{L}^{Camelot}$ and $\mathbf{E}^{Camelot}$ are sets of genotype from selected marker features and the expression of selected transcript features, respectively; and $\beta_{\mathbf{L}^C}$ and $\beta_{\mathbf{E}^C}$ are the associated coefficients.

Considering all the feature sets we have obtained ($\mathbf{L}^{\mathcal{A}}$, $\mathbf{L}^{\mathcal{AE}}$, $\mathbf{L}^{Cause}$, $\mathbf{L}^{Indirect}$, $\mathbf{E}^{Cause}$, and $\mathbf{E}^{Cor}$), we derive $\mathbf{L}^{Camelot}$ and $\mathbf{E}^{Camelot}$ according to the following criteria:

$$\mathbf{L}^{Camelot} = (\mathbf{L}^{\mathcal{A}} \cup \mathbf{L}^{\mathcal{AE}} \cup \mathbf{L}^{Cause}) \setminus \mathbf{L}^{Indirect},$$
$$\mathbf{E}^{Camelot} = \mathbf{E}^{Cause}.$$

These criteria are aimed at enriching the final feature set for those that are more likely to be causative. Transcripts that pass the triangle test are more likely to act directly, while the linked upstream sequence variation ($\mathbf{L}^{Indirect}$) is likely to be indirect and act through $\mathbf{E}^{Cause}$. For transcripts ($\mathbf{E}^{Cor}$) that fail the triangle test, it is more likely that $\mathbf{L}^{Cause}$ are the common factors responsible for both D and $\mathbf{E}^{Cor}$, and the correlation between them.

Including the subset of $\mathbf{L}^{Cause}$ might introduce markers that are not significantly associated with D, so we only include those markers that have selection-frequency $\gamma_m^{\mathcal{A}}$ larger than 0.3. In addition, neighbouring regions in the final $\mathbf{L}^{Camelot}$ are corrected by only choosing the one with the highest selection-frequency. Once the final set of features have been selected, the regression coefficients were re-optimized using robust regression (robustfit function in Matlab).

Zoom-in Score

To pinpoint the causal variant responsible for the linkage signal, for each marker feature selected ($\mathbf{L}^{Camelot}$), we developed a Bayesian prioritization score that ranks genes within a linked region according the likelihood of their causal potential. The method integrates three cues: "Is the gene expression a good predictor of drug resistance?" (*i.e.*, if the gene expression correlates with the drug response), "Is the gene cis-linked?" (*i.e*, if the gene expression is linked to its own locus), and "how well is the gene sequence conserved?" which is consistent with our basic intuition that if nature conserved a residue across millions of evolutionary years, its change is more likely to have a causative influence on the associated phenotype.

Assume we have marker m linked to D (*i.e.* $L_m \in \mathbf{L}^{Camelot}$). To calculate the causal potential of each gene g located within the region around marker m, we define a zoom-in score as follows:

$$P(L_g, E_g, D) = P(D \mid L_g, E_g)\, P(E_g \mid L_g)\, P(L_g),$$

where E_g and L_g are the expression and genotype of gene g, respectively. Note that L_m can be used as an approximation of L_g since gene g is in proximity to marker m. In this model, we assume that sequence variation in gene g affects both E_g and D, and the expression of transcript g (E_g) also affects the phenotype D (Figure 4A in main text).

The decomposed probability consists of three parts. The first term $P(D \mid L_g, E_g)$ assesses how well D can be explained by both the genotype and expression profile of gene g; the second part $P(E_g \mid L_g)$ scores the degree of *cis*-linkage, how well E_g is explained by L_g; and the last term is the prior probability of g having a causal effect. $P(D \mid L_g, E_g)$ and $P(E_g \mid L_g)$ are calculated using the probability density function with normal distribution, where the mean and variance are estimated using linear regression ($D \sim L_g + E_g$ and $D \sim E_g$). $P(L_g)$ was estimated based on the conservation of the coding sequence, as follows.

We assume genes with polymorphisms between BY and RM are more likely to affect the phenotype, especially if the polymorphisms are in positions where amino-acid residues are conserved throughout evolution. Therefore, considering the fungal alignment of orthologs (Wapinski *et al*, 2007), we calculated a conservation score for each mismatched/gap in amino-acid sequence between BY and RM based on a quality score defined for multiple sequence alignment (Thompson *et al*, 1997). Let position j in the alignment has a mismatch/gap between BY and RM. We define a score s_j as follows:

$$s_j = exponential\,(-\, similarity(R_{j,BY}, R_{j,RM}) - min(D_{j,BY}, D_{j,RM})\,),$$

where $R_{j,BY}$ ($R_{j,RM}$) is the residue at position j in BY (RM), $D_{j,BY}$ ($D_{j,RM}$) is the distance (Thompson *et al*, 1997) defined for multiple sequence alignment between BY (RM) and other fungal species, and *similarity* is a similarity metric based on Gonnet PAM 250 matrix (Benner *et al*, 1994).

For each gene g, we then define

$$P(L_g) = sigmoid(\theta \sum s_j),$$

where s_j is defined as above and θ is a parameter chosen to adjust the distribution of $P(L_g)$.

When neighbouring regions were linked to the phenotype, we merged them into a larger region and ranked all potential genes within the merged region based on the zoom-in score. We calculated the joint probability defined above for each gene residing within 30,000 base pairs up-/down-stream of a linked region ($L_m \in \mathbf{L}^{Camelot}$). Genes without polymorphisms in coding and non-coding regions between BY and RM were disregarded.

Moreover, when calculating the probabilities, we used the original genotype data (Brem and Kruglyak, 2005) (missing values were imputed according to the distance between markers) to represent the locus more accurately. That is, instead of using the 526 merged markers, L_g was obtained from the nearest locus among the original 2957 markers, according the genomic location of *g*. The three components of decomposed probability were weighted so that $P(D \mid L_g, E_g)$ has the strongest effect and prior $P(L_g)$ the weakest. Finally, we ranked genes based on their zoom-in scores in the region.

Statistical Evaluation

The robustness and statistical significance of Camelot's performance were systematically evaluated. In this section, we demonstrate that bootstrapping enhances the precision with which we can identify causal factors using synthetic data. Moreover, using cross-validation we demonstrate superior performance of Camelot's ability to predict the response to drug treatment compared to classical linkage analysis.

Bootstrapping enhances accurate retrieval of causal factors
Because elastic net aims to minimize the fitting error, it often results in models with too many features that are not necessarily causal of the phenotype. In order to select robust and causal features, we use non-parametric bootstrap with elastic net regression (see *Feature Selection*). Here we evaluate if bootstrapping helps elastic net to select correct features using synthetic datasets.

First, we generated several synthetic datasets as follows. We used the LARS implementation of elastic net (Efron *et al*, 2004; Zou *et al*, 2005) to generate the full path of solutions for 20 conditions of growth data (randomly chosen from the original data), using a feature pool **X,** containing both **L** and **E**. Then we randomly chose 10 of the first 30 features that enter each solution path. Using these 10 features with their coefficients from elastic net, we generated the synthetic growth data for each condition. We added different levels of noise from Gaussian distribution $N(0, \sigma^2)$, with $\sigma^2 = 0.2, 0.4, 0.6, 0.8,$ or 1 to the data to mimic the noise found in real data. The advantage of the synthetic data is that we know the true features and can therefore get an accurate evaluation of Camelot's ability to retrieve causal features.

We applied elastic net to the synthetic growth data with and without bootstrapping. Feature selection with bootstrapping is as described in *Feature selection*. We compared the precision of feature retrieval between elastic net and bootstrapped elastic net. Precision is defined as the number of true features selected, divided by the number of total features selected in the procedure. Supplementary Figure 1 shows precision of bootstrapped elastic net verses elastic net and stepwise regression (used in (Chen *et al*, 2008)). For almost all the models, bootstrapped elastic net shows dramatically higher precision.

Evaluating our results on the actual drug data (rather than the synthetic set), we compare the number of features selected in elastic net with and without bootstrapping. Supplementary Figure 5 shows the histograms of number of features selected. Elastic net models without bootstrapping include many features (mostly between 10 to 30 features per model). On the other hand, bootstrapping significantly reduces the number of features. Most models derived from bootstrapped elastic net contain fewer than eight features, which is more biologically plausible. The final models and selected features are listed in Supplementary Table IV.

Cross-Validation analysis demonstrates the robustness of Camelot

We use ten-fold cross-validation to evaluate the performance of Camelot. We randomly split the data into ten equal parts. Each part contains growth, genotype and expression data for ~10 segregants. Holding out each part as test data, we took the rest of the data as training data (~94 segregants), and applied Camelot to obtain a linear regression model (Box 1B). This model was then used to generate predictions for the held-out segregants. Note that no data from the test segregants was used during model construction; therefore, ten-fold cross-validation provides a good way to evaluate the predictions of the models and their potential performance on additional new strains.

For comparison, the same ten-fold cross-validation procedure was also applied to the elastic net *L* model (see *Feature Selection*) and linkage analysis. Linkage analysis was performed using the Wilcoxon rank-sum test, with FDR=2% ($p<5.6\times10^{-5}$) (Perlstein *et al*, 2007) to determine significant linkages genome-wide. That is, during each fold of the cross-validation, significant linkages were obtained from the training data (n=93~94) with Wilcoxon rank-sum test and these linkages (markers) were treated as predictors in a linear regression model. Regression coefficients associated with these markers were obtained through robust regression (robustfit function in Matlab).

First, we use accuracy of classification (*Acc*) to evaluate the prediction from these three different models (*i.e.* Camelot, elastic net *L* and linkage analysis). Three classes are defined by discretisation of the normalised drug response: resistant to the drug (standardised growth > 1), no significant response to the drug (-1≤ standardised growth ≤ 1) and sensitive to the drug (standardised growth < -1). Similarly, prediction of class for a segregant in the test data is determined by the predicted value from the regression model. Accuracy (*Acc*) is defined as the number of correct predictions divided by the total number of segregants tested. Figure 1 (main text) shows *Acc* of models from Camelot, elastic net *L* and linkage analysis. As discussed in the main text, Camelot shows superior classification accuracy compared to the other two models.

Because classification could be biased due to discretisation, we further seek to compare the performance of models in continuous-value space. We use correlation coefficients to evaluate the prediction as they can be used to assess whether the prediction accords with the original growth data. We use both Pearson (r) and Spearman (ρ) correlation coefficients between the predicted response to drug and original growth data to evaluate the prediction. In addition, the median of growth data in the training set was used in place

of the predicted growth if an algorithm failed to generate a model (*i.e.* failed to link or select any features). As shown in the figures (Supplementary Figure 6 and 7), Camelot's prediction correlates better with the original growth data than prediction from linkage analysis or elastic net *L*. This shows that Camelot's predictions are robust and correlate with the observed data.

Taken together, cross-validation provides statistical evidence to support the robustness and superior performance of Camelot, compared to linkage analysis and elastic net *L* models, across a number of evaluation metrics. We show that Camelot's performance is robust, and not due to over-fitting, using ten-fold cross-validation. This demonstrates the Camelot's potential to predict the response of unseen strains, since only training data is used during cross-validation.

Supplementary Table I.

Strains used

Strain	Background	Genotype	Reference/source
FY1333	BY4724	*Mat alpha leu2Δ0 ura3Δ0*	(Kanta *et al*, 2006)
HCY413	BY	*Mat a leu2Δ0 ura3Δ0*	This study*
RM11-1a	RM	*Mat a leu2Δ0 ura3Δ0 ho::KanMX*	(Yvert *et al*, 2003)
HCY503	RM	*Mat alpha leu2Δ0 ura3Δ0 ho::KanMX*	This study*
HCY467	BY	*Mat a leu2Δ0 ura3Δ0* $GPB2_{RM}$	This study
YAD350	BY	*Mat alpha his3Δ0 leu2Δ0 lys2Δ0 ura3Δ0 MKT1(D30G)*	(Deutschbauer and Davis, 2005)
BY4722 L259P	BY	*Mat alpha leu2Δ0 ura3Δ0 PHO84* (L259P)	(Perlstein *et al*, 2007)

*Made by switching the mating types of FY1333 and RM11-1a respectively.

Supplementary Table II.

Primer

Name	Sequence (5' to 3')	Use
ERV25_F	ttcgtgttgcgtttactgct	RT-PCR
ERV25_R	gtgtctcttaatctctcctctct	RT-PCR
GPB2_F	ccgtcggcgttgccttatt	RT-PCR
GPB2_R	agtctgtcgacttggagatctt	RT-PCR
BY.GPB2_pGSKU_F	taaagattgtgattcattggcaggtccattgtcgcattactaaatcataggctagggataacagggtaatttggatggacgcaaagaagt	PCR
BY.GPB2_pGSKU_R	ttatattctactactaaacaaagtttacaaagtgaaagcattgaaaactgccttttttcgtacgctgcaggtcgac	PCR
5'UTR.GPB2_F	cgataagacggaatagaatagtaaagattgtgattcattggc	PCR
BY3'UTR.RMGPB2ORF_R	ctactactaaacaaagtttacaaagtgaaagcattgaaaactgctttttatgcactaggatttacactag	PCR
MKT1_F	ttggttgggcaagaaagatt	RT-PCR
MKT1_R	tttcgcagcatttagctcct	RT-PCR
PHO84_F	ctttgttctgtgtcatcggttt	RT-PCR
PHO84_R	agttggttggcttaccgtct	RT-PCR

References

Benner SA, Cohen MA, Gonnet GH (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng* **7:** 1323-1332.

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

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, Lusis AJ, Schadt EE (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature* **452:** 429-435.

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

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

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

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.

Kanta H, Laprade L, Almutairi A, Pinto I (2006) Suppressor analysis of a histone defect identifies a new function for the hda1 complex in chromosome segregation. *Genetics* **173:** 435-450.

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 U S A* **103:** 14062-14067.

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.

Sjöstrand K (2005) Matlab implementation of LASSO, LARS, the elastic net and SPCA. Informatics and Mathematical Modelling, Technical University of Denmark, DTU.

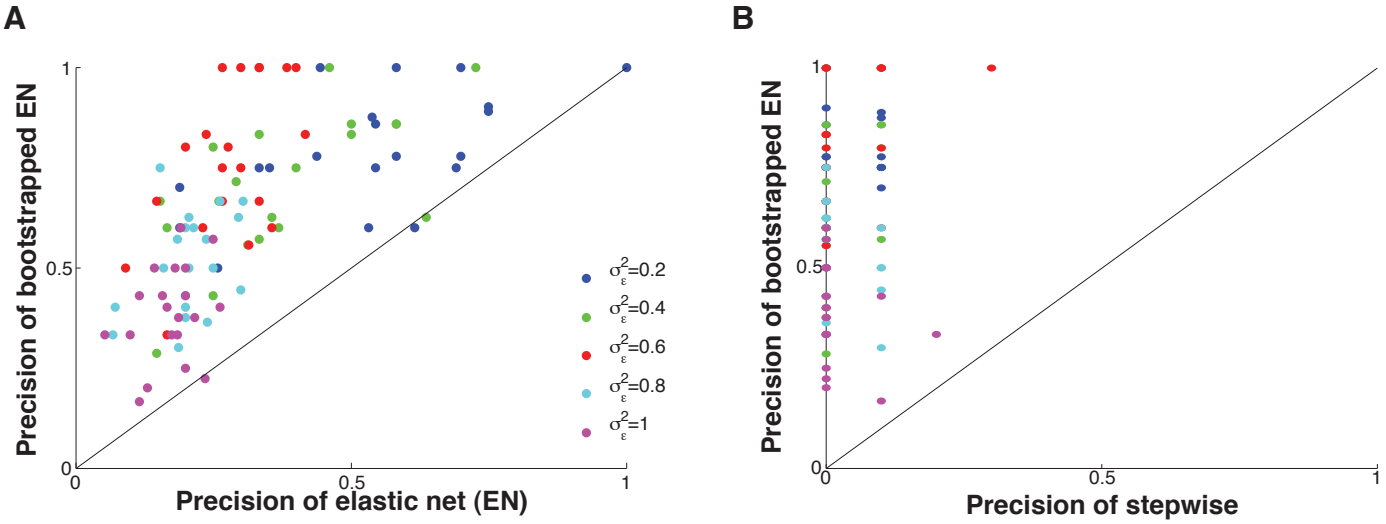
Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876-4882.

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

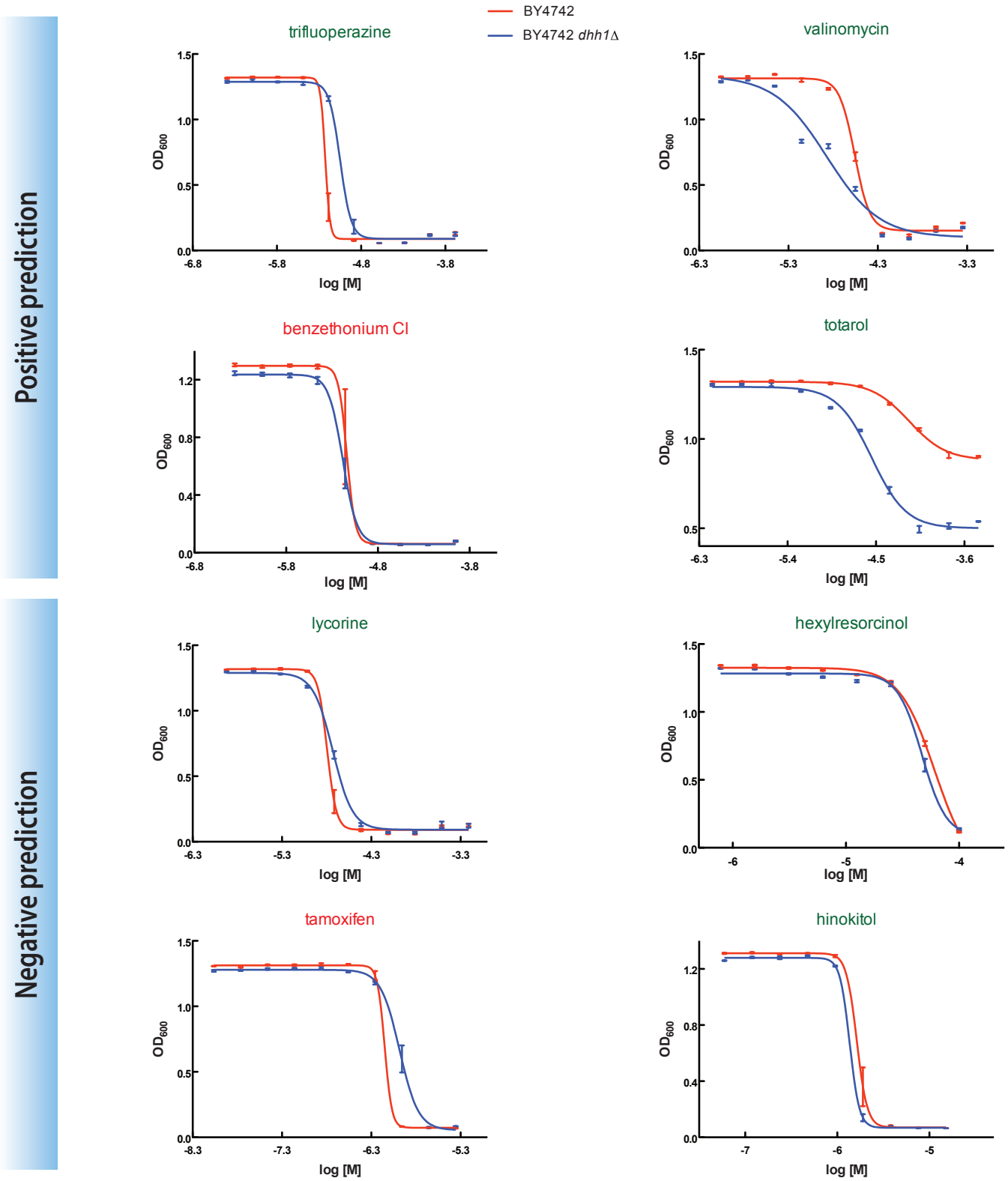
Xie MW, Jin F, Hwang H, Hwang S, Anand V, Duncan MC, Huang J (2005) Insights into TOR function and rapamycin response: chemical genomic profiling by using a high-density cell array method. *Proc Natl Acad Sci U S A* **102:** 7215-7220.

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L (2003) *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat Genet* **35:** 57-64.

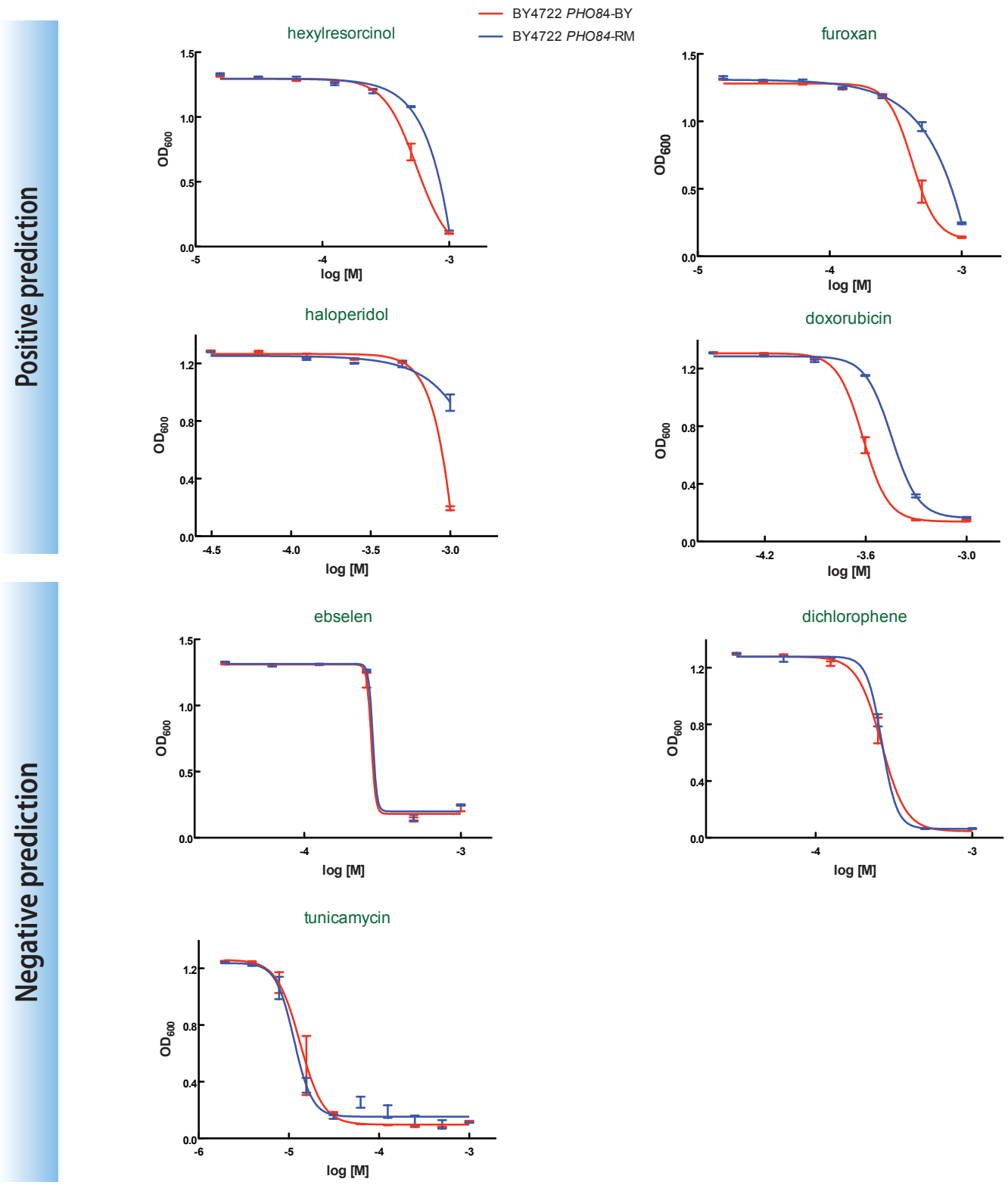
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Statist Soc* **B:** 301-320.



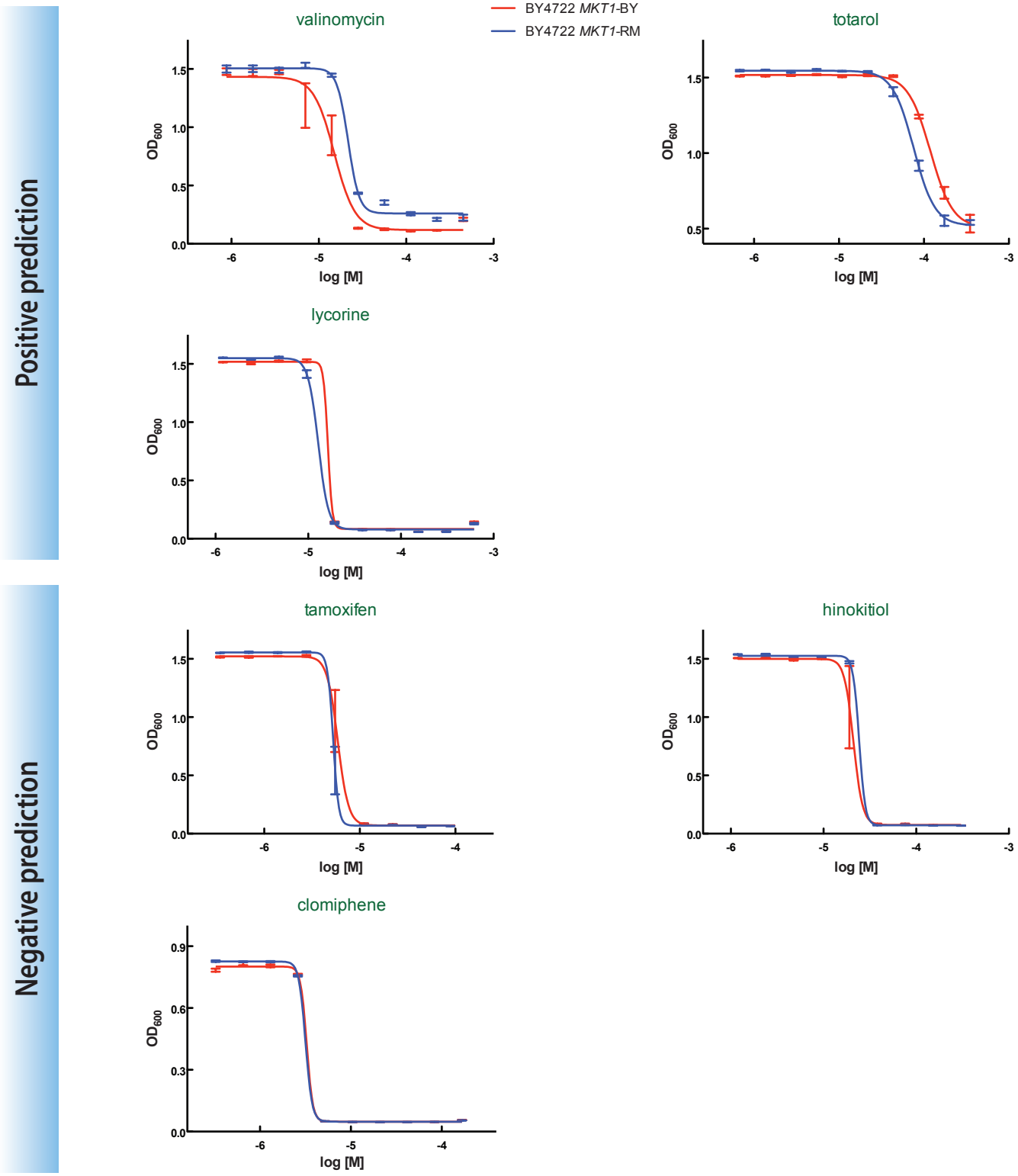
Supplementary Figure 1: Bootstrap improves precision in retrieving true factors. A: Bootstrapping improves the precision with which true genetic factors can be retrieved from a synthetic dataset. Precision obtained using elastic net with bootstrapping (y-axis) is compared with that from elastic net without bootstrapping (x-axis). Each dot represents a synthetic growth condition. The diagonal line shows where the two models have the same precision. The plot shows that bootstrapping significantly improves the precision with which the true features can be retrieved. Different coloured dots represent different levels of noise added to the synthetic data. **B:** Bootstrapped elastic net retrieves true factors more precisely than stepwise regression. Similar to **A**, precision with which factors are retrieved using elastic net with bootstrapping (y-axis) is compared with that from stepwise regression (x-axis). Bootstrapping is used to select robust features from elastic net whereas stepwise regression takes the top 10 significant features (the exact number of true factors used to generate the data). The plot demonstrates the superior precision of bootstrapped elastic net compared to that of stepwise regression. For example, when synthetic growth data contains noise generated from Gaussian distribution N(0, 0.2), the precision of stepwise regression to retrieve the true factors is limited between 0 and 0.3, while the precision of bootstrapped elastic net ranges from 0.5 to 1.0 (with an average of precision 0.8).



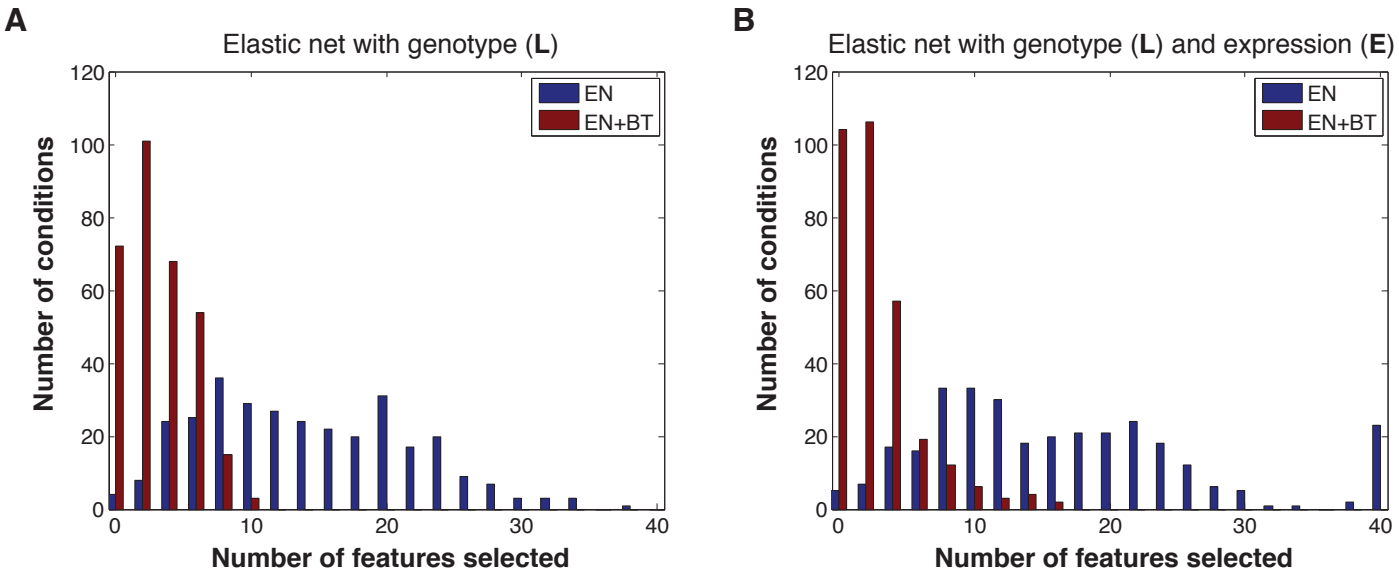
Supplementary Figure 2: Growth yield of BY and BY *dhh1*Δ strains in the presence of drugs. Validation of the causal effect of *DHH1* on the growth yield in the presence of specific drugs. Averaged OD_{600} absorbance growth measurements of BY (red) and BY *dhh1*Δ mutant (blue) are plotted against a twofold dilution series for each drug. The causal effect of *DHH1* was confirmed for our positive predictions, with the exception of benzethonium chloride. Negative controls show the specificity of the causal effect. Camelot predicted that the response to tamoxifen would be the same for BY and RM. Drugs written in green match Camelot's prediction, whereas drugs written in red do not.



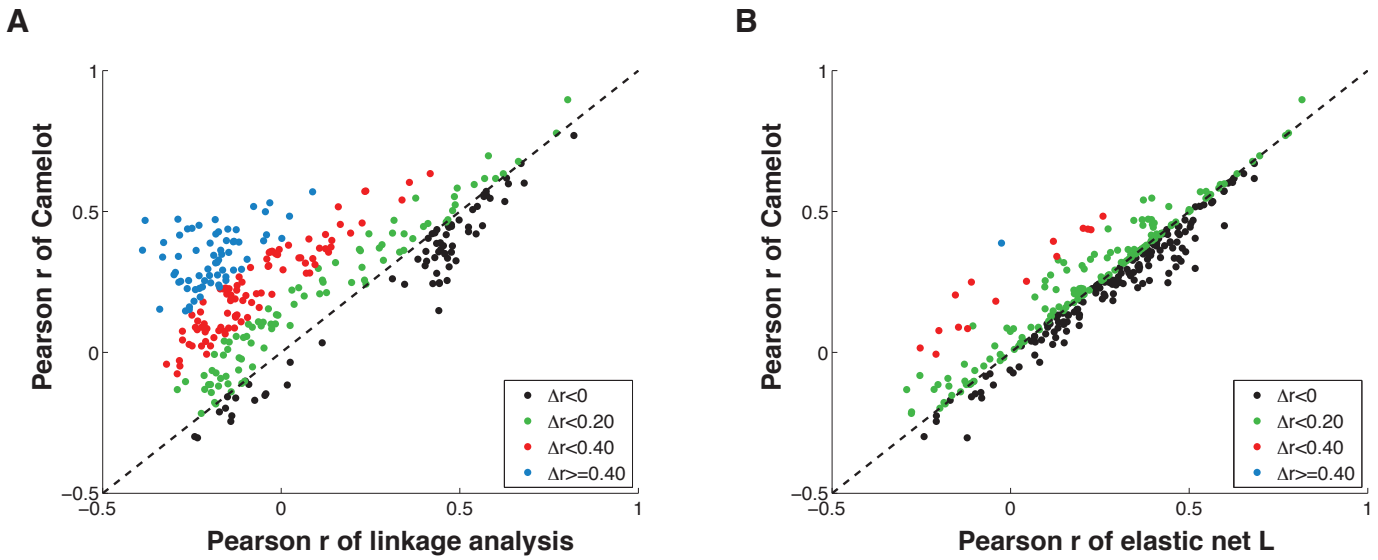
Supplementary Figure 3: Validation of Camelot's *PHO84* predictions. Validation of the causal effect of *PHO84* on the growth yield in the presence of specific drugs. Averaged OD_{600} absorbance growth measurements of BY (red) and BY with an allele swap for *PHO84*-RM (blue) are plotted against a twofold dilution series for each drug. All positive predictions from Camelot were confirmed, showing *PHO84* variant is the cause of the difference in response to these drugs between BY and RM. Negative controls show the ability of Camelot to distinguish drugs that are not affected by *PHO84*, even though they show significant linkage to the locus.



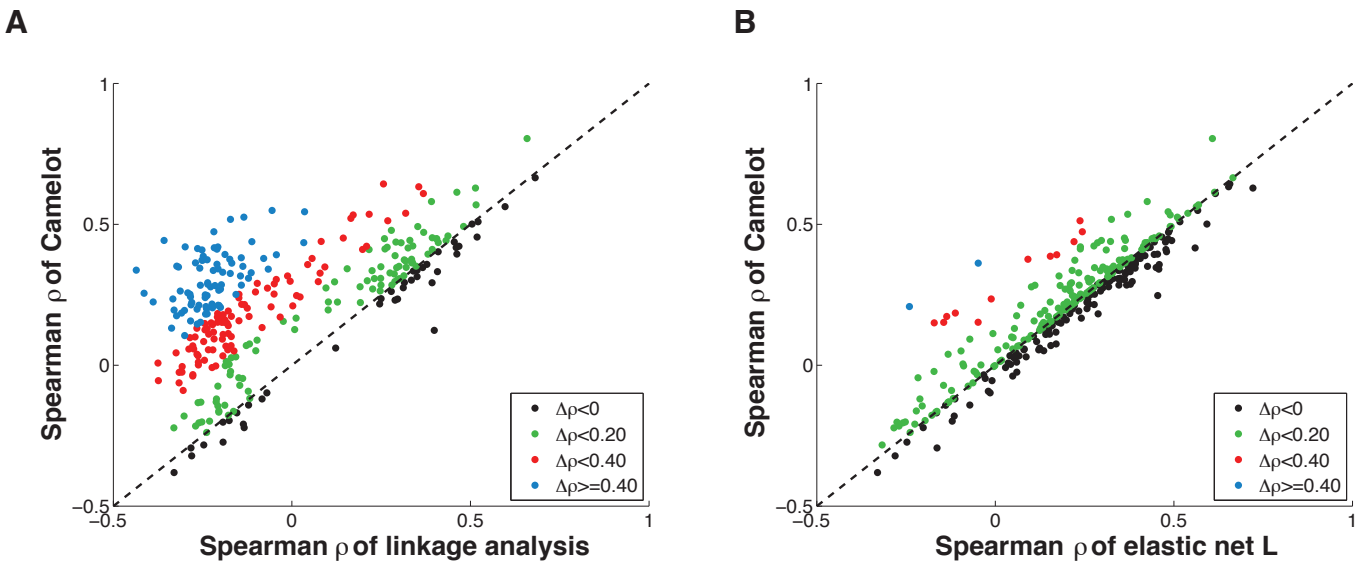
Supplementary Figure 4: Validation of Camelot's *MKT1* predictions. Validation of the causal effect of *MKT1* on the growth yield in the presence of specific drugs. Averaged OD_{600} absorbance growth measurements of BY (red) and BY with an allele swap for *MKT1*-RM (blue) are plotted against a twofold dilution series for each drug. All positive predictions from Camelot were confirmed, showing that the *MKT1* variant is the cause of the difference in response to these drugs between BY and RM. Negative predictions were also confirmed showing that Camelot can distinguish drugs that are not affected by *MKT1*, even though they show significant linkage to the locus.



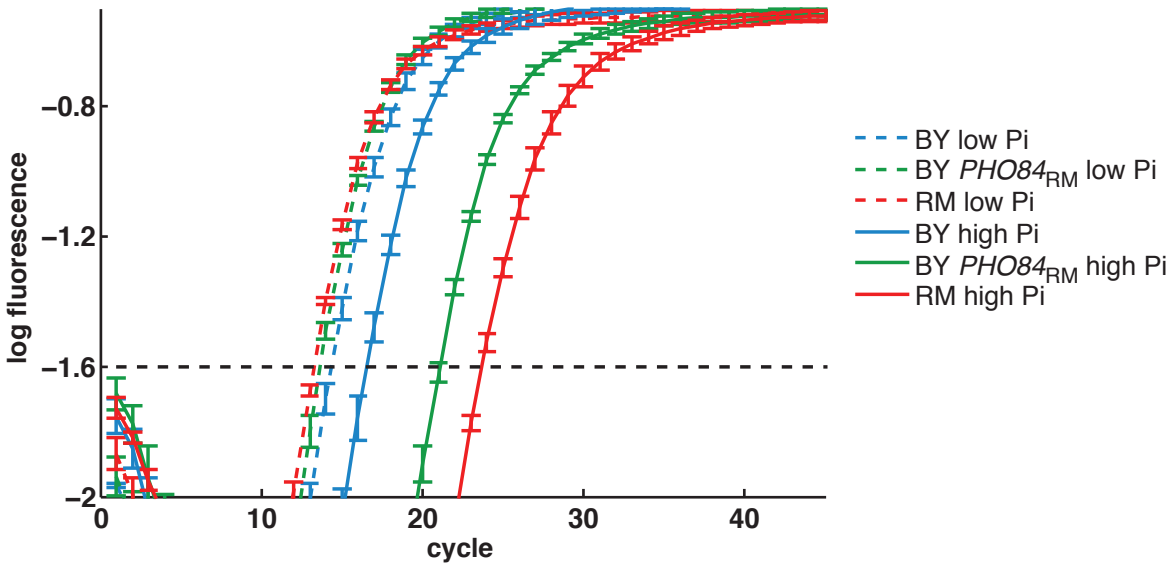
Supplementary Figure 5: Bootstrapping reduces the number of features. Histograms of the number of features selected in elastic net without bootstrapping (EN) compared to elastic net with bootstrapping (EN+BT). **A:** Number of features selected when only genotype data (**L**) are used in the feature pool (**X**). **B:** As **A**, but for models using genotype (**L**) and expression data (**E**) in the feature pool.



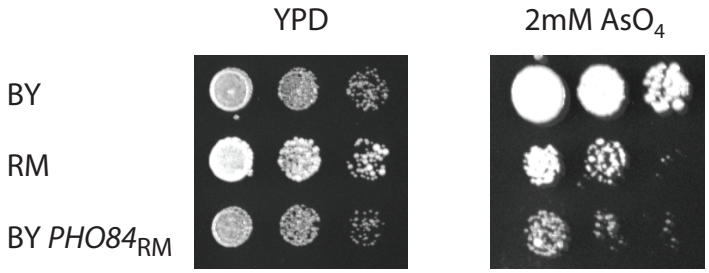
Supplementary Figure 6: Comparison of prediction – Pearson correlation. All predictions represented in this figure are based on held out test data. **A:** Camelot compared with linkage analysis. Each dot represents a condition (growth yield in the presence of a drug), showing the Pearson correlation coefficient between the original growth data and the prediction from Camelot (y-axis) plotted as a function of the correlation coefficient between the original growth data and the prediction from linkage analysis (x-axis). Dots above the diagonal indicate the superior performance of Camelot and are color coded to indicate the degree of improvement. **B:** As **A**, but the prediction by Camelot is compared with that of the elastic-net L model lacking transcript features.



Supplementary Figure 7: Comparison of prediction – Spearman correlation. Same as Supplementary Figure 6, but using Spearman correlation coefficients.



Supplementary Figure 8: Feedback regulation of *PHO84* is stronger in RM than BY. Abundance of *PHO84* for each strain in high and low phosphate media measured using RT-PCR (see Materials and Methods). *PHO84* is expressed at similar level in all three strains under SC+low phosphate conditions. The addition of phosphate results in repression of *PHO84* expression as expected; however, in the allele-swapped and RM strains, *PHO84* is repressed to a greater extent than in the BY strain. The abundance of the control gene (*ERV25*) is similar for all three strains in high and low phosphate media (data not shown).



Supplementary Figure 9: Growth in arsenate. Strains were grown overnight in YPD medium, diluted to OD_{600} ~0.2 and plated with 10-fold dilution on YPD (control) or YPD+2mM AsO_4 media (see Materials and Methods). The RM and allele-swapped (BY $PHO84_{RM}$) strains are more sensitive to arsenate, suggesting that the RM version of Pho84 transports phosphate more efficiently than the BY version of Pho84.