Personalized Medicine:

Studies of Pharmacogenomics in Yeast and Cancer

Bo-Juen Chen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

Columbia University
2013

ABSTRACT

Personalized Medicine: Studies of Pharmacogenomics in Yeast and Cancer

Bo-Juen Chen

Advances in microarray and sequencing technology enable the era of personalized medicine. With increasing availability of genomic assays, clinicians have started to utilize genetics and gene expression of patients to guide clinical care. Signatures of gene expression and genetic variation in genes have been associated with disease risks and response to clinical treatment. It is therefore not difficult to envision a future where each patient will have clinical care that is optimized based on his or her genetic background and genomic profiles.

However, many challenges exist towards the full realization of the potential personalized medicine. The human genome is complex and we have yet to gain a better understanding of how to associate genomic data with phenotype. First, the human genome is very complex: more than 50 million sequence variants and more than 20,000 genes have been reported. Many efforts have been devoted to genome-wide association studies (GWAS) in the last decade, associating common genetic variants with common complex traits and diseases. While many associations have been identified by genome-wide association studies, most of our phenotypic variation remains unexplained, both at the level of the variants involved and the underlying mechanism. Finally, interaction between genetics and environment presents additional layer of complexity governing phenotypic variation.

Currently, there is much research developing computational methods to help associate genomic features with phenotypic variation. Modeling techniques such as machine learning have been very useful in uncovering the intricate relationships between genomics and phenotype. Despite some early successes, the performance of most models is disappointing. Many models lack robustness and predictions do not replicate. In addition, many successful models work as a black box, giving good predictions of phenotypic variation but unable to reveal the underlying mechanism.

In this thesis I propose two methods addressing this challenge. First, I describe an algorithm that focuses on identifying causal genomic features of phenotype. My approach assumes genomic features predictive of phenotype are more likely to be causal. The algorithm builds models that not only accurately predict the traits, but also uncover molecular mechanisms that are responsible for these traits. . The algorithm gains its power by combining regularized linear regression, causality testing and Bayesian statistics. I demonstrate the application of the algorithm on a yeast dataset, where genotype and gene expression are used to predict drug sensitivity and elucidate the underlying mechanisms. The accuracy and robustness of the algorithm are both evaluated statistically and experimentally validated.

The second part of the thesis takes on a much more complicated system: cancer. The availability of genomic and drug sensitivity data of cancer cell lines has recently been made available. The challenge here is not only the increasing complexity of the system (e.g. size of genome), but also the fundamental differences between cancers and tissues.

Different cancers or tissues provide different *contexts* influencing regulatory networks and signaling pathways. In order to account for this, I propose a method to associate contextual genomic features with drug sensitivity. The algorithm is based on information theory, Bayesian statistics, and transfer learning. The algorithm demonstrates the importance of context specificity in predictive modeling of cancer pharmacogenomics.

The two complementary algorithms highlight the challenges faced in personalized medicine and the potential solutions. This thesis detailed the results and analysis that demonstrate the importance of causality and context specificity in predictive modeling of drug response, which will be crucial for us towards bringing personalized medicine in practice.

Table of Content

List of Figures

List of Tables

Acknowledgements

I am much indebted to the people in our lab and collaborators, for my research work and personal growth during the pursuit of my doctoral degree.

I have joined Dana Pe'er's lab since 2007. I am grateful for the opportunity of working with her and learning from her. Dana has taught me the principles of research in systems biology. She guided me through various research projects, helping me to learn, to innovate and to grow. It has always amazed me how much Dana has taught me through the years. From Dana, I have learned the beauty of Bayesian statistics, the power of advanced machine learning, the great insights of biological data, and the attitude of conducting solid and rigorous research. Dana has taught me to be a Bayesian, a system biologist, and a rigorous scientist. I owe Dana my skills, my work, and my research attitude. I would not have come so far without her guidance.

I am grateful for Lyle Ungar, who introduced me to information theory and transfer learning theory. My collaboration with Lyle has given me the chance to explore another field of machine learning, which I enjoy as much as biology. Lyle's patient guidance broadened my view of the field, opening a new chapter of research for me. I am deeply indebted to Lyle for everything he has taught me.

I cannot thank Helen Causton enough for her mentorship on biological bench work. Coming as a computer scientist, I was clueless in experimental design and execution. Helen patiently introduced me to various kinds of assays and techniques, explained the logic and reasons behind the design of each experiment, and offered me tremendous amount of help for my research and experiments. From her, I learned to appreciate the elegance of experimental designs and the efforts devoted to biological data collection.

I am also deeply indebted to my thesis committee: Frederick Cross, Itsik Pe'er, Raul Rabadan, and Saeed Tavazoie. I am grateful for their valuable input to my work and their guidance. They have helped me improve my research and my thesis work. I would not have finished this dissertation without them.

I would not have come this far without the support of my dearest friends, Oren Litvin, Han-Yu Chuang and Felix Sanchez-Garcia. I have enjoyed working with Oren in the last six and half years. We often exchange ideas for our research and have stimulating discussions. And most of all, Oren has supported me through the ups and downs in my life. He is always there to give me a hand, to encourage me, to give me advice. I will never be grateful enough for his wonderful friendship. Han-Yu and I have been friends since college and we are in each other's life since then. I am grateful for her support, her advice, and her friendship through the years. Felix has been a dear friend and a great coworker. I love discussing science and exchanging ideas with Felix. His warm friendship and passionate attitude toward research have been a blessing for me. I am truly thankful for his friendship and cherish the experience of working with him.

I owe a debt of gratitude to my collaborators and all the past and present members of Dana's lab. I have enjoyed working with every one of them and am thankful for their support and help for my research. I thank Noel Goddard and Ethan Perlstein for our collaboration on the yeast work. I am grateful for the input and friendship of Uri David Akavia, El-ad David Amir, Ambrose Carr, Jiyoung Kim, Dylan Kotliar, Jacob Levine, Denesy Mancenido, Rachel Melamed, Anna Starikov, and Michelle Tadmor. My gratitude also goes to Ambrose Carr and Jacob Levine for helping edit this thesis.

Finally, I am grateful for my parents' support through the years. They have encouraged and supported me to pursue my career and training. Their love and support are the foundation of who I am and what I have achieved.

To Dad, Mom, and Grandpa.

Chapter 1 Introduction

The emergence of high throughput molecular profiling technologies has revolutionized biological and clinical research. Array-based assays have been used to measure the expression of thousands of genes in a biological system, allowing glimpses of transcriptional regulation inside cells. Genetic variants such as single nucleotide polymorphisms (SNPs) or structural variants can now be profiled for cohorts of large populations (Altshuler et al., 2008; Balding, 2006; Frazer et al., 2009; Metzker, 2010), offering unprecedented resolution of genetic variation in populations. Analysis and interpretation of these high-throughput data have had groundbreaking influence on the field of genetics.

One prospect of these technologies is that of personalized medicine, treatments that are tailored for individuals based on genetic background and genomic expression profiles. Pioneer studies such as Alizadeh et al. (2000); Kutalik et al. (2008); van de Vijver et al. (2002) have demonstrated the possibilities of utilizing gene expression profiles to predict prognosis and clinical outcomes. Signatures of gene expression have been identified to associate with pathway activation in cancers (Segal et al., 2004). These studies demonstrate the predictive values of gene expression profiles to drug response (Bild et al., 2006) and raise hope for clinical applications.

Another approach is genome-wide association studies (GWAS), which have facilitated discoveries of association between genetic and phenotypic variation (Wellcome Trust Case Control, 2007). Based on the assumption that common variants may underlie common diseases, GWAS have demonstrated success for uncovering potential hereditary risk factors for diseases such as diabetes (Cox et al., 1999; Diabetes Genetics Initiative of Broad Institute of et al., 2007; Scott et al., 2007; Zeggini et al., 2007), Crohn's disease (Barrett et al., 2008; Fisher et al., 2008; Fransen et al., 2010; Mathew, 2008; Wang et al., 2009) and Parkinson's disease (Satake et al., 2009; Simón-Sánchez et al., 2009).

In addition to disease associations, genome-wide data imply the possibility of statistical pharmacogenomics (Wang et al., 2011). Using GWAS, several genetic variants have been associated with response to cardiovascular drugs and treatments of infectious diseases (Ge et al., 2009; Suppiah et al., 2009; Takeuchi et al., 2009; Tanaka et al., 2009). For example, using genome-wide association, Tanaka et al. (2009) identified genotypes of *VKORC1* (vitamin K epoxide reductase complex, subunit 1) and *CYP2C9* (cytochrome P450, family 2, subfamily C, polypeptide 9) that are associated with the effective dose of the anticoagulant warfarin.

Pharmacogenomic approaches are currently an active area of research in cancer therapeutics (Lee and McLeod, 2011; McLeod, 2013). In a recent study, targeted sequencing of 145 genes in colon and non-small cell lung cancers has identified at least one genetic variant associated with target therapy in 59% of samples (Lipson et al., 2012). Each of these genetic alterations suggests one option of treatments for these individuals.

These studies indicate that future clinical strategy may routinely involve evaluation of patient genomic profiles in order to optimize treatment. Maximized effectiveness and minimized adverse effects both imply an ability to predict patient response in advance. For example, a diagnostic score based on a dozen genes has been deployed to predict the response to adjuvant endocrine therapy for ER+/HER2+ breast cancer patients (Filipits et al., 2011). In colon cancer, testing of *KRAS* (v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog) mutation is used to predict the efficacy of EGFR (epithelial growth factor receptor) inhibitors (Amado et al., 2008; Karapetis et al., 2008). Testing for KRAS mutation is now recommended for colon cancer patients before treatment with EGFR inhibitors (Allegra et al., 2009; Dietel et al., 2013). Researches for devising such a predictive model have been under development (Chen et al., 2009; Chen et al., 2008; Huang et al., 2007; Jelier et al., 2011; Lee et al., 2008b; Schadt et al., 2005).

Despite the recent success in genomic medicine, we are still far from the scenario where predictive models give optimal, personalized treatments. GWAS results have so far explained only the tip of the iceberg for human phenotypic variation (Manolio et al., 2009; Purcell et al., 2009; Weedon and Frayling, 2008; Zuk et al., 2012). Moreover, associated genomic regions can contain a large number of genes. Identifying the exact causal genes for the phenotype is often difficult (Chen et al., 2009). Much of the complex interactions between signaling pathways and regulatory networks underlying the phenotypic variation remains unknown. In addition, many drugs may elicit a variety of responses beyond the intended mechanism of action. The complexity of biological systems can yield unanticipated responses, including drug resistance in cancer (Dietel et al., 2013; Rosenzweig, 2012). For instance, lung cancer cells harboring *MET* (met proto-oncogene) amplification have been shown to activate PI3K (Phosphatidylinositide 3-kinases) via an alternative route, despite these cells are treated with EGFR inhibitors. The alternative route of signaling acts via ERBB3 (v-erb-b2 erythroblastic leukemia viral oncogene homolog 3) and confers resistance to EGFR inhibitors such as erlotinib (Bean et al., 2007; Engelman et al., 2007).

The complication of drug sensitivity or phenotypic variation aggregates when we consider “context.” The effectiveness of targeted therapy in cancers often depends on tissue type, cancer type, and mutation status (Bollag et al., 2010; Dietel et al., 2013; Heiser et al., 2012). In other words, the complex interactions of cellular signaling are modulated by contextual factors such as tissue type and microenvironment (Bissell and Labarge, 2005) often resulting in altered response to environmental cues and drugs. In yeast, it has also been shown that genetic variation may create context and dictate different regulatory programs (Litvin et al., 2009). This additional complexity due to the interaction between genetics and context poses a great challenge for biological research and for the realization of personalized medicine.

In this thesis, we propose two algorithms to tackle these challenges. In the first, we use a yeast dataset to demonstrate how the rich information from gene expression can be leveraged to improve predictions of drug response. The critical insight is that, although the gene expression profiles are measured before any treatment, they encode information of the regulatory network that governs mechanisms of subsequent drug response. Principled by rigorous statistical tests and Bayesian statistics, the algorithm Camelot (CAusal Modelling with Expression Linkage for cOmplex Traits) aims to uncover the genes that are causal and therefore deterministic of the phenotypic variation. Anchored on the interplay between causality and predictability, Camelot builds models from genotype and gene expression profiles and is able to accurately predict the response to 87/94 drugs (Chen et al., 2009).

Camelot's causal modeling further elucidates how gene expression helps uncover the mechanisms underlying complex traits. We show that the expression level of a transcript may reveal complicated regulatory mechanisms such as feedback loops, or encode genetic effects from multiple loci.

In the second part of the thesis, we tackle the analogous problem in the more complicated system of cancer cell lines. We propose a novel algorithm to consider context specificity between cancer types for predictive modeling of drug sensitivity. The novelty of the algorithm is to transfer knowledge between drugs and tissue types, and therefore enhance our capability to identify both predictive features that are shared across cancers and that are specific to individual cancer types. We demonstrate the importance of context in predictive modeling and the complexity of drug response in cancer. Our results reveal that the activities of several pathways are associated with drug sensitivity, elucidating potential mechanisms of drug resistance.

We believe both studies will shed light onto mechanisms underlying pharmacogenomics. Currently, most treatment choices in genomic medicine are made by the profiles of one or two genes. Our predictive models suggest the complexity of drug response may be associated with many more interactions between pathways. We hope our results will contribute to the field, brining closer the realization of personalized medicine in practice.

Chapter 2 Genetics, genomics and phenotype

2.1 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; Kutalik et al., 2008; van't Veer et al., 2002). 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) have recently been used to show that gene expression can be used to associate genes with disease phenotypes (Chen et al., 2008; Emilsson et al., 2008; Mehrabian et al., 2005; Schadt et al., 2005), 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, i.e., 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, both generated in drug free conditions, to detect causal genes and predict 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 analyze 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 in 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.

2.2 Datasets

We used a dataset containing genotype and gene expression information from 104 segregants that arose from the mating of two genetically diverse strains, 'BY' and 'RM' (Brem and Kruglyak, 2005). The gene expression profiles of 6189 genes were measured in rich media for each segregant (Brem and Kruglyak, 2005). The phenotype, growth yield from each segregant grown in the presence of one of 94 chemicals ('drugs') is from Perlstein et al. (2007). In total 313 growth conditions (different concentrations of chemicals and time points) are considered as phenotype.

For the genotype, 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 zero and variance of one. We compiled a list of candidate gene expression features based on two sources. One contained genes with potential regulatory effects, including transcription factors, signaling molecules, chromatin factors and RNA factors as in Lee et al. (2009). 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 SD ≤ 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.

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 (Figure 2-1).

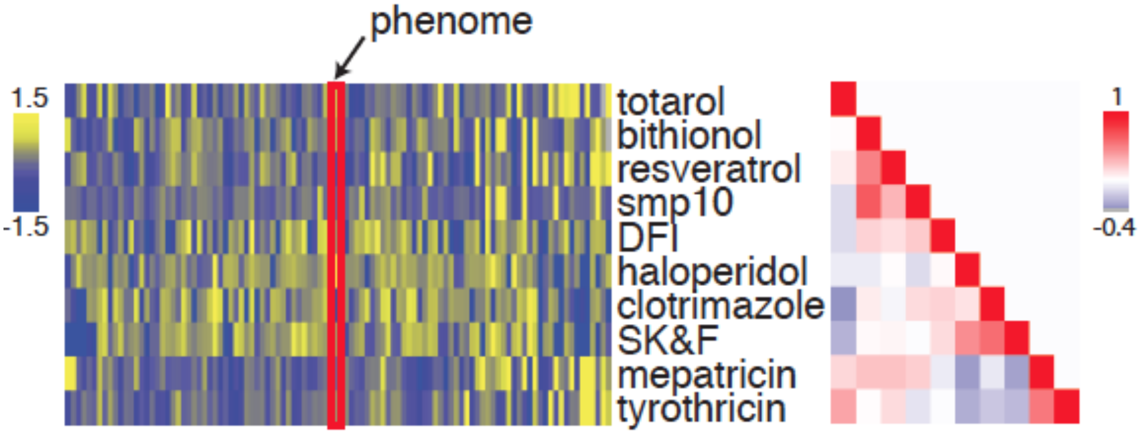

Figure 2-1: Diversity of drug response.
Growth in the presence of a subset of drugs is represented by the heat map on the left (blue corresponds to low growth yield and yellow to high growth yield). Each row represents the data for a single drug (smp10 is 1,9-pyrazoloanthone, DFI is diphenyliodonium 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 PCC 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 anti-correlation (-0.40). The same scale is used for all Figures.

Our goal is to take baseline information about a strain: genotype and gene expression data measured from each segregant grown in the absence of drug and to 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 to be 'causal' if genetic manipulation of this feature, e.g., by allele swap or gene deletion, changes the phenotype, as predicted by the model.

Chapter 3 Camelot Algorithm

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. While the true relationship may not be linear, we use regularized linear models as these can be robustly inferred from the data (Hastie et al., 2001). That is, Camelot selects a sparse set of features, markers $\{L\}$ and transcripts $\{E\}$ to build a linear regression model $D \sim \{L\} + \{E\}$.

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) (Section 3.1.1), non-parametric bootstrap (Efron, 1979) (Section 3.1.2) and tests designed to further select only those that are most likely causal (Section 3.2). The selected features are then used to optimize a linear prediction function (Section 3.3). Overview of Camelot algorithm is shown in **Error! Reference source not found.**

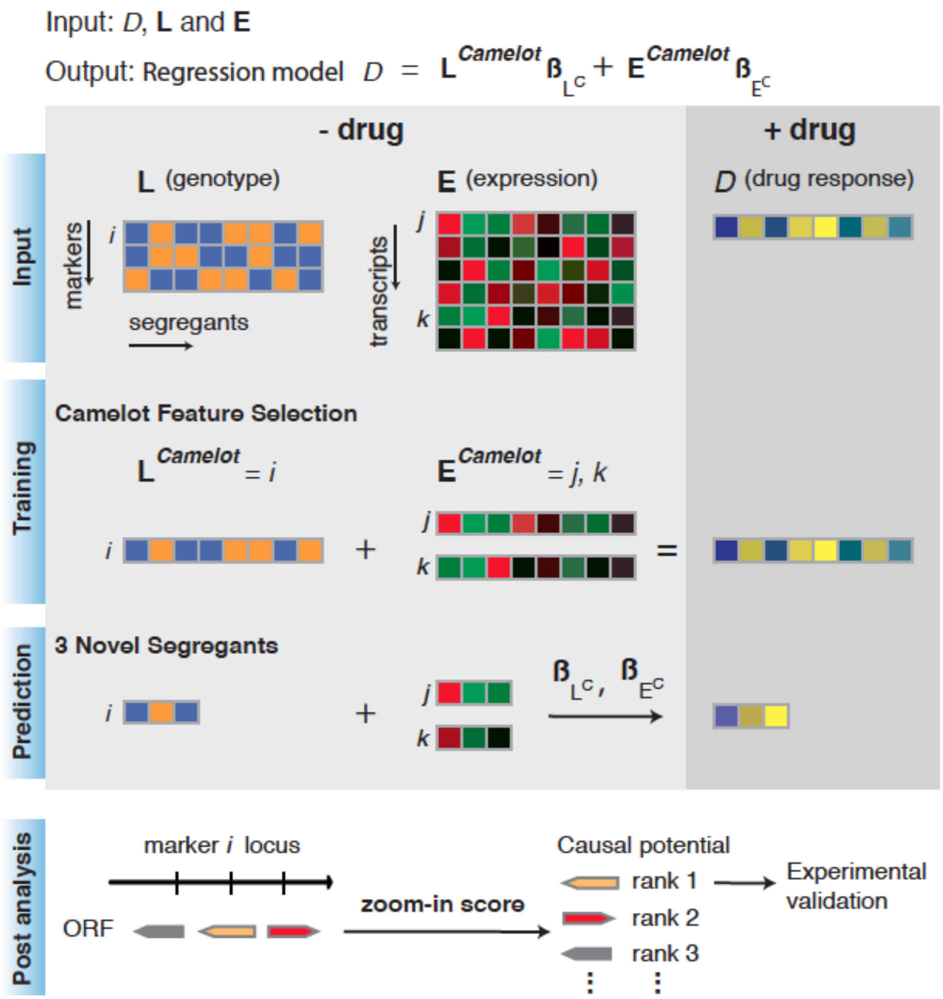

Figure 3-1: Overview of Camelot.

The input data includes 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 $\mathrm{L}^{Camelot}$ and $\mathrm{E}^{Camelot}$, representing selected markers and transcripts, respectively. A linear regression model is then built on $\mathrm{L}^{Camelot}$ and $\mathrm{E}^{Camelot}$. In the prediction stage, Camelot uses the model built on the training data (regression coefficients β_{L^C} and β_{E^C}) 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 ($\mathrm{L}^{Camelot}$) and uses the zoom-in score to prioritize the likelihood that each gene within the linked region is causal.

3.1 Feature Selection

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 Materials and Methods). Camelot selects features using a bootstrap procedure on coefficients of regularized regression (see Methods).

3.1.1 Regularized linear regression

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:

$$D = \boldsymbol{X}^{*}\beta^{*} \qquad \textbf{Eq. 3.1}$$

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

However, here we want to identify features ($\boldsymbol{X}^{*}$) 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 X^*. In brief, elastic net regression solves the following optimization problem:

$$\hat{\beta} = \operatorname{argmin}_{\beta}|y - \boldsymbol{X}\beta|^2 \qquad \textbf{Eq. 3.2}$$

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

where y represents response (*D*), $\boldsymbol{X}$ is the feature matrix (containing marker ***L*** and/or transcript ***E***). Both y and X are standardized to mean of 0 and variance of 1. β (from here we use β 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 β, 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 (MathWorks), including a modified implementation (Sjöstrand, 2005) of elastic net. Parameters of elastic net were chosen using standard cross-validation techniques.

3.1.2 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 and Hastie, 2005) to generate the full path of solutions for 20 conditions of growth data (randomly chosen from the original data), using a feature pool $\boldsymbol{X}$**,** containing both $\boldsymbol{L}$ and $\boldsymbol{E}$. 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. 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.

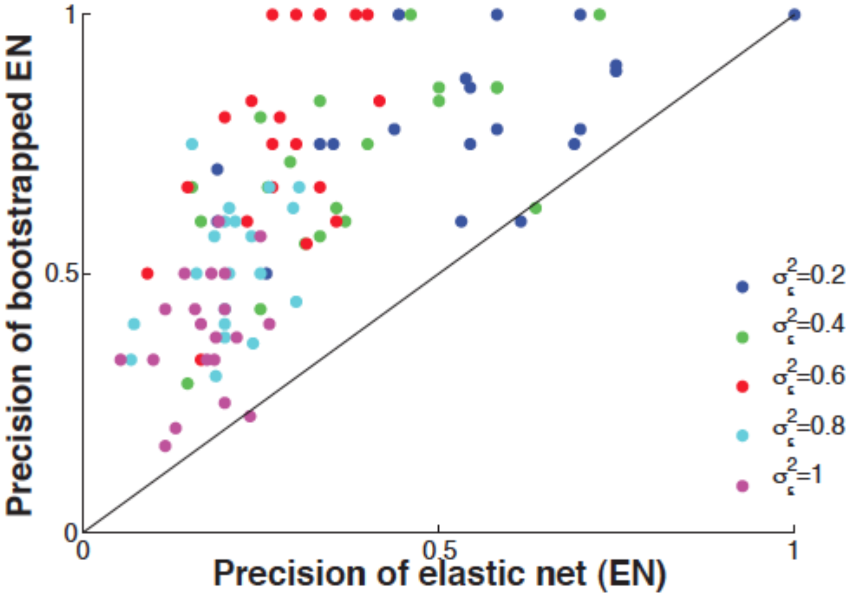

Figure 3-2: Bootstrap improves precision in retrieving true factors.
Bootstrapping improves the precision with which true genetic factors can be retrieved from a synthetic dataset. Precision obtained using elastic net with bootstrapping (y-axis) is compared with that from elastic net without bootstrapping (x-axis). Each dot represents a synthetic growth condition. The diagonal line shows where the two models have the same precision. Different colored dots represent different levels of noise added to the synthetic data. 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).

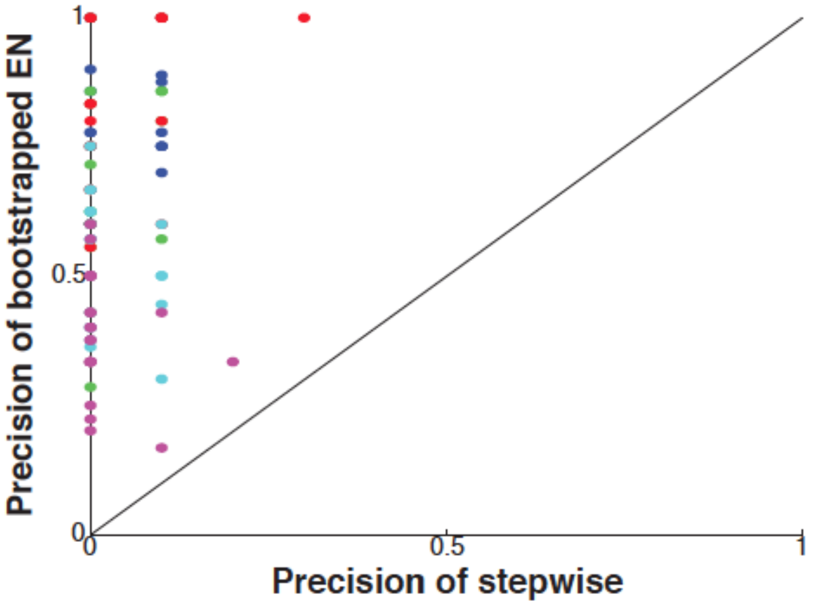

Figure 3-3: bootstrapped elastic net versus stepwise regression.
Bootstrapped elastic net retrieves true factors more precisely than stepwise regression. Similar to Figure 3-2, precision with which factors are retrieved using elastic net with bootstrapping (y-axis) is compared with that from stepwise 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.

Figure 3-2 shows bootstrapped elastic net has higher precision than elastic net without bootstrapping. In addition, we compare bootstrapped elastic net to stepwise regression, which is commonly used as a feature selection algorithm for regression problem (used in Chen et al. (2008); Emilsson et al. (2008); Mehrabian et al. (2005); Schadt et al. (2005)). Figure 3-3 shows the precision retrieved for the synthetic datasets from both methods. 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. Figure 3-4 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.

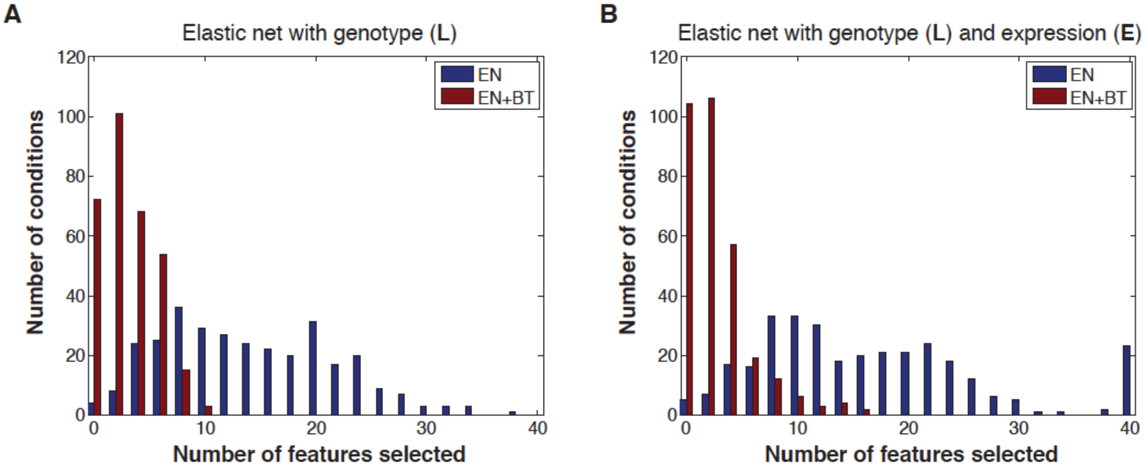

Figure 3-4: Bootstrapping reduces the number of features.
Histograms of the number of features selected in elastic net without bootstrapping (EN) compared to elastic net with bootstrapping (EN+BT). A: Number of features selected when only genotype data (L) are used in the feature pool (X). B: As A, but for models using genotype (L) and expression data (E) in the feature pool.

3.2 Resolving Causality

Although bootstrapped elastic net helps pruning spurious association between genetic factors and the phenotype, the causal roles of the selected predictors are not guaranteed.

The models proposed in the previous section involve two types of features: genotype and gene expression. For genotype, its causal role is straightforward and therefore when a genotype feature is chosen as a predictor, we can assume the causal factor resides in the DNA region. However, the causal role of a gene expression predictor is less clear, since its association with the phenotype could be due to the influence of its regulating genotype, such as mutation. Identification of the causal genes provides insight into the biological processes and stresses involved in the response to a drug, and has practical implications for identifying alternative drug targets in resistant strains.

Fortunately, the use of gene expression provides clues for us to perform statistical tests to dissect the potential causal roles of selected features. In this section, I will present two methods for uncovering the causal genetic factors of the phenotype.

3.2.1 Triangle Test

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 3-5A), (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 3-5B), 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.

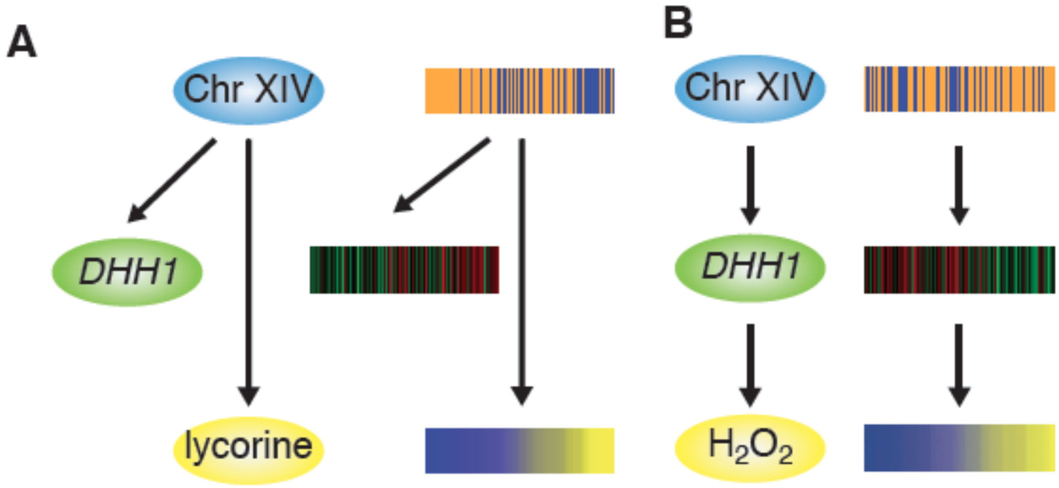

Figure 3-5: Correlation versus causality.
A. Sequence variation at the Chromosome XIV locus is the cause of variation in both *DHH1* expression and the response to lycorine; however, in this situation *DHH1* expression does not causally influence the drug response. *DHH1* was a candidate feature for 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. B. Example of a causal chain where polymorphisms in 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}. Some notation for all 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).

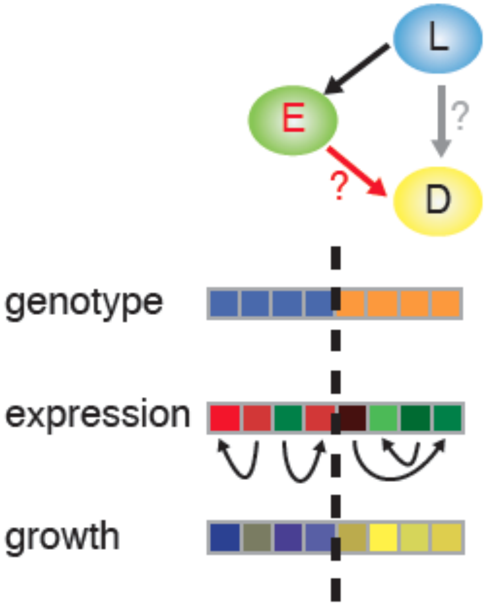

Figure 3-6: Determine the causal role of gene expression features.
The triangle test evaluates the likelihood that each transcript feature causally explains the phenotype (red edge). It distinguishes between causal chain (left, red edge) and co-regulation structures (right, grey edge) using permutation 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. Orange represents RM and blue BY for the genotype at the locus.

In order to determine if a gene expression predictor is causal, 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 3-6) We assume 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 transcripts features that are more likely causal and enriches the final selection with causal features.

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, $E_g \in \mathbf{E}^{\mathcal{L}\mathcal{E}}$). 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

$$E_g{\sim}\boldsymbol{L}_g^{e\mathcal{L}}, \tag*{\textbf{Eq. 3.3}}$$

$$\boldsymbol{L}_g^{eL} = \{L_m | \gamma_m^{e\mathcal{L}(g)} \geq 0.5, m\text{:index of marker}\},$$

$$\gamma_m^{e\mathcal{L}(g)} = \sum_B^{N_B} I\left(\beta_m^{B,e\mathcal{L}(g)} \neq 0\right) \Big/ N_B$$

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:

$$\boldsymbol{E}^{Cause} = \left\{ E_g \middle| \begin{matrix} E_g \in \boldsymbol{E}^{\mathcal{L}\mathcal{E}} \text{ and } E_g \text{ passes test for at least} \\ \text{one set of linked neighboring markers} \end{matrix} \right\} \tag*{Eq. 3.4}$$

$$\boldsymbol{L}^{Indirect} = \left\{ L_m \middle| \begin{matrix} L_m \in \boldsymbol{L}_g^{e\mathcal{L}} \text{and } E_g \text{ passes} \\ \text{test for the triplet } (L_m, E_g, D) \end{matrix} \right\} \tag*{Eq. 3.5}$$

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 $\mathbf{E}^{Cor}$:

$$\boldsymbol{E}^{\boldsymbol{Cor}} = \{E_g | E_g \in \boldsymbol{E}^{L\mathcal{E}} \text{ and } E_g \text{ fails test for all involved triplets}\} \tag*{Eq. 3.6}$$

$$\boldsymbol{L}^{\boldsymbol{Cause}} = \{L_m | L_m \in \boldsymbol{L}_g^{eL} \text{ and } E_g \text{ fails test for } the\ triplet(L_m, E_g, D)\} \tag*{Eq. 3.7}$$

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 in Chen et al. (2009).

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 in Chen et al. (2009)). 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.

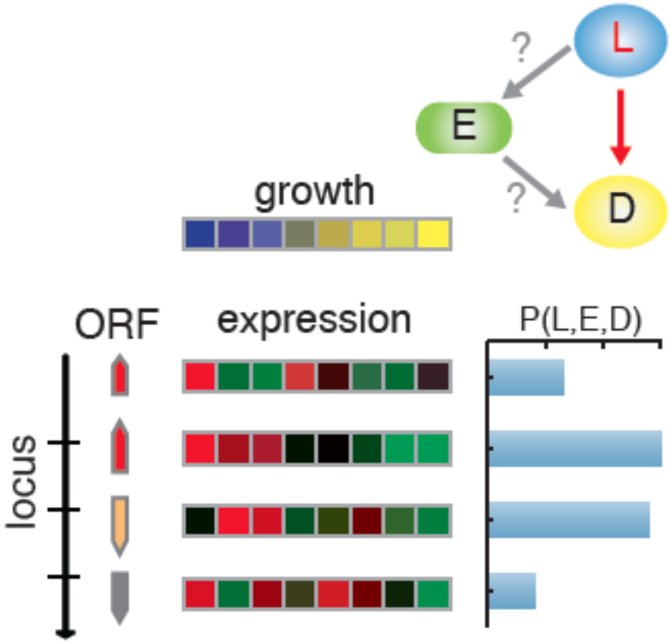

Figure 3-7: Identify causal gene in a genomic region.
The zoom-in score for each locus to drug association (red arrow), evaluates the likelihood, P(L,E,D), that each gene in the locus causally influences the growth of the strain in response to drug, based on its expression (each row represents a gene at the locus and each column a segregant; each row represents one gene).

3.2.2 Zoom-in Score

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 (Figure 3-7).

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:

$$P(L_g, E_g, D) = P\left(D \middle| L_g, E_g\right) P\left(E_g \middle| L_g\right) P(L_g) \qquad \textbf{Eq. 3.8}$$

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 3-7).

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:

$$s_j = \exp\left(-similarity\left(R_{j,BY}, R_{j,RM}\right) - \min\left(D_{j,BY}, D_{j,RM}\right)\right) \qquad \textbf{Eq. 3.9}$$

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

Eq. 3.10

$$P(L_g) = sigmoid(\theta \sum_j s_j)$$

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.

3.3 Build Predictive Model

As we believe the causal factors should provide better predicting power for phenotype than non-causal factors, we revise our final feature set based on the results of the triangle test.

For a phenotype *D,* our goal is to obtain the following model

$$D = \boldsymbol{L}^{Camelot}\ \beta_{L^C} + \boldsymbol{E}^{Camelot}\ \beta_{E^C} \qquad \textbf{Eq. 3.11}$$

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 β_{L^C} and β_{E^C} are the associated coefficients.

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

$$\boldsymbol{L}^{Camelot} = (\boldsymbol{L}^{\mathcal{L}} \cup\ \boldsymbol{L}^{\mathcal{LE}} \cup \boldsymbol{L}^{Cause}) \backslash \boldsymbol{L}^{Indirect} \qquad \textbf{Eq. 3.12}$$

$$\boldsymbol{E}^{Camelot} = \boldsymbol{E}^{Cause} \qquad \textbf{Eq. 3.13}$$

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 $\gamma_m^{\mathcal{L}}$ 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).

Chapter 4 Results From BYxRM Data

We applied Camelot to the previous describe yeast dataset ((Brem and Kruglyak, 2005), (Perlstein et al., 2007)), processed as described in Section 2.2. In this Chapter, we will first evaluate the performance of Camlot's prediction (Section 4.1). Then we will discuss the biological findings (Section 4.4 and 4.5). In addition, we explain why gene expression encodes such rich information for dissection the relationship between genotype and phenotype (Section 4.5).

4.1 Cross-Validation analysis demonstrates the robustness of Camelot

Camelot, the elastic net *L* model and linkage analysis are evaluated with ten-fold cross-validation: the data are randomly split 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, elastic net, or linkage analysis, to obtain a linear regression model (Section 4.7.4). 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.

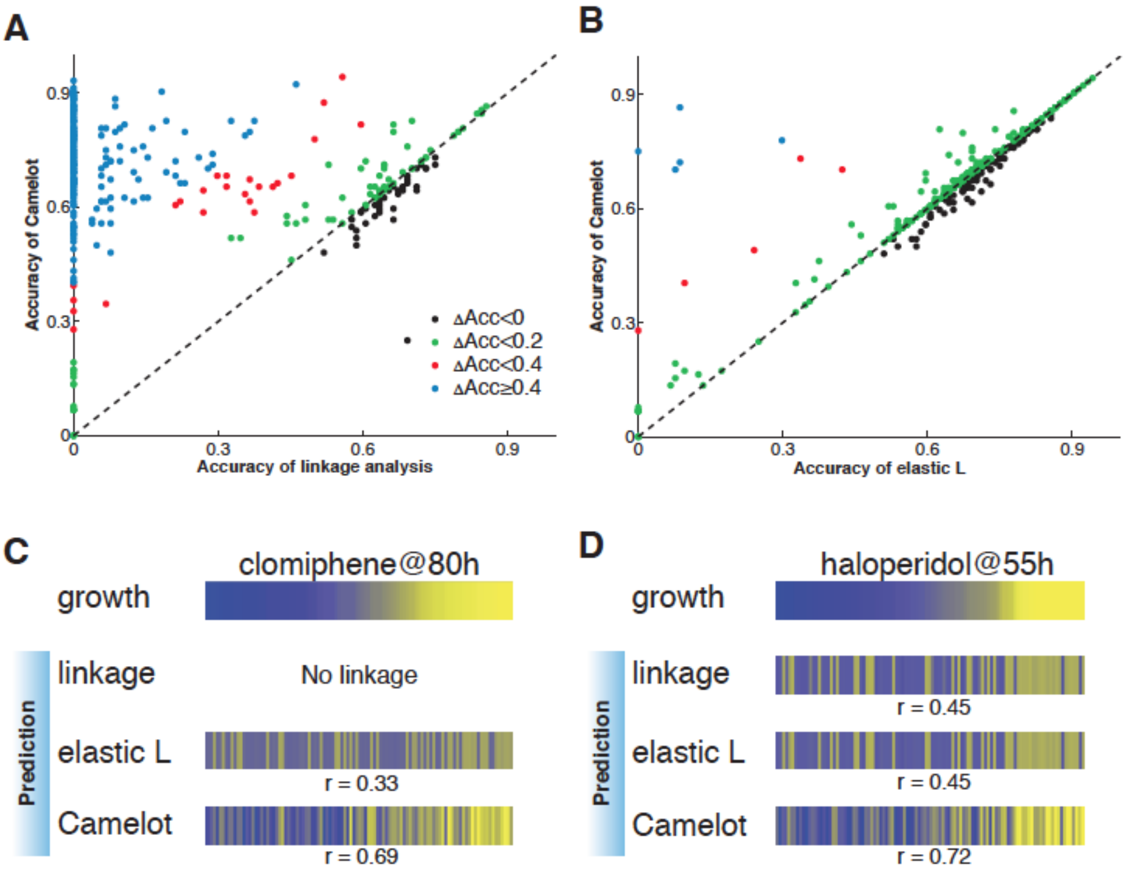

Figure 4-1: Camelot has superior predictive ability.

Comparison of prediction methods on held out test data from different models. A. Classification accuracy (See 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 A, but the classification accuracy by Camelot is compared with that of the elastic-net L model lacking transcript features (See 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 Figures. D. As C, but for haloperidol.

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%; $p<5.6\times10^{-5}$) (Perlstein et al., 2007). Linear regression models are built on significant linkages using robust regression (robustfit function in Matlab).

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's (*r*) and Spearman's (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).

In addition, 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 (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. Predictions of responses to drugs were made based on the predicted values from regression models. Accuracy (*Acc*) is defined as the number of correct predictions divided by the total number of segregants tested.

Camelot outperforms association and linkage analysis in providing a set of features that yield significantly more accurate prediction of drug response (see Figure 4-1, Figure 4-2 and Figure 4-3). 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 4-1A) and in many cases led to a dramatic improvement in the accuracy of prediction, *e.g.,* for clomiphene and haloperidol (Figure 4-1C and D).

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

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 4-1B-D). Note that the gene expression data was obtained from cells grown in nutrient-rich, non-perturbed conditions, whereas the growth data were measured in the presence of different drugs and that the expression features chosen differ between 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 counterintuitive (Figure 4-1B). 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).

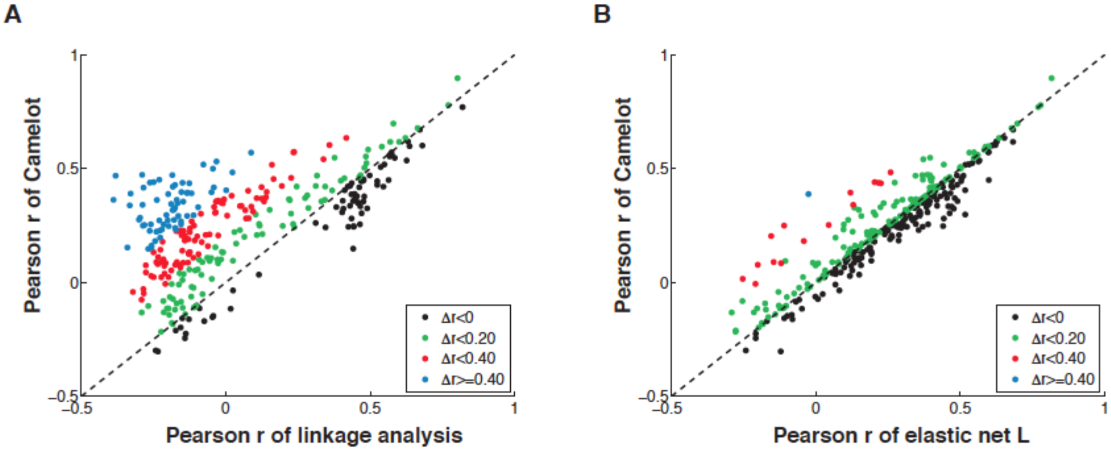

Figure 4-2: Comparison of prediction – Pearson correlation.
All predictions represented in this figure are based on held out test data. A: Camelot compared with linkage analysis. Each dot represents a condition (growth yield in the presence of a drug), showing the Pearson correlation coefficient between 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.

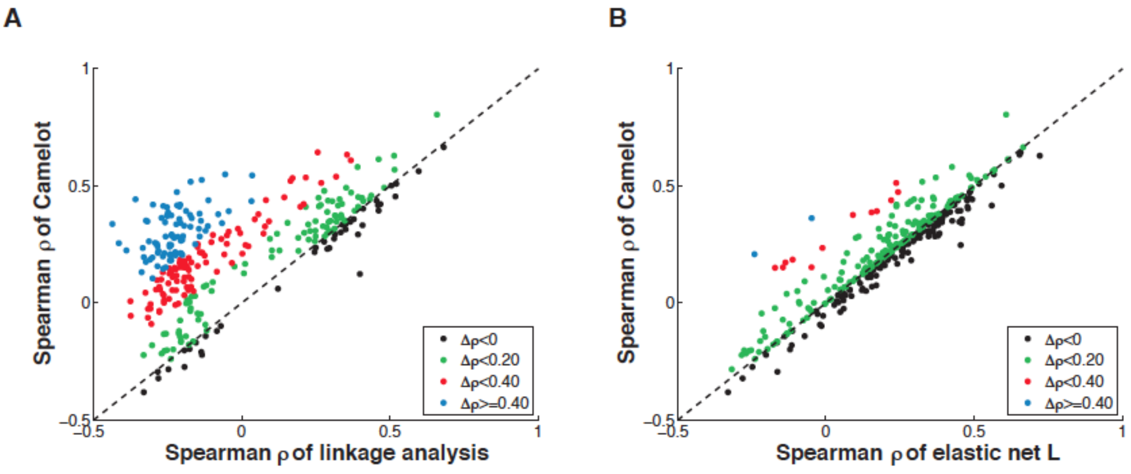

Figure 4-3: Comparison of prediction – Spearman correlation.
Same as Figure 4-2, but Spearman correlation coefficients are used to evaluate the models.

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.

4.2 Identifying features that actively influence the phenotype

As described in Chapter 3, Camelot aims to find a model that is not only predictive, but that also identifies genes that are responsible for the phenotypic variation. Identification of these genes provides insight into the biological processes and stresses involved in the response to a drug, and has practical implications for identifying alternative drug targets in resistant strains.

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, two possibilities exist (see Section 3.2.1). In order to test the causality, we apply the permutation-based *triangle test* (Section 3.2.1) to all transcript features chosen with significant confidence after bootstrapping.

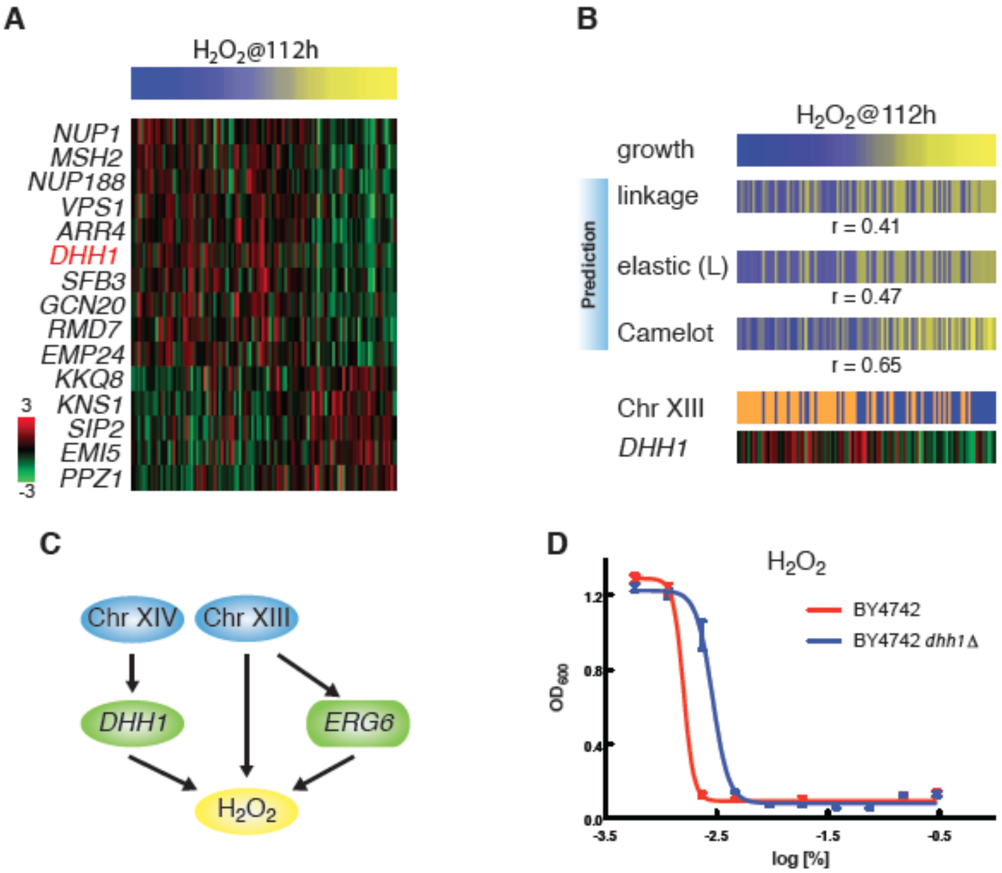

Figure 4-4: Causal role of *DHH1* in response to hydrogen peroxide.
A. Growth yield in the presence of hydrogen peroxide and correlated expression profiles for genes in the candidate pool (absolute Pearson correlation coefficients≥0.36, $p<2\times10^{-4}$), showing that the expression of multiple genes correlates with growth. Each column is associated with a different segregant (matched horizontal position across panel) sorted by growth yield (as in Figure 4-1) and gene expression on a red-green scale. GO-based filtering reduced the list of 123 candidate transcripts to the 15 genes are shown here. B. 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 4-1, demonstrating superior prediction by Camelot. Camelot chose a Chromosome XIII locus (227254-243624) 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. C. 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 over-expression of *ERG6* leads to decreased resistance to hydrogen peroxide (Khoury et al., 2008). The Chromosome XIV locus is at position 449639-486861. Some notation for all Figures: green rectangles (such as *ERG6*) represent expression of a gene within a linked region. D. Averaged OD_{600} absorbance growth measurements of BY (red) and BY *dhh1Δ* 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*.

4.2.1 From Prediction to Mechanism: DHH1's causal role in response to hydrogen peroxide as an 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 4-4A): 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. Second, Camelot selects features using a bootstrap procedure on coefficients of regularized regression (Section 3.1). Bootstrapping further reduced the list of expression features to a single gene, *DHH1*. In the next stage, Camelot uses the *triangle test* to test the causal role of *DHH1* in responses to drugs.

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

4-4). The *dhh1*Δ strain grew better than the wild type, confirming that *DHH1* negatively influences the phenotype.

This result complements the finding that Dhh1 colocalises 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 up-regulated 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).

Table 4-1: *DHH1* predictions and validation results

Prediction	Control
totarol	hinokitiol
H2O2	lycorine
valinomycin	tamoxifen
trifluoperazine	hexylresorcinol
dequalinium Cl	
benzethonium Cl	

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, one false negative in red.

4.2.2 Testing the causal role of transcript features

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 6 of these drugs. These were subsequently validated experimentally (Table 4-1, Figure 4-4 and Figure 4-5). 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.

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 (Table 4-1 and Figure 4-5). 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 (Cardoso et al., 2001; Tuquet et al., 2000). This demonstrates the stringency of our approach which is designed to minimise 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 (Evans et al., 2000; Lee et al., 2005; Lee et al., 2008a; Nicolson et al., 1999; Nulton-Persson and Szweda, 2001; Safiulina et al., 2006; Sancho et al., 2007; Yip et al., 2006). 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 (Friedman, 2002; Hasumi et al., 1992; Vance et al., 1972). 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 stabilise or destabilise 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.

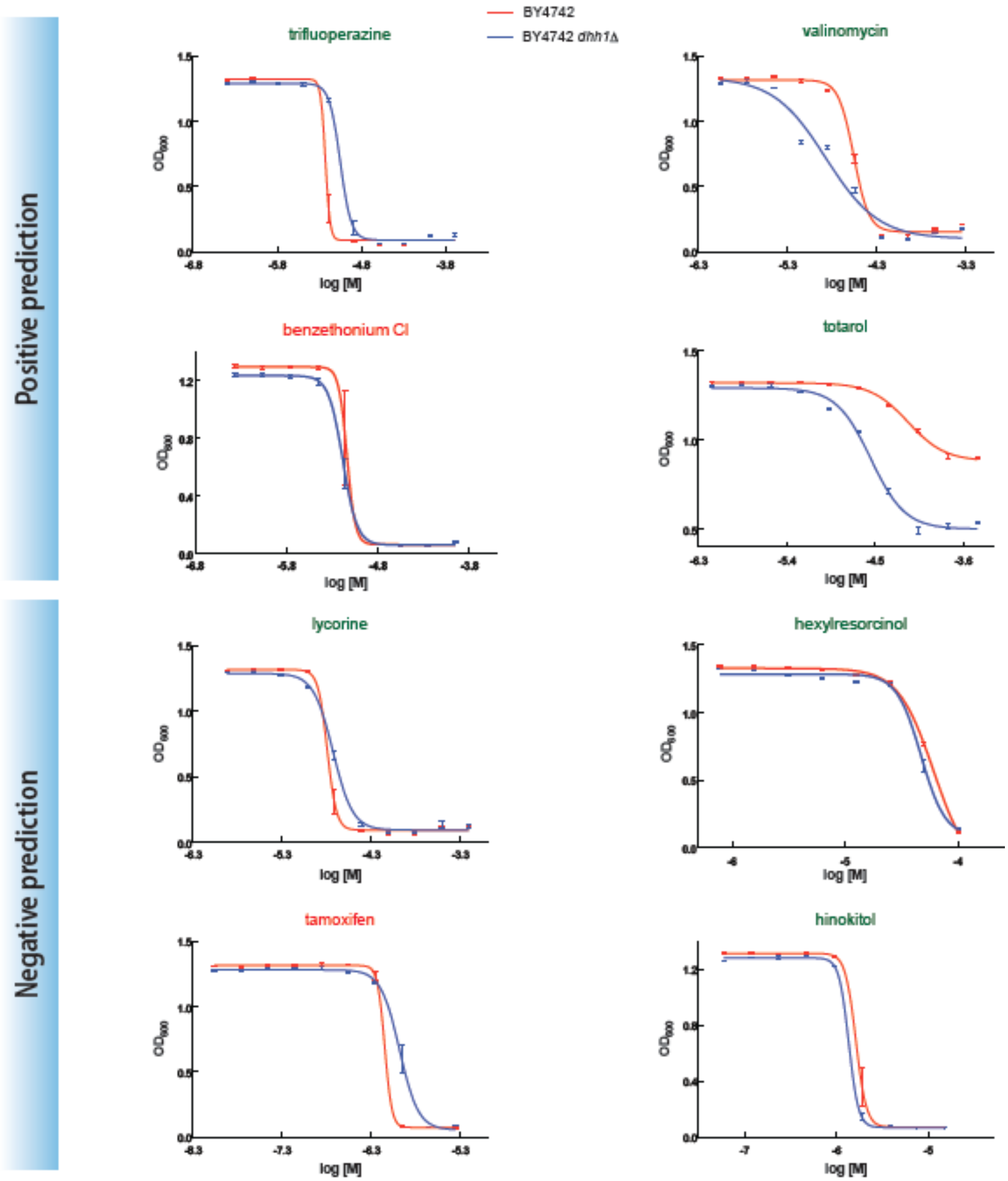

Figure 4-5: Growth yield of BY and BY dhh1Δ strains in the presence of drugs. Averaged OD600 absorbance growth measurements of BY (red) and BY dhh1Δ mutant (blue) are plotted against a twofold dilution series for each drug. The causal effect of DHH1 was confirmed for our positive predictions, with the 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.

4.3 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* (Section 3.2.2) uses gene expression to prioritise 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 prioritise genes within a locus (Figure 3-7).

Camelot chose two features for hydrogen peroxide, the *DHH1* transcript and a region on chromosome XIII (locus 227254-243624) containing 88 genes (Figure 4-4C). The zoom-in score identified *ERG6* as the causal gene within this region, i.e., polymorphism in the *ERG6* sequence between the BY and RM strains is likely to be responsible for the differences in the response to hydrogen peroxide between the parent strains. Over-expression of *ERG6* leads to decreased resistance to hydrogen peroxide (Khoury et al., 2008), matching Camelot's prediction. These results demonstrate how the triangle test and zoom-in score combine to provide a better understanding of the cellular response to each drug.

Table 4-2: ***PHO84*** **predictions and validation results**

Prediction	Control
doxorubicin	ebselen
hexylresorcinol	dichlorophene
haloperidol	tunicamycin
furoxan	
pentachlorophenol	
TCPN	

To the left are high scoring drugs predicted to be influenced by *PHO84*, all 6 drug were validated including two previously associated with *PHO84* (Perlstein et al., 2007) (light green), and 4 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.

Similar to linkage analysis (Perlstein et al., 2007), Camelot identified the two largest marker hotspots, a region on Chromosome XIII (locus 27644-33681), linked to 25 drugs and a region on Chromosome XIV (linked to 12 drugs). While 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 Chomosome 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 strains (BY strain containing *PHO84* with one amino acid substitution, L259P, from the RM strain) individually in the presence of one of 9 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 (Table 4-2, Figure 4-6). We performed a similar analysis for the drugs linking to the Chromosome XIV region and identified 3 drugs likely to respond to *MKT1* and 3 linked drugs that are unlikely to be affected by *MKT1*. Again Camelot correctly predicted the response to the drugs in an allele swapped (BY *MKT1*-RM) strain 6/6 times (Figure 4-7). 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.

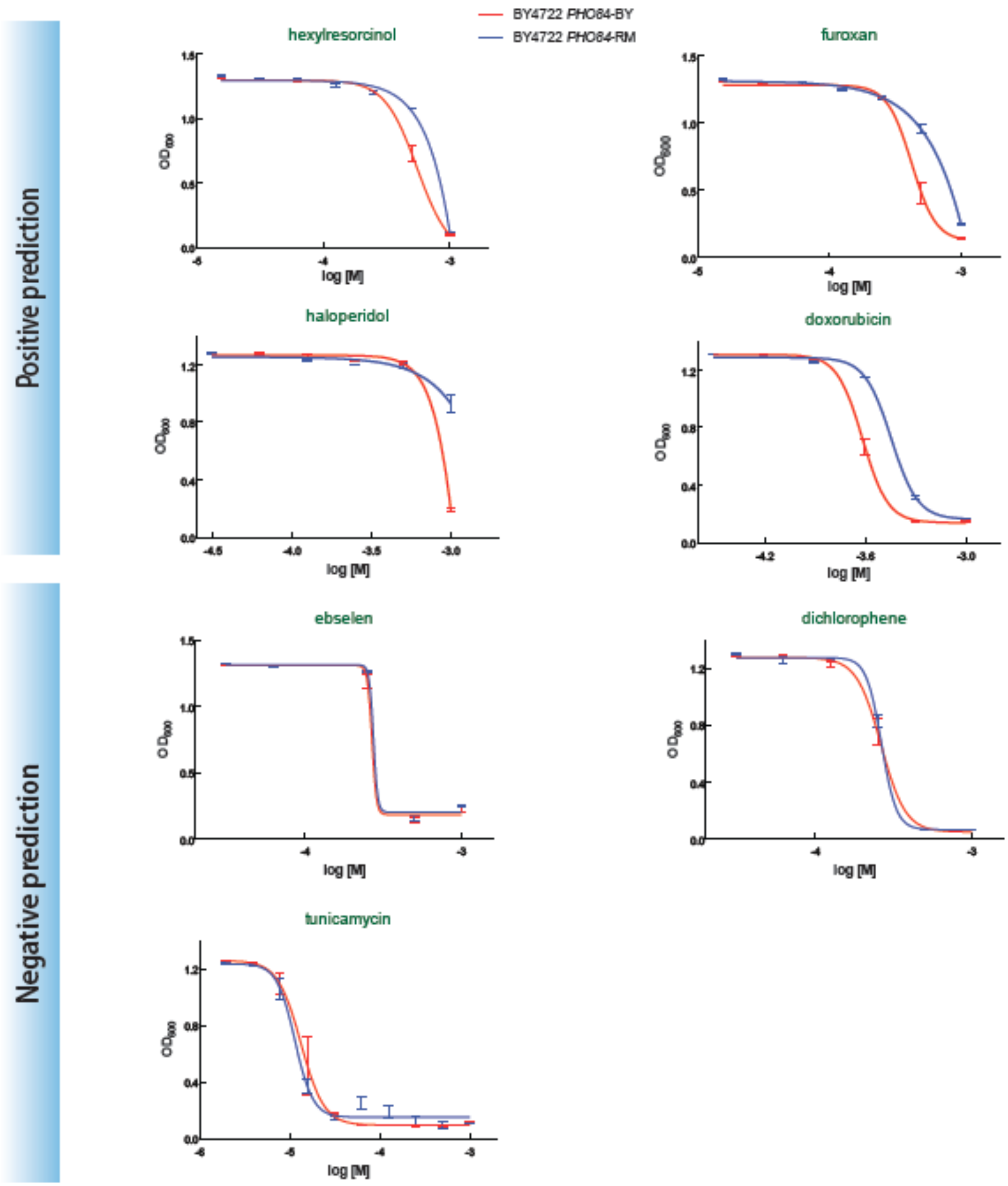

Figure 4-6: Validation of Camelot's PHO84 predictions.

Averaged OD600 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.

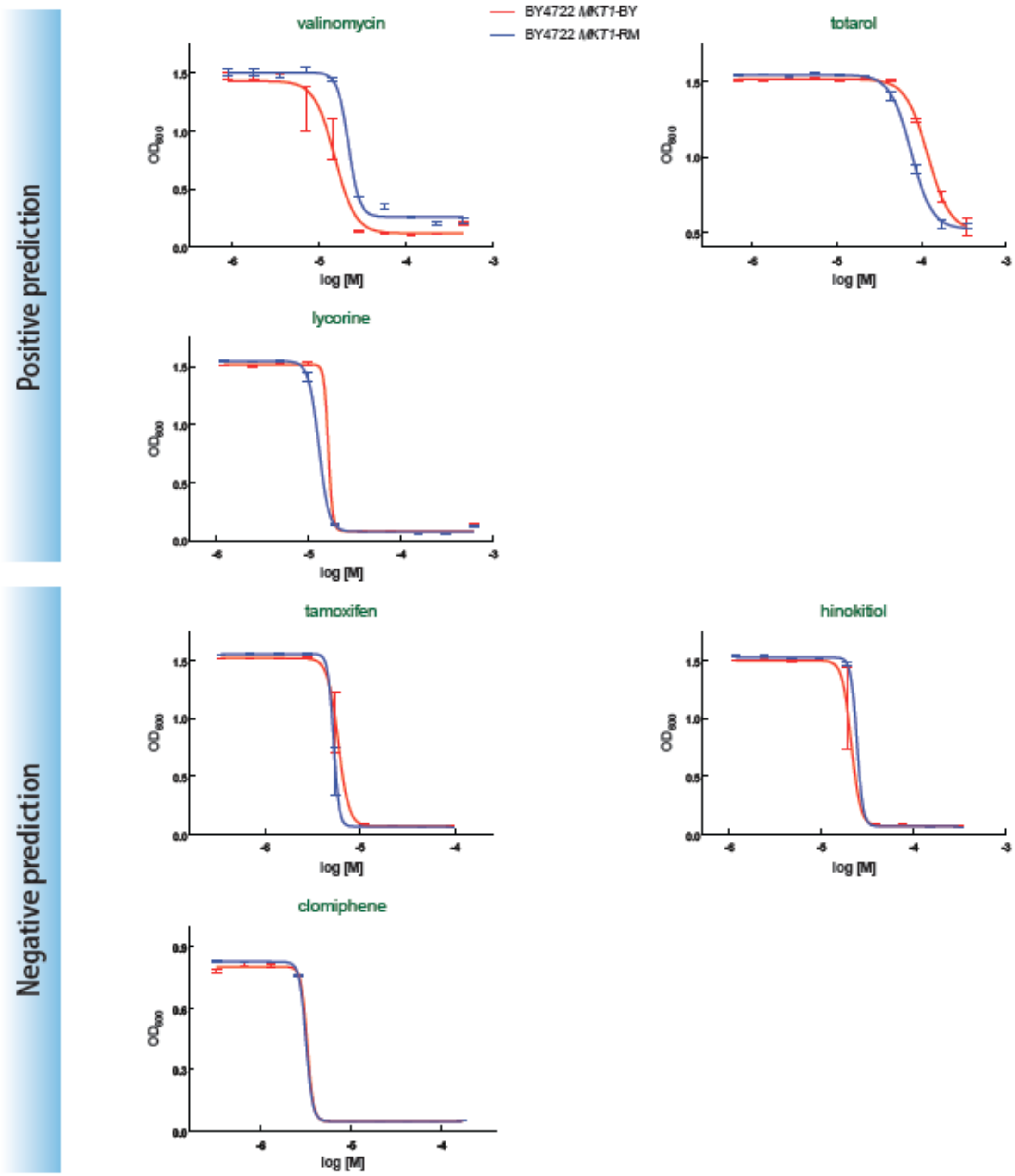

Figure 4-7: Validation of Camelot's MKT1 predictions.

Averaged OD600 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.

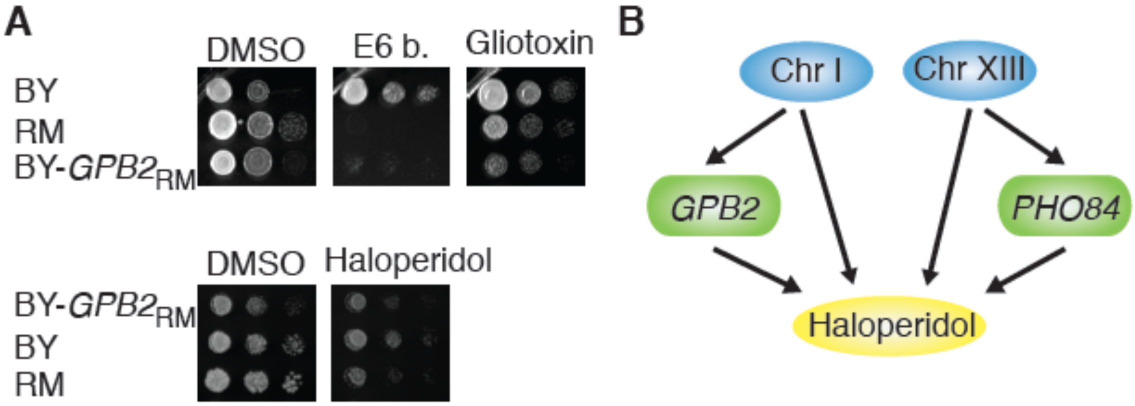

Figure 4-8: Causal role of *GPB2* in response to drugs.
A. Strains were grown overnight in YPD medium, diluted to OD_{600} ~0.2 and plated with 10-fold dilution on YPD+drug media (see Materials and Methods). The top 3 panels are photos of YPD plates containing DMSO (control), E6 berbamine or gliotoxin. 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 allele-swapped strain (BY *GPB2*-RM) grows at a similar rate to the RM strain. B. Camelot identifies two loci (Chromosome I: 1-55329 and Chromosome XIII: 27644-33681) and causal genes encoded within these loci, *GPB2* and *PHO84*, that are responsible for the response to haloperidol.

4.4 *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 4 new phenotypes (response to drug) to *PHO84* and 3 new phenotypes to *MKT1*. To further test Camelot, we looked to see 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:1-55329, 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 allele-swapped strain (BY *GPB2*-RM), in which the entire BY *GPB2* coding region was replaced with that from the RM strain (see Section 4.7.2) 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 allele-swapped strain is highly similar to the RM strain on E6 berbamine and haloperidol suggesting variation in the *GPB2* sequence accounts for much of the difference in the response to these drugs (Figure 4-8A).

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) , while 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 suggests that the mechanism of action involves G-protein signaling. 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 4-8B). 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 4-9) 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.

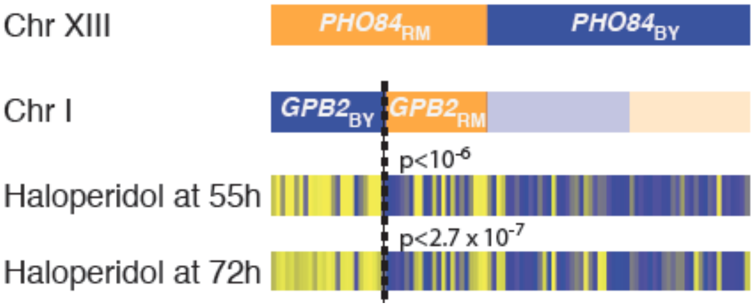

Figure 4-9: Analysis shows *GPB2* and *PHO84* interact with each other to influence the growth in the presence of haloperidol.
Shown are the genotypes for *PHO84* and *GPB2*, and growth in the presence of haloperidol. Segregants with both the *PHO84*-RM and *GPB2*-BY alleles have significantly better resistance (p-value from Wilcoxon rank-sum test) to haloperidol compared with other segregants.

4.5 *PHO84* gene expression and feedback

In total, we validated 18/18 predictions made using the zoom-in score, including 9/9 for *PHO84*. While *PHO84* has 2 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 allele-swap 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 condition (no drug) to monitor *PHO84* expression in the BY, RM and the allele-swapped (BY *PHO84-RM*) strains.

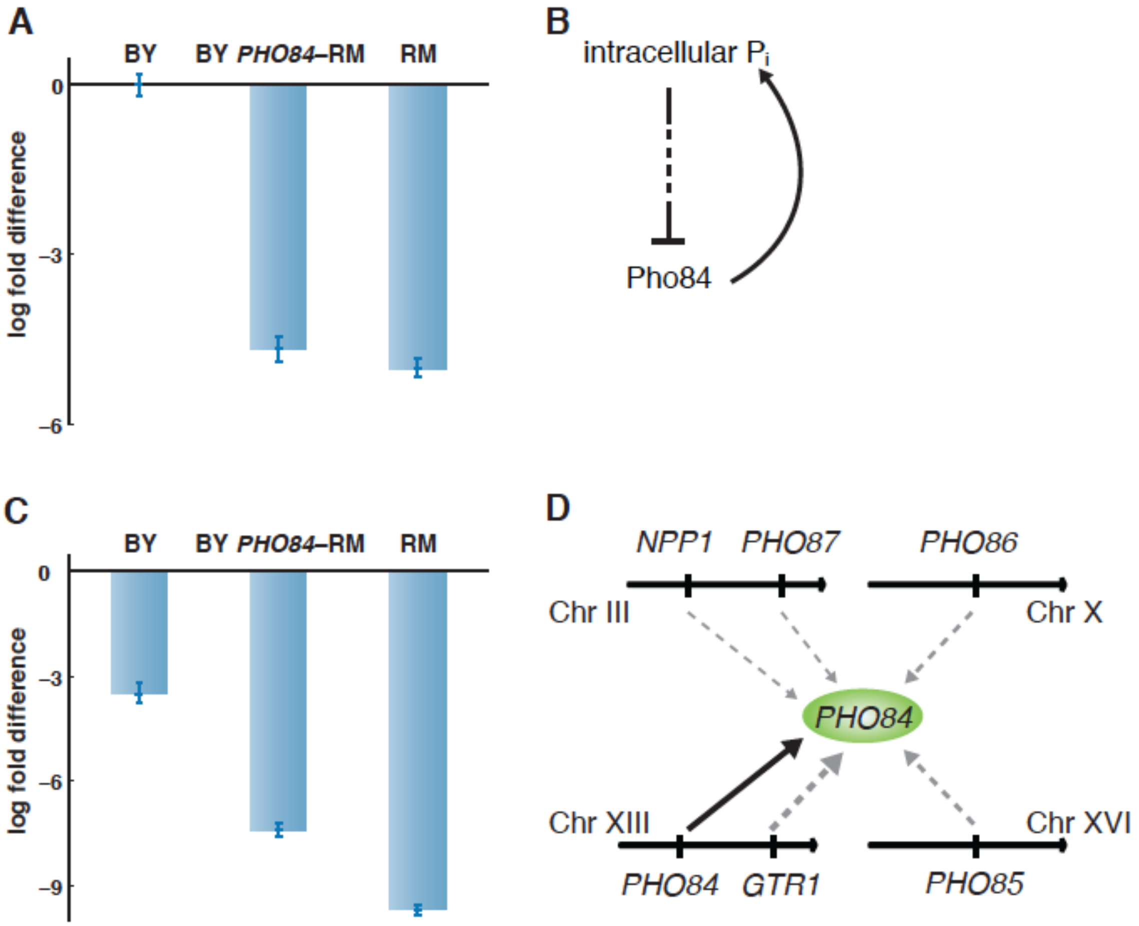

Figure 4-10: Feedback in *PHO84* expression.

A. Expression levels of *PHO84* in the BY, RM and the *PHO84*-RM allele-swapped strain. The expression of *PHO84* in the allele-swapped and RM strains is similar and significantly lower than in BY. The fold difference is calculated relative to the BY strain. Since the allele-swapped strain only differs from the BY in the *PHO84* coding sequence, feedback regulation must act via the *PHO84* gene itself. The error bars represent the standard deviation 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 in SC+low phosphate media for each strain. The cells were harvested 90 minutes after the addition of phosphate. The allele-swapped 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 Materials and Methods).

Though the allele-swapped (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 4-10A). The difference in expression results from negative feedback that acts via the Pho84 protein in high phosphate (Figure 4-10B) (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 down-regulated in high phosphate, relative to low phosphate in all 3 strains (Figure 4-10C). 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. In low phosphate, the gene expression for all 3 strains is similar, suggesting that the loop is not active. This implies that the relative strength of the negative feedback differs between strains in the high phosphate conditions that activate this loop.

We used arsenate, a toxic non-metabolisable 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 (Figure 4-11). 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 via a feedback mechanism; the expression level itself is unlikely to be causal directly.

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 in Chen et al. (2009)). 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 Materials and Methods) weakly influence *PHO84* and the expression data represents the combined influence of these factors (Figure 4-10D).

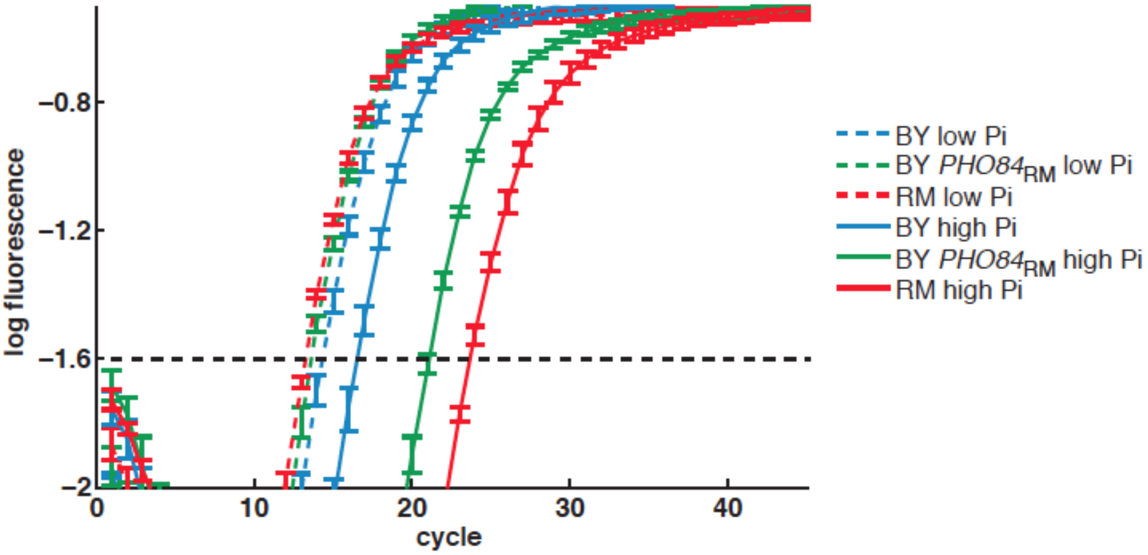

Figure 4-11: Feedback regulation of PHO84 is stronger in RM than BY. Abundance of PHO84 for each strain in high and low phosphate media measured using RT-PCR. PHO84 is expressed at similar level in all three strains under SC+low phosphate conditions. The addition of phosphate results in repression of 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.

4.6 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 anti-depressant haloperidol. 25/27 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, *e.g. SORT1* associated with lipid metabolism (Willer et al., 2008) and multiple genes associated with metabolic traits (Emilsson et al., 2008).

Others have demonstrated how integrating genotype and gene expression data can be used to better understand the relationship between genotype to phenotype in populations and towards understanding disease (Chen et al., 2008; Emilsson et al., 2008; Mehrabian et al., 2005; Schadt et al., 2005) and the predictive value of gene expression towards classifying phenotype (Alizadeh et al., 2000; Golub et al., 1999; Huang et al., 2007; Kutalik et al., 2008; van't Veer et al., 2002). 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 in 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 4-10D). 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 can potentially identify the mechanism of action and the biological processes involved, thereby expanding the number of drug targets, e.g., 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 accommodating a genome of greater scale and complexity carries a heavy statistical burden, Camelot provides another step towards the realization of personalised medicine, as well as highlighting the power to be gained by exploiting gene expression data for this application.

4.7 Materials and Methods

4.7.1 Validation Growth Experiments

Strains used in this study are listed in Table 4-3. The *MKT1*-SK1 (D55N) and *PHO84* (L259P) allele-swapped strain is as described (Deutschbauer and Davis, 2005; Perlstein et al., 2007). The *dhh1Δ* strain was a gift from Liz Miller (Columbia University). *MKT1*, *PHO84* and *dhh1Δ*, growth yield experiments were performed in multi-well (96- or 384-well) plates as described in (Perlstein et al., 2007). Serial dilutions were carried out in replicate, and the resulting growth yield and IC50 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 OD600~0.1 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 to 2 days for *GPB2* or 5 days for the arsenate assay. Final concentrations of drugs were as (Perlstein et al., 2007), For *GPB2*: gliotoxin (15.3µM), E6-berbamine (16.5µM) and haloperidol (66.6µM). For *PHO84*: arsenate was used at a final concentration of 2mM and methylphosphonate at 10mM as Mouillon and Persson (2005).

Table 4-3: 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.

4.7.2 Generation of the GPB2$_{RM}$ allele-swapped 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 allele-swapped strain matched that of RM, while the upstream region remained that of BY.   Primers used in this study are listed in Table 4-4.

4.7.3 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 RiboPureYeast kit according to the manufacturers instructions, with the exception that 10µg sample was digested twice each for 1 hour at 37°C with 2U DNase I. cDNA was made using the Stratagene AffinityScript kit and random primers. RT–PCR was performed on a Chromo4 machine (BioRad) using iQ SYBR Green Supermix (Biorad) and primers listed in Table 4-4. 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 (250µM). After at least two doublings cultures were split in two and phosphate was added to 15mM final (‘high phosphate’) to one flask. The cultures were grown for a further 80 minutes before harvesting. Phosphate media was made using YNB – potassium phosphate (Sunrise Science) supplemented with amino acids, glucose, ammonium sulphate and potassium phosphate. Potassium chloride (10mM) was added to low phosphate media.

Table 4-4: Primers used

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	ttatattctactactaaacaaagtttacaaagtgaaagcattgaaaactgcctttttcgtacgctgcaggtcgac	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

4.7.4 Statistical Analysis

Camelot, the elastic net *L* model and linkage analysis are evaluated with ten-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%; $p<5.6\times10^{-5}$) (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 40kb for the purposes of enrichment analysis. The linkages for *PHO84* gene expression were obtained with Wilcoxon rank-sum test ($p < 0.01$).

Chapter 5 Drug Sensitivity in Cancer Cell Lines

5.1 Introduction

Personalized medicine for human disease is a holy grail in modern clinical care. Clinicians envision that they can determine the optimal treatments for each individual patient based on his or her genomic profiles. With the recent advances in microarray and next-generation sequencing technologies, the prospects of personalized medicine look brighter than ever (Altelaar et al., 2013). Many efforts have been devoted to associate genomic profiles with diseases and clinical outcomes. For example, a large collection of gene expression profiles of human cell lines treated with chemical compounds has been used to associate gene signature with diseases and drugs (Lamb et al., 2006). Recently, a study reported that by monitoring an individual over 14 months with multiple molecular-omics profiling, they were able to predict the onset diabetes (Chen et al., 2012).

The use of genomics data to improve clinical treatment is perhaps most studied in cancer. Many pioneer studies have shown how one can use signatures of gene expression to predict clinical outcome for individual patients (Alizadeh et al., 2000; Kutalik et al., 2008; van't Veer et al., 2002). However, primary tumors are experimentally intractable for many functional studies. Collections of human cancer cell lines therefore provide an excellent system for us to study the genomic characteristics of cancers and how these characteristics can be associated with drug sensitivity (Heiser et al., 2012; Weinstein, 2006). Recently, two large collections of drug screens and genomics profiles of cancer cell lines have been published (Barretina et al., 2012; Garnett et al., 2012). Such resources are invaluable for the study of pharmacogenomics in cancer. Indeed, the two studies accompanying the data have shed light into how some drugs are potent for specific cancer types or cancers with specific mutations.

These data also offer the opportunity to develop methods that can be used in a personalized medicine scenario. By associating genomic features with drug sensitivity in cancer cell lines, we may build models based on the associated features to predict the drug response of an individual (Heinemann et al., 2013; McLeod, 2013). Moreover, attribution of drug sensitivity to genomic features allows us to understand the mechanism of drug action and discover the underlying reasons for resistance to the treatment.

A key challenge towards linking genetic characteristics to drug sensitivity is the role of context in biological systems. It has long been known that interaction between genetics and context, such as environment and tissues, plays an important part in phenotypic variation (Dimas et al., 2009; Emilsson et al., 2008; Gerke et al., 2010). It is as if environment or tissues condition cells to activate different regulatory mechanisms, providing the context for diverse cellular phenotypes in addition to the influence of genetics. For example, regulation of gene expression has been shown to have patterns specific to tissues and cell-types (Dimas et al., 2009; Fu et al., 2012; Nica et al., 2011; Price et al., 2011). In tumorigenesis, diverse patterns of mutation, gene expression, and epigenetic regulation have also been observed in cancer-specific or tissue-specific manner (Bissell and Labarge, 2005; Wolff et al., 2010). For example, germ-line mutations in DNA-repair genes *BRCA1* or *BRCA2* are mostly associated with breast and ovarian cancers, but rarely other tumors. Several hypotheses have been proposed to explain such tissue specificity (Monteiro, 2003). In addition, many studies have shown the induction of the tumor suppressor p53 and the regulation of its targets can be cell-type, tissue-type and stress dependent (Bouvard et al., 2000; Fei et al., 2002; Feng et al., 2007). Another example is the tissue- and gender-dependent mutation frequency of the oncogenic *PIK3CA*: higher frequency of mutation is observed in females than in males in colon cancers but not in breast cancers (Benvenuti et al., 2008).

This context dependency of tumorigenesis impacts the efficacy of treatment. For example, PLX4732, a RAF inhibitor targeting oncogenic $BRAF^{V600E}$, is a potent treatment for melanoma patients with the mutation (Bollag et al., 2010). However, colon cancer patients with the same mutation do not respond to PLX4732 (Kopetz et al., 2010). It is therefore important to take into account such context specificity, created by cancer types, when we analyze the genomics of drug sensitivity.

Studies of such context-specificity may be the key to understanding the variation of drug response. Recent reports have shown that though *BRAF* is mutated frequently in both cancers, colon cancer expresses higher EGFR than melanoma. The difference of EGFR expression between these two cancers may mediate the drug resistance to PLX4732 in colon cancer (Corcoran et al., 2012; Prahallad et al., 2012). Moreover, the expression of EGFR in melanoma may also contribute to the variation of response to PLX4732. It is the understanding of such variation that can help us make accurate predictions of response to treatments.

It is no surprise that predictive models built with the consideration of context perform better than those without. For example, models built using only melanoma data give better prediction for drug sensitivity of melanoma samples than those built using data of mixed cancer types (Barretina et al., 2012). This argues that we should focus on one cancer type for predictive models for drug sensitivity. While such strategy allows us to avoid confounding influence of context, it constrains us to the data of a small number of samples. This poses a great challenge for model learning due to lack in statistical power.

We utilize common grounds between cancer types to overcome this paucity of data. For examples, basal-like breast cancer and ovarian cancer shares many molecular signatures (Cancer Genome Atlas, 2012). Shared tumor-associated antigens (Chi et al., 1997) and dysregulated pathways (Prickett and Samuels, 2012) have been also reported in melanoma and glioma.

Taking advantages of common genomic features, which may be shared between similar cancer types, we propose an algorithm that builds predictive models by selecting genomic features that are shared between cancer types, or are specific only in some cancer types. By pooling samples of similar cancer types together, we increase the power to uncover biomarkers predictive of drug sensitivity. Meanwhile, we take context-dependency into account by allowing only genomic features that contribute predictive effects to specific types of cancers.

In the next chapter, we propose an algorithm that aims to build predictive models for drug sensitivity of similar cancer types. Contrary to previous methods that assume all samples have the same predictive features, we also explicitly learn if the samples of a subtype have different mechanism contributing to the phenotype. For data with multiple subtypes of samples, the algorithm also identifies the relevant subtype that dictates the context specificity. This suggests that it could be applied to data such as breast cancer where multiple subtypes have known to respond treatment differently (Heiser et al., 2012).

5.2 Datasets and Goals

Our goal is to devise a model that can predict drug sensitivity from genomic profiles of a tumor, taking into account cancer or tissue types. Here, cancer or tissue types act as context to influence the predictive programs of drug sensitivity. We aim to identify predictive features for drug sensitivity that are shared across context and predictors that are specific only for a certain context.

We take the data from Cancer Cell Line Encyclopedia (Barretina et al., 2012) to study the association between drug sensitivity and both shared and context-specific genomic features among cancers. Data from CCLE includes large collection of cell lines (n=504) of various cancers and tissue types. The cell lines were treated with one of 24 drugs at various concentrations. Sensitivity to each drug was characterized for each cell line from the growth under the influence of the drug. For our analysis, two phenotypes are considered for each compound: concentration that inhibits 50% of proliferation (IC50) and activity area above the curve fitted from the drug response data (ACT). In addition, mutation, copy number and gene expression profiles of the cell lines are used for the analysis. Three sets of data are extracted from the CCLE collection, including samples of breast carcinoma, ovary carcinoma, melanoma, glioma, and samples derived from haematopoietic and lymphoid tissues (see Section 7.2.1 for details).

In the next chapter, we will first describe the split-regression algorithm. Statistical analysis of the algorithm will be discussed in Chapter 7 and biological results from the CCLE datasets will be discussed in Chapter 8.

Chapter 6 Split-Regression Algorithm

Our aim is to identify relevant genomic features that are predictive of the phenotype, drug sensitivity. As in the problem posed in Chapter 2 and Camelot algorithm described in Section 3.1, the key to build a predictive model for drug response is to select relevant features from the genomic profiles.

As we know, context plays an important role in biology. Cells of a person share the same genetic information, and yet each tissue is dramatically different from each other. Even when two tumors have the same mutations, the clinical outcome may be very different. For example, *BRAF*^{V600E} is frequently observed in both melanoma and colon cancers, but the BRAF inhibitor PLX4732 only inhibits tumor growth in melanoma patients (Corcoran et al., 2012; Prahallad et al., 2012). Moreover, the context relevant to the response to a drug may not be known in advance.

Due to the confounding influence of cancer and tissue types, we cannot simply apply linear regression or a Camelot-like algorithm to the data (Chapter 3). One straightforward solution is to limit the analysis to one cancer type at a time, since this will constrain the analysis within a context. The caveat with this strategy is that then the sample size is too small. For example, all cancers in the CCLE collection have fewer than 40 samples with drug sensitivity data, except non-small-cell lung cancer and melanoma. The lack of statistical power, due to small sample size, is exacerbated by the size and complexity of the human genome. Compared to the yeast data described in Section 2.2 (~100 yeast strains and about 6000 genes in yeast genome), the human genome contains more than 20,000 genes and 60 million DNA variants. This makes the feature selection a very challenging problem.

In order to gain statistical power and still account for some context specificity, we pool samples of similar cancer types together for the analysis. The assumption is similar cancer types might share similar mechanisms contributing to drug sensitivity, as well as specific mechanisms to each cancer. We developed an algorithm that aims to uncover predictive features that are shared across contexts and features that are predictive only a certain context. A context can be a cancer type, tissue type, or cancer subtype. We refer to this context as the *relevant subtype* that separates individuals into two groups where the predictive program of drug sensitivity can be different.

The intuition behind our algorithm is based on transfer learning theory (Raina et al., 2006). First, because we learn models from samples of similar cancers, we essentially share the information between cancers by assuming they may share the same genomic features responsible for drug sensitivity. This is what enables us to retain enough samples to learn predictive models for sensitivity to each drug.

Second, we transfer information between drugs. We assume that if two drugs induce similar response, their predictive programs would be similar as well. For examples, if two drugs induce highly correlated phenotypes and we have observed gene *A* as a predictor for sensitivity to one drug, it is more likely gene *A* is also predictive for the other drug. Therefore, we can boost the learning power of our algorithm by sharing model information between the phenotypes.

Here we employ *L0-norm* regularized regression to select features. L0-norm regularized regression has properties well founded in information theory and Bayesian statistics, and has recently shown to perform excellently in prediction tasks (Dhillon et al., 2011). In this chapter, a brief background of *L0-norm* regularized regression is presented in Section 6.1, followed by details of our algorithm extended from *L0-norm* regularized regression (Section 6.2 and 6.3). Finally, statistical analysis of our algorithm on simulated and CCLE data are presented in Chapter 7.

6.1 Background: L0-norm regularized regression

There is a large family of regularized regression that aims to obtain a sparse model. They usually follow the form:

$$\min_{\beta}(y - X\beta)^2 + penalty(\beta) \qquad \textbf{Eq. 6.1}$$

where y is the phenotype ($\mathbb{R}^{n\times 1}$), X is the feature matrix ($\mathbb{R}^{n\times p}$) and β is a vector of coefficients ($\mathbb{R}^{p\times 1}$) for the regression model. The goal is to obtain a sparse solution to optimize Eq. 6.1 via enforcement of the regularization function $penalty(\beta)$.

Ideally, the penalty function should be a function of the number of nonzero entries in β. This corresponds to an L0-norm penalty function:

$$\min_{\beta} \ (y - X\beta)^2 + f(\|\beta\|_0) \qquad \textbf{Eq. 6.2}$$

where $f(\|\beta\|_0)$ is a function of L0-norm of ß.

However, Eq. 6.2 is difficult to optimize, as it is a NP-hard problem (Dhillon et al., 2011; Natarajan, 1995). Many approximate penalty functions have been proposed. For example, lasso (Tibshirani, 1996) uses L1 norm as its penalty function whereas elastic net (Zou and Hastie, 2005) uses a combination of L1- and L2-norm. The relaxation of penalty function makes them convex problems and therefore can be solved analytically. Characteristics of these two algorithms in feature selection are described in Section 3.1.1.

However, algorithms based on Eq. 6.2 can be very powerful. They correspond well with information theory and have a straightforward Bayesian interpretation. From an information theoretic perspective, the L0 norm penalty can be interpreted as the model description length. That is, when an entry in β is nonzero, we can describe it with $\log_2(p)+2$ bits, where $\log_2(p)$ bits are used to identify the index of the nonzero entry, and 2 bits are used to encode the actual coefficient (Dhillon et al., 2011). The optimization problem Eq. 6.2 then corresponds to finding a model of minimum description length (MDL) (Rissanen, 1978, 1999), minimizing the fitting error while constraining the complexity of the model.

From Bayesian viewpoint, each feature has $1/p$ probability to be selected and have a nonzero coefficient (Cover and Thomas, 2012). This corresponds to a non-informative prior over the feature space while the minimization of sum of square errors correspond to the maximum of likelihood (Bickel and Li, 1977):

$$\max_{\beta} \ \log \ P(y|X,\beta) + \log P(\beta) \qquad \textbf{Eq. 6.3}$$

where $P(y|X,\beta) \sim N(y|X\beta,\sigma^2)$ with σ^2 as the variance of random noise. The solution to Eq. 6.3 is equivalent to an MAP (maximum *a posteriori*) estimate.

Despite the lack of a closed form solution, L0-norm regularization has several advantages. First, the algorithm is nonparametric, since the sparse selection of predictors is guided by the minimum description length principle. Second, the correspondence between MDL (Eq. 6.2) and Bayesian prior (Eq. 6.3) lends the flexibility to adjust the belief of a feature being relevant or not. That is, we can employ different prior probabilities over β to present our beliefs or preferences in features selection. The prior probability would be translated into number of bits required to encode each β, and therefore allows us to select features using the MDL principle. Third, a greedy algorithm has been recently proposed to obtain a solution to Eq. 6.2 and the resulting models are demonstrated to have excellent performance (Dhillon et al., 2011). The implementation of greedy search lends the capability for the algorithm to search for the relevant predictors in the large space spanned from the interaction of genomics and context. In the following section, we will see how to extend Eq. 6.2 to allow the encoding of context-specificity into the model.

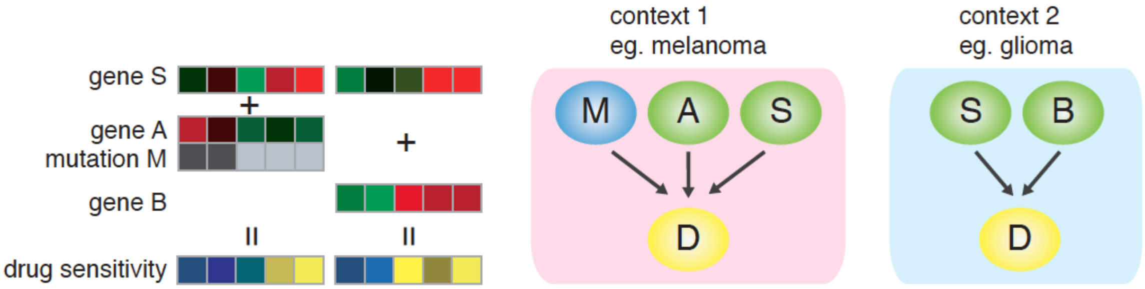

Figure 6-1: Concept of split-regression.
Two cancer types may have different genomic factors that are predictive of drug sensitivity. For example, the drug sensitivity of melanoma samples can be predicted by mutation of M and gene expression of A and S, whereas in glioma, expression of gene S and B are the predictors. Split-regression takes advantage of pooling samples together to gain statistical power, identifying both shared (gene S) and context-specific features (A, B and M). In cases where the relevant context is unknown, the algorithm searches for the best “split”, if any, to separate samples into two groups.

6.2 Extended Model to Consider Context Specificity: Split-Regression

Here we consider samples that may be a mixture of two populations, for example, a collection of cell lines of two cancer types. If these two cancer types share similar molecular characteristics, some genomic features may be predictive of drug sensitivity in both cancers. For example, because basal-like breast cancer and ovarian cancer have similar genomic signatures, it is likely same predictors can be used to predict drug sensitivity in both cancers. When we pool samples of these two cancers together, we increase the power to uncover predictors that have weak effects on drug sensitivity.

Additionally, cancer-specific genomic features may be responsible for the drug response in each cancer. In order to capture this, we extend Eq. 6.2 to

$$\min_{t,\beta^S,\beta^{t1},\beta^{t0}} \sum_i^n (y_i - \sum_j^p x_{ij}\beta_j^S - \sum_j^p \delta(z_{it} = 1)x_{ij}\beta_j^{t1}$$
$$- \sum_j^p \delta(z_{it} = 0)x_{ij}\beta_j^{t0})^2 \qquad \textbf{Eq. 6.4}$$
$$+ penalty(t) + f(\|\beta^S\|_0)$$
$$+ f(\|\beta^{t1}\|_0) + f(\|\beta^{t0}\|_0)$$

where z_{it} is a binary variable representing if sample *i* is of cancer type *t*, and $\delta(.)$ is an indicator function that returns 1 when sample *i* is of the cancer type queried. β^S are the coefficients for shared genomic features, whereas β^{t1} and β^{t0} are the coefficients for cancer-type-specific features.

Note that we also select *t* in addition to $\beta_j^{\cdot}$. The optimal *t* represents the *relevant subtype*, the best classification of samples, which allows us to optimize Eq. 6.4. Samples in a dataset may be classified in various ways. By treating z_{it} as binary we can extend the variable *z* to represent many different classifications of samples. For example, in breast cancer, we can classify if a sample “is of basal subtype or not” or “is of luminal subtype or not.” This translates into two binary variables. The binary representation also allows classifications to be hierarchical or overlapping.

As the subtype variables are binary, we also refer to the relevant subtype as a "*split.*" It indicates the subtype splits the samples into two groups, each of which may have its own set of predictors. We refer to the regression problem in Eq. 6.4 as *split regression*. Figure 6-1 illustrates the concept of split regression.

The regularization terms in Eq. 6.4 correspond to the penalty term in Eq. 6.2. As we use MDL to guide the feature selection, we need to decide a coding scheme to describe the features. Assuming no prior knowledge or preference for feature selection, we encode each feature using the same number of bits. For each subtype t, $\log_2(T)$ bits are required to index t among all T classifications. For each coefficient in β^S, β^{t1} or β^{t0}, we only need to describe those that are non-zero in the model. Each nonzero coefficient requires $\log_2(p)$ bits to index the genomic features among all p features. In addition, two bits are used to encode the actual coefficient[1] (Dhillon et al., 2011). These coding costs correspond to uniform prior probabilities of these variables being relevant or not (Section 6.1).

6.3 Transferring Knowledge Between Phenotypes

Biological datasets are often of small sample size. This poses as a challenge for statistical analysis, since the power of computational methods largely relies on sample size. Here,

[1] In the actual implementation, two additional bits is used for each $\beta_j^{\cdot}$ to specify whether it is in β^S, β^{t1} or β^{t0}.

we take advantage of the similarity between multiple drug responses. That is, when two drugs induce similar drug response across samples, we hypothesize their predictive programs would share common features. This approach is termed *transfer learning*. The idea behind transfer learning is that by sharing information between similar phenotype data, we increase statistical power to detect the features that are predictive of similar drugs. For example, the activity area data (Section 5.2) of samples are highly correlated among the drugs of the same family (Figure 6-2), such chemotherapy drugs paclitaxel, topotecan, and irinotecan. It is very likely the same genomic features are predictive of sensitivity to these drugs.

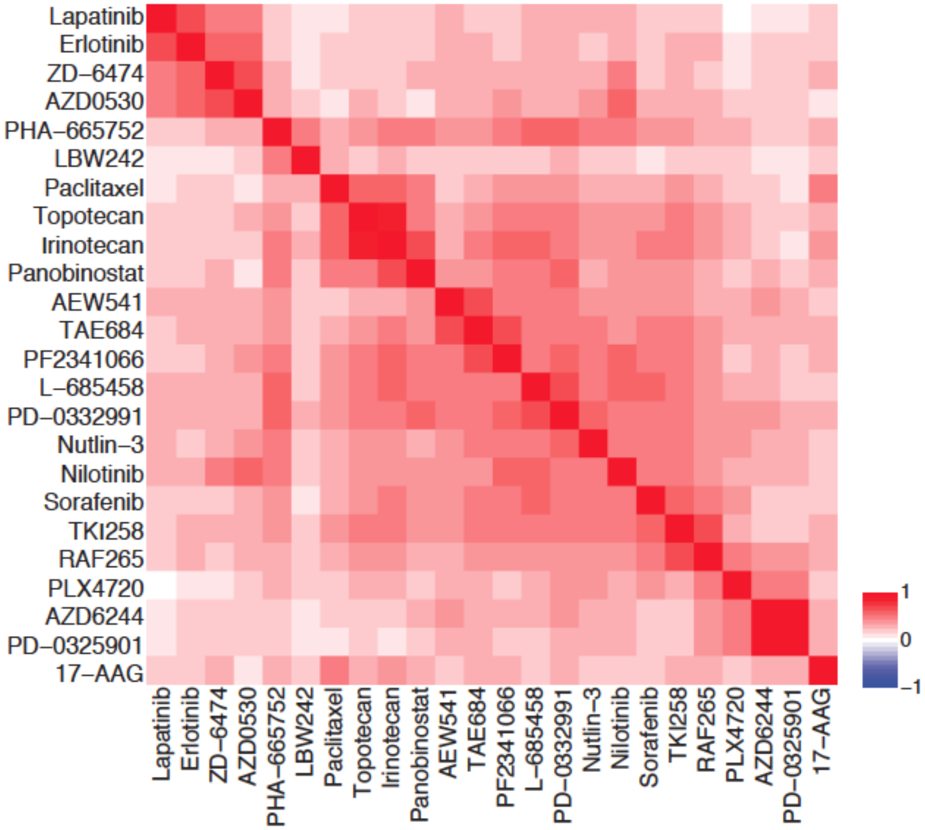

Figure 6-2: Similarity between phenotype data.
The symmetric matrix represents Pearson correlation coefficients between the responses of each pair of drugs. The data is from CCLE. Activity area data of samples treated with each drug are used to calculate the correlation coefficients. Responses to drugs targeting EGFR (laptinib, erlotinib, ZD-6474, and AZD0530) are highly correlated.

Transfer learning is therefore a technique that allows the algorithm to boost the confidence of feature selection for similar drugs. In our algorithm, this means the cost of a feature can be adjusted to reflect the probability of it being chosen to model the response to a drug. For example, if a feature is chosen for two similar phenotypes, it is more likely to be a relevant predictor for another similar phenotype. With this knowledge, we can change the penalty of this feature to reflect the probability of this feature being predictive of drug sensitivity.

In order to transfer the knowledge between phenotypes, we adopt an iterative learning procedure: we obtain approximate solutions to Eq. 6.4 in the first iteration, where each feature has the same penalty (uniform prior, Section 6.2). In the following iterations, the penalty of each feature for each phenotype is adjusted according to the frequency of the feature being selected in the previous iterations. The outline of this iterative learning procedure is described in Table 6-1.

Table 6-1: Outline of Transfer Learning in the Algorithm

```
for each y in Y
    estimate $\{\beta\}^{\mathbf{y}}$ with uniform prior;
do until stop
    estimate prior $P^{y}(\beta)$ for each y based on $\{\beta\}^{\mathbf{Y}}$;
    for each y in Y
        estimate $\{\beta\}^{\mathbf{y}}$ with $P^{y}(\beta)$;
    repeat;
```

6.3.1 Estimation of Prior

In Section 6.1 we describe a mapping between MDL and a Bayesian prior. The penalty of each feature can be interpreted as $-\log\left(P\left(\beta_j\right)\right)$. This mapping allows us to adjust the penalty by a prior $P\left(\beta_j\right)$ learned from the previous iteration. Specifically, we estimate the prior for phenotype *y* by

$$P^y(\beta_j \neq 0) = \frac{\sum_{k=1}^{K} w_{yk} f_{jk} + a}{\sum_{k=1}^{K} w_{yk} + b} \qquad \textbf{Eq. 6.5}$$

where *k* enumerates phenotype, $f_{jk}$is the frequency of feature *j* being selected for phenotype *k* in the previous iteration, *a* and *b* are hyper-parameters of a non-informative beta prior to regulate this Bernoulli distribution, and w_{yk} is a similarity score between phenotypes *y* and *k*. We refer to f_{jk} as the *relevant frequency* of feature *j* for phenotype *k*.

Eq. 6.5 represents our belief of a feature *j* being relevant to phenotype *y*. This belief is iteratively updated according to models learned in the previous iteration and the similarity between the phenotypes.

Similarly, the penalty of the relevant subtype *t* for phenotype *y* corresponds to

$$P^y(t\ is\ the\ relevant\ subtype) = \frac{\sum_{k=1}^{K} w_{yk} f_{tk} + a}{\sum_{k=1}^{K} w_{yk} + b} \qquad \textbf{Eq. 6.6}$$

where f_{tk} represents the frequency of subtype *t* being chosen as the relevant subtype for phenotype *k*.

It is straightforward to estimate $P^y(\beta_j^S \neq 0)$ with Eq. 6.5, using f_{jk}^S (the relevant frequency of feature *j* for phenotype *k* across all samples). $P^y(\beta_j^{t1} \neq 0)$ and $P^y(\beta_j^{t0} \neq 0)$ can be derived similarly from Eq. 6.5 for subtype-specific features. Ideally,

$P^y(\beta_j^{t1} \neq 0)$ and $P^y(\beta_j^{t0} \neq 0)$ should be estimated for each possible subtype *t*. However, when the number of possible subtypes *T* is large, it is impractical to estimate all the priors and infeasible to keep them in memory if the number of features *p* is large. Nevertheless, assuming the transference only occurs between similar phenotypes, we approximate $P^y(\beta_j^{t1} \neq 0)$ and $P^y(\beta_j^{t0} \neq 0)$ with Eq. 6.5 using f_{jk}^t (the relevant frequency of feature *j* for phenotype *k* for samples of subtype *t*) regardless of which subtype *t* is chosen as relevant in each run.

6.3.2 Estimation of Relevant Frequency of Features

In Eq. 6.5, f_{jk} is either 0 or 1, since feature *j* can only be selected or not for phenotype *k*. To obtain a better estimation of f_{jk}, we use bootstrapping techniques (Efron, 1979). For each iteration, we subsample the cell lines multiple times, with replacement, and obtain $\beta's$ for each sampling. Thus, f_{jk} can be now estimated from the multiple bootstrapping results:

$$f_{jk} = \frac{\sum_{l=1}^{B} I(\beta_j^{kl} \neq 0)}{B} \qquad \textbf{Eq. 6.7}$$

where *B* is the number of bootstrapping runs, β_j^{kl} is the coefficient of feature *j* for phenotype *k* in bootstrapping run *l* and $I(.)$ is an indicator function that returns one when the criterion in the function is satisfied. Compared to a single run, bootstrapping allows us to estimate the relevant frequency of features robustly.

In next chapter we will evaluate the models obtained by split-regression. First we will demonstrate the algorithm can accurately retrieve relevant features in simulated data (Section 7.1). Then, details of applications to CCLE data are described in Section 7.2.

Chapter 7 Statistical Analysis of Split-Regression

7.1 Simulation

We generated a synthetic dataset to test the proposed algorithm. In order to generate the data, the algorithm is first applied to CCLE blood cancer samples (*n*=70) where seven binary subtypes are defined (Section 7.2.1). Mutation and gene expression features are used and 100 bootstrapping datasets are drawn to train the models for *activity area* of each drug (Section 7.2.1). The whole training and bootstrapping procedure is repeated for ten iterations, and the priors (Eq. 6.5) are updated after each iteration. 0.15 is used as a frequency cutoff to select features after the final iteration. One phenotype did not have any features passing the threshold, so it is dropped from further analysis.

Of 23 phenotypes, 6 are set to have only shared features among all samples. That is, no subtype-specific features are used to simulate the sensitivity data for the 6 drugs. This is to test if the algorithm would only select shared features or if it would be biased to select subtype-specific features. The number of features used to synthesize phenotypes ranges from 2 to 11, excluding intercepts. After the features are selected, the coefficients are estimated using ridge regression with $\lambda = 0.01$:

$$\min_{\beta} \ (y - X\beta)^2 + \lambda \|\beta\|_2^2 \qquad \textbf{Eq. 7.1}$$

The choice of ridge regression is to regularize the coefficients of collinear features.

Gaussian-distributed noise is added to the simulated phenotypes. The noise is drawn from $N(0, \sigma^2)$, where $\sigma^2 = \varepsilon^2 \sigma_y^2$, proportional to the variance (σ_y^2) of synthetic data. ε^2 is set to 0.2, 0.4 ,0.6, and 0.8 to generate different levels of noise.

The algorithm is then applied to the simulated data. 100 bootstrapping datasets are drawn for modeling during each iteration, and 10 iterations are repeated to adjust the prior according to Eq. 6.5. The final model for each phenotype is determined by the relevant frequency of a feature (Eq. 6.7). For the relevant subtype, only the one with the maximum relevant frequency passing a predefined threshold, θ, is considered significant and as the final relevant subtype. Three different threshold values, 0.3, 0.5 and 0.8, are used for θ to see the effects of the threshold.

In order to evaluate the algorithm, three metrics are calculated: split accuracy, precision and recall. Split accuracy is defined as:

$$split\,accuracy = \frac{\sum_{k=1}^{K} I(\widehat{t_k} = t_k)}{K} \qquad \textbf{Eq. 7.2}$$

where $\widehat{t_k}$ is the relevant subtype retrieved for phenotype *k* and t_k is the true relevant subtype used to synthesize the phenotype.

Precision and recall are defined for each synthetic phenotype *k*:

$$precision^k = \frac{\sum_{j=1}^{p} I(f_{jk} \geq \theta \; and \; \beta_j^k \neq 0)}{\sum_{j=1}^{p} I(f_{jk} \geq \theta)} \qquad \textbf{Eq. 7.3}$$

$$recall^k = \frac{\sum_{j=1}^{p} I\left(f_{jk} \geq \theta \; and \; \beta_j^k \neq 0\right)}{\sum_{j=1}^{p} I\left(\beta_j^k \neq 0\right)} \qquad \textbf{Eq. 7.4}$$

where θ is the threshold for relevant frequency, f_{jk} is the relevant frequency of feature j, and β_j^k is the coefficient used to synthesize the phenotype. Here $\beta_{\cdot}^{k}$ includes both shared and subtype-specific features. Intuitively, precision measures how many of the selected features are truly relevant whereas recall indicates how many relevant features are recovered. There is usually a trade-off between precision and recall. For example, a lenient algorithm can select as many features as possible to achieve high recall, but its precision would be low. In biology, high precision is often desired since it implies low false positive rate.

7.1.1 Effects of transfer learning

Figure 7-1 shows the split accuracy and Figure 7-2 shows the precision and recall of the results from the simulated data. For clear illustration, the precision and recall is averaged across all phenotypes. The figures show the evaluation metrics as a function of iterations. We can see that all three metrics improve after the first iteration. Interestingly, the improvement often stops after three to five iterations. After three to five iterations, split accuracy and recall remain the same. For precision, there is a dependency between the number of iterations and the threshold chosen for the relevant frequency. When a conservative threshold is chosen ($f \geq 0.8$), precision increases and stabilizes through iterations. However, when a lenient threshold is applied ($f \geq 0.3$ or 0.5), precision starts to deteriorate after two or three iterations. This is caused by a problem called *negative transfer* (Caruana, 1998), meaning the transfer learning could actually negate the performance of models. This indicates that as we adjust the feature prior after each iteration, we also decrease the penalty for many features that are not relevant for the simulated phenotype. The sharing between phenotypes reinforces this adjustment of prior after each iteration. As a result, false positives start to be selected in bootstrapping runs. When a lenient threshold is applied to the relevant frequency to determine significance of features, these false positives would be chosen. This behavior is also emphasized when the noise level increases.

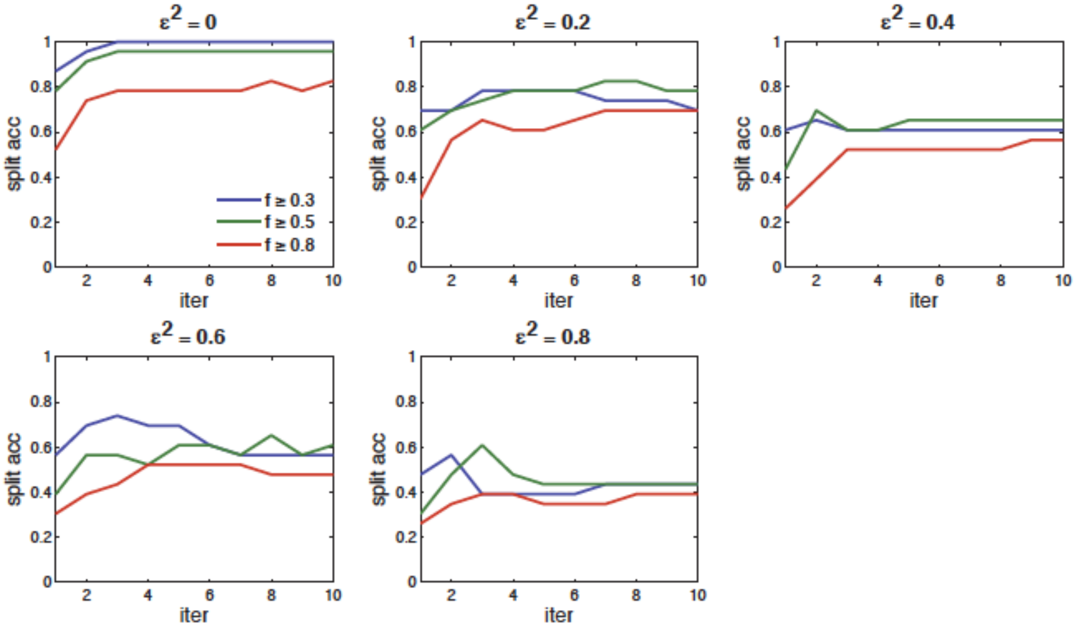

Figure 7-1: Accuracy of identifying the relevant subtypes (splits) in the synthetic data.
The colored lines depict the accuracy of identifying the subtypes used to generate the data, as a function of the number of iterations. In order to decide if a chosen subtype is robust among bootstrapping runs, three different thresholds are used for the frequency of a subtype being chosen among bootstrapping runs. Only the most frequently chosen subtypes that pass the threshold are called relevant. Each panel shows the results from data with different levels of noise.

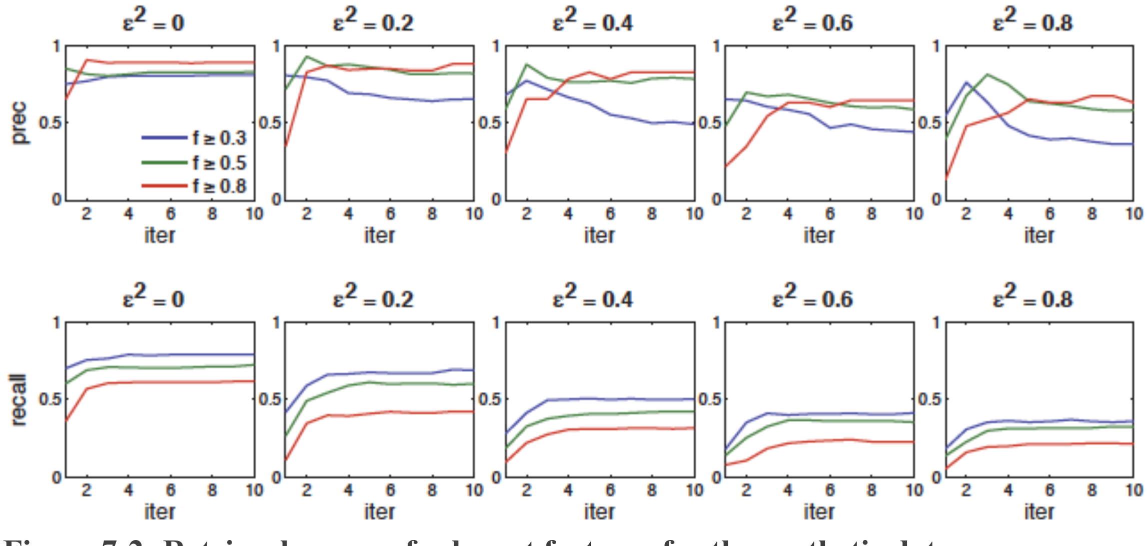

Figure 7-2: Retrieval scores of relevant features for the synthetic data.
Precision (top) and recall (bottom) are used to evaluate the retrievals of features used to generate the synthetic data. The scores are plotted as a function of the number of iterations. Similar to Figure 7-1, three different thresholds are used for the frequency of a feature being regarded as relevant after the bootstrapping runs.

7.1.2 Effects of bootstrapping

Figure 7-3 shows the comparison of precision and recall from models with and without bootstrapping. The compared models are from the first iteration, in order to control for the effect of transfer learning.

As shown in the figure, bootstrapping significantly increases the precision of the algorithms regardless of the threshold for the relevant frequency. Recall also benefits from bootstrapping when a lenient threshold is applied to relevant frequency. However, as the threshold increases, the algorithm becomes conservative so recall benefits little from bootstrapping in the first iteration. This is remedied by the following iterations as shown in Figure 7-2.

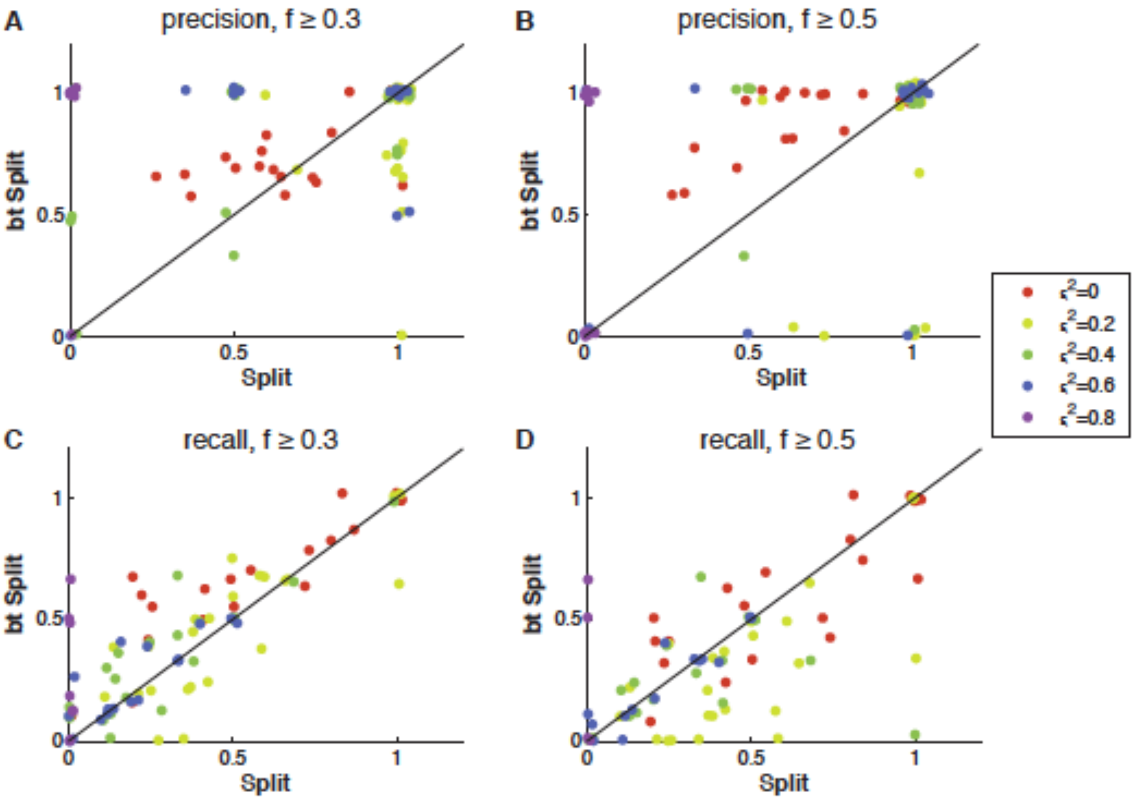

Figure 7-3: Effects of bootstrapping on precision.
For each phenotype, the precision from the first iteration without bootstrapping (x-axis) is plotted against that from bootstrapped models (y-axis). Each dot represents a phenotype, colored according to the noise level added. A. A threshold of 0.3 is applied to the relevant frequency to determine significant features. B. Similarly as A., but a threshold of 0.5 is applied. C. and D. Recalls are plotted instead of precisions.

7.1.3 Comparison with elastic net

Barretina et al. (2012) and Garnett et al. (2012) applied elastic net regression to all samples. To evaluate the effect of considering context specificity, we applied elastic net to the synthetic data. For elastic net, 10-fold cross-validation is used to choose the parameters that optimize mean square errors (Barretina et al., 2012). The Matlab implementation of elastic net, *lasso*, is used here . The range of 0.2 to 1 with increment of 0.1 is used for selecting the parameter *alpha*, which controls the ratio of L1 to L2 norm penalty (*alpha=1* means no L2 norm penalty is used). The second parameter *lambda,* which controls sparsity, is chosen from the smallest to the largest possible values that give non-empty models. 100 bootstrapped runs of elastic net are applied to the each synthetic phenotype with the optimal parameters. All seven possible subtypes are treated as binary features and included for model training. For evaluation, we calculate precision and recall for elastic net models. Since elastic net does not select subtype-specific features, we simply evaluate the models regardless of subtype specificity.

Figure 7-4 shows the comparison of precision and recall obtained from the split-regression algorithm and elastic net. Although many models from the first iteration of the split-regression algorithm have zero precision (Figure 7-4A), most of them outperform elastic net after the second iteration (Figure 7-4B). However, in exchange for high precision and hence low false positive rate, the split-regression is more conservative than elastic net. As a result, its recall is not as good as elastic net (Figure 7-4C and D). This is also due to the fact that elastic net has the capability to select collinear features. Since the features used including gene expression, highly correlated features are expected. The results from Figure 7-4 suggests elastic net selects many more correlated features than the split-regression algorithm to achieve high recall, and therefore suffers from low precision.

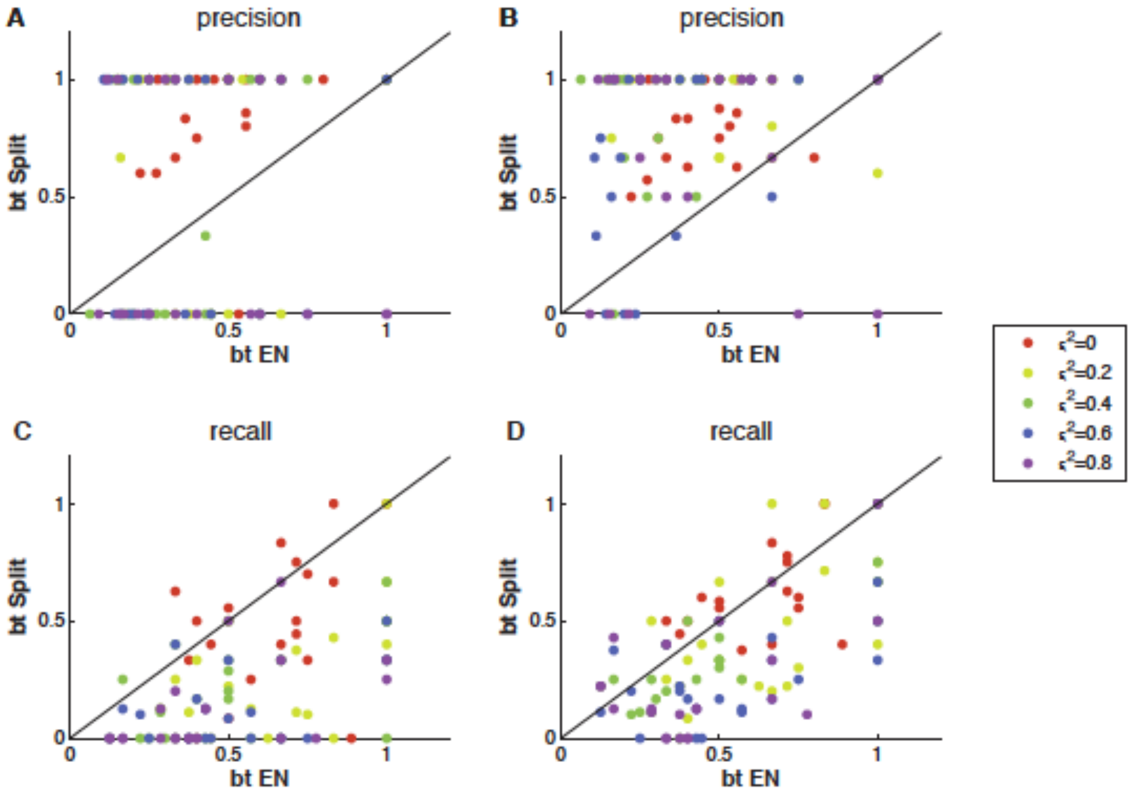

Figure 7-4: Comparison between *split-regression* and elastic net.
Bootstrapped elastic net (*bt EN*) is compared to bootstrapped split-regression (*bt Split*). A threshold $f \geq 0.5$ is applied to call significant features. A. The precision of each phenotype from *bt EN* (x-axis) is plotted against that from *bt Split* after the first iteration (y-axis). Each dot represents a phenotype, colored by the noise level added. B. As in A, but the precisions of *bt Split* are from the second iteration. C. and D. The recalls from both algorithms are plotted in the same manner, with the data of *bt Split* from the first iteration in C. and that from the second iteration in D.

The results from the simulation validate the correctness and robustness of the proposed algorithm. It also sheds lights on the effects of bootstrapping and transfer learning. These results suggest the algorithm is relatively conservative due to the sparsity constraints. This is important since false positives can be costly in biology, especially when we want to pursue experiments to validate the roles of the selected features.

However, the split-regression algorithm also suffers a similar issue as L1-norm regularized regression *lasso*. They cannot handle collinear features well. Correlated features would compete with each other during bootstrapping and decreases their relevant frequency to the phenotype. This argues for a lenient threshold for the relevant frequency. We will discuss this in the next section when the algorithm is applied to the real data.

7.2 Predictive Modelling for Drug Sensitivity from CCLE Data

7.2.1 Data Processing

Three sets of data are extracted from the CCLE collection. The first set includes breast carcinoma (n=27) and ovary carcinoma (n=25) samples because of the genomic similarities between basal-like breast cancers and high-grade serous ovarian cancers (Cancer Genome Atlas Research, 2011). The second set includes malignant melanoma (n=38) and glioma (n=25) samples. Melanocytes and neuroglia are both embryologically derived from ectoderm. Shared tumor-associated antigens (Chi et al., 1997) and dysregulated pathways (Prickett and Samuels, 2012) have been reported in melanoma and glioma. In addition, high similarity between samples of central nervous system and skin tissues is shown in the projection of samples on principal components derived from gene expression profiles (Figure 7-5). Therefore, it is possible that these two cancers share some biological pathways or genomic features contributing to drug sensitivity. For these two datasets, only one possible “split” is allowed, as we consider the cancer type provides context governing the underlying mechanism of drug sensitivity. We refer to these two sets as CCLE-BreastOvary and CCLE-SkinGlioma hereafter.

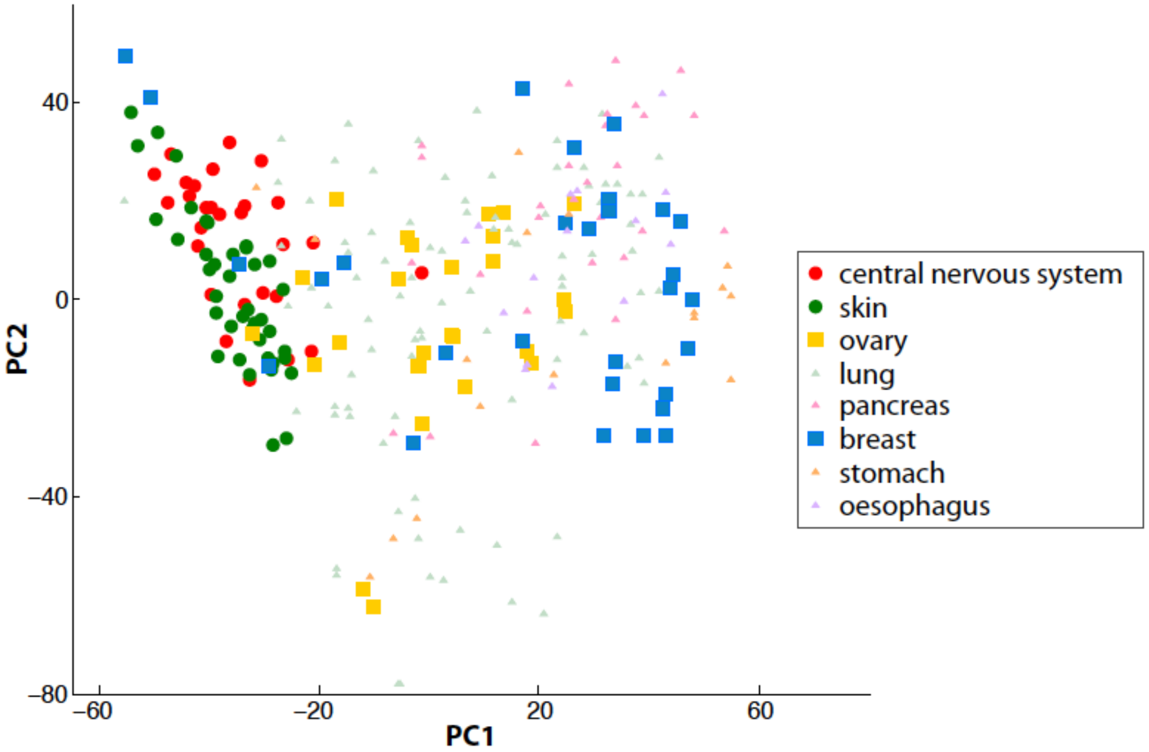

Figure 7-5: Similarity of gene expression profile between central nervous system and skin samples.

Samples are projected to the first two principal components (PC1 and PC2) using principal component analysis with gene expression profiles. Samples of other tissues are omitted in the figure for clarity.

Table 7-1: Sample classification used as subtypes in CCLE-Blood.

Subtype	No. samples	Cancer types included
ALL	10	acute lymphoblastic leukaemia, B or T cell
AML	10	acute myeloid leukaemia
NHL	28	non-Hodgkin lymphoma
NHL-B	22	non-Hodgkin lymphoma (B-cell)
DLBCL	10	diffuse large b cell lymphoma
Plasma	14	plasma cell myeloma
Haema	15	haematopoietic neoplasm

Each subtype is defined by a binary classification. Only samples of the *cancer types included* are treated as samples of the subtype. Subtypes are not mutually exclusive. *Haema* subtype is defined for samples of haematopoietic neoplasm, as opposed to samples of lymphoid neoplasm.

The third set of samples extracted from CCLE includes cell lines derived from haematopoietic and lymphoid tissues (n=70). This set includes samples of many different subtypes, including ALL (acute lymphoblastic leukemia), AML (acute myeloid leukemia), CML (chronic myeloid leukemia), plasma cell myeloma, and etc. Seven binary classifications are defined for samples and treated as potential relevant subtypes for the analysis (Table 7-1). This set of data is designated as CCLE-Blood in the following text.

Processed mutation and copy number data of the cell lines are obtained from the CCLE website (http://www.broadinstitute.org/ccle/home). Raw data of gene expression microarray are processed as follows. Samples are processed together with MAS5 algorithm (Hubbell et al., 2002). Probes that are annotated to hybridize to multiple genes are discarded. Signals are then log transformed, and then we discard probes that have low

(< 6) or high (> 15.5) values in more than 20% of the samples in a dataset. Principal component analysis is applied to z-scores for the remaining probes. Outlier samples are identified from the projection of samples on the first three principal components. Log-transformed data are then normalized according to the 75th percentile within each sample. Finally, values of probes hybridized to the same gene are averaged to derive expression of the gene.

Since samples may be very resistant to the drug at the screened concentration in the CCLE data, many samples did not reach 50% inhibition of growth (Barretina et al., 2012) and hence do not have valid IC50. Therefore, when more than 80% of samples do not have valid IC50 of a drug, the IC50 phenotype of the drug is excluded from the analysis. IC50 values are then log transformed and treated as phenotype for modeling. Activity area (ACT) above the growth curve is treated as another phenotype associated with the drug sensitivity.

To limit the search space of predictors, we compile lists of genes associated with each cancer from literature and the Disease database (Frankild and Jensen). Only mutation, copy number and gene expression of the genes associated with the analyzed cancers are included as potential predictors. Moreover, mutation data of genes with <10% mutated samples or wild types are excluded. For gene expression, genes with variance < 0.2 are filtered. For copy number, genes with variance < 0.15 are filtered. Table 7-2 summarizes the number of phenotypes, features, and available samples in each dataset.

Table 7-2: Summary of datasets for analysis.

Dataset	No. phenotypes	No. features	No. samples
CCLE-BreastOvary	39	492	35-52
CCLE-SkinGlioma	33	248	45-63
CCLE-Blood	44	278	50-70

No. samples is a range since the availability of phenotype data depends on each drug.

7.2.2 Evaluation of Predictive Models

We applied *split-regression* to the datasets CCLE-SkinGlioma, CCLE-BreastOvary, and CCLE-Blood. The details of data processing are described in Section 7.2.1. For each dataset, ACT (activity area above drug response curve) and log-transformed IC50 (concentration at which 50% inhibition occurs) are used as phenotypes. Drugs that are inactive or induce small variation in response are filtered. Mutation, copy number and gene expression data of a pre-compiled list of genes are used as features to predict the phenotypes (see Section 7.2.1).

Ten-fold cross-validation is used to evaluate the prediction. During each "fold", 9/10 of samples are used to learn predictive models for the phenotype, and predictions are made for the 1/10 of samples held out during training. The procedure is repeated ten times so that predictions for all samples are obtained. For *split-regression*, features are selected through 10 iterations of transfer learning and 100 bootstrapping runs (Section 6.3).

Features with relevant frequency ≥ 0.3 after each iteration of transfer learning are used to build regression model to predict the test samples. This allows us to evaluate the performance of the algorithm at the end of each iteration. The use of the lenient threshold 0.3 is to compensate for lack of ability to include collinear features. Ridge regression is used to obtain regression coefficients of the selected features (Eq. 7.1).

For comparison, elastic net models are evaluated with the samples in the abovementioned ten-fold cross-validation. The subtypes used in split-regression are included as features. The two parameters of elastic net are selected within each "fold" using a nested ten-fold cross-validation optimizing mean square errors. 100 bootstrapping runs are also applied using the optimal parameters to select the final set of features. Same threshold for the relevant frequency is used to obtain the features. Predictions of test samples are made using models obtained by ridge regression (Eq. 7.1).

Both models are evaluated with Pearson and Spearman correlation coefficients between the predictions and the phenotype data. When an algorithm fails to select any feature for a phenotype, the average of training data is used as prediction.

First we observe the effects of transfer learning in split-regression (Figure 7-6). Predictions are improved through iterations, especially between the first two iterations. This is because uniform prior is used in the first iteration and the algorithm fails to obtain models for many phenotypes. However, the performance is improved through the following iterations. As the information of models is exchanged between similar phenotypes, priors of features being relevant are adjusted. The algorithm starts to obtain relevant features, and the correlation coefficients are improved. The improvement generally stops after four to five iterations.

Figure 7-7 to Figure 7-9 shows the comparisons between split-regression and elastic net. Predictions from elastic net correlate better with the phenotypes than those from the first iteration of split-regression. However, after transfer learning, split-regression outperforms elastic net for most phenotypes in CCLE-Blood (Figure 7-7) and CCLE-SkinGlioma datasets (Figure 7-8), although the performances of the two algorithms are comparable for CCLE-BreastOvary data (Figure 7-9).

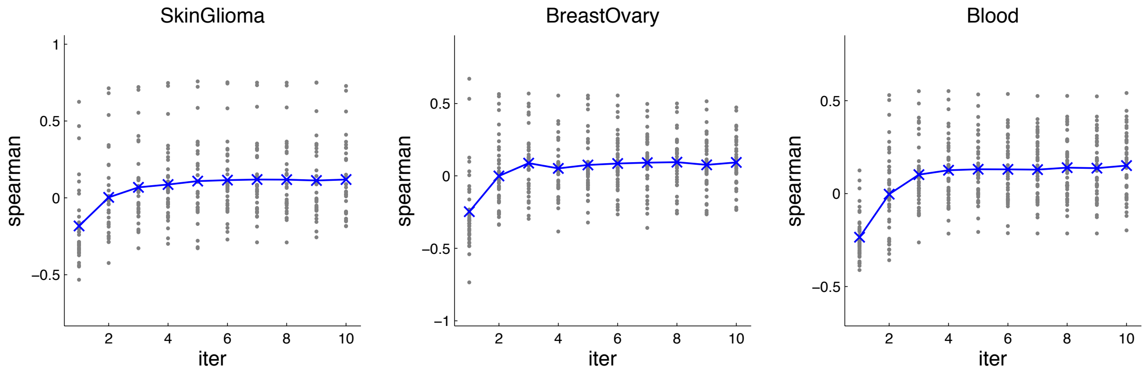

Figure 7-6: Effect of transfer learning.
Spearman's correlation coefficients between predictions and phenotypes are plotted as a function of iterations. Blue dots illustrate the trend of mean performance, and grey dots represent the performance of each phenotype in each iteration.

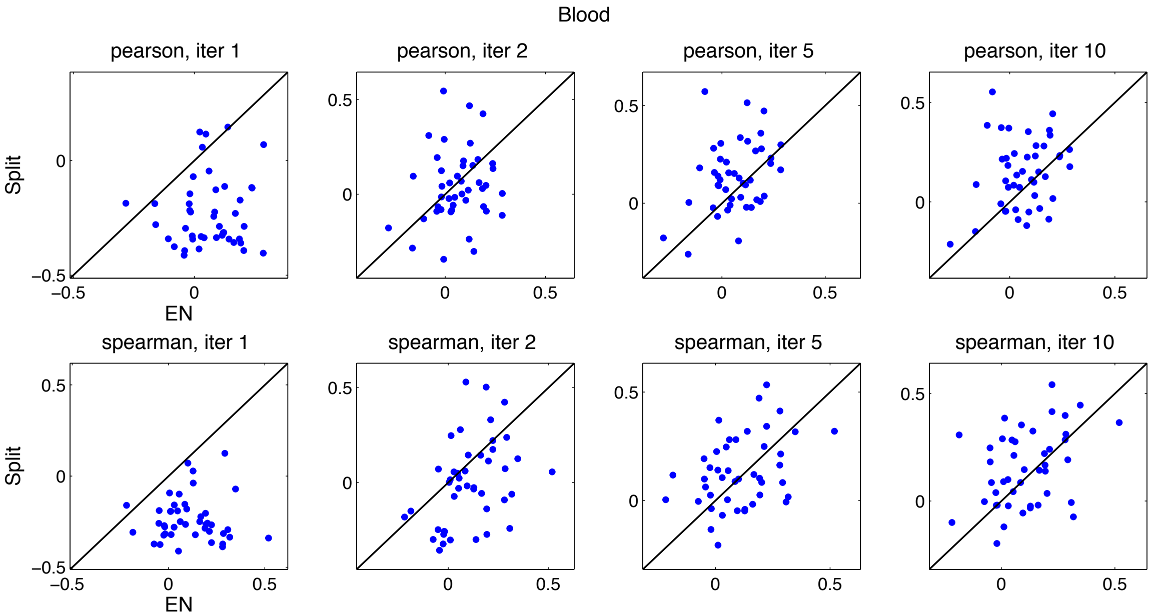

Figure 7-7: Comparison between split-regression and elastic net using CCLE-Blood data.

Predictions are obtained via 10-fold cross-validation. Each dot represents the prediction score of a phenotype from elastic net (x-axis) and split-regression (y-axis). Predictions from the 1st, 2nd, 5th, and 10th iterations of split-regression are compared to elastic net.

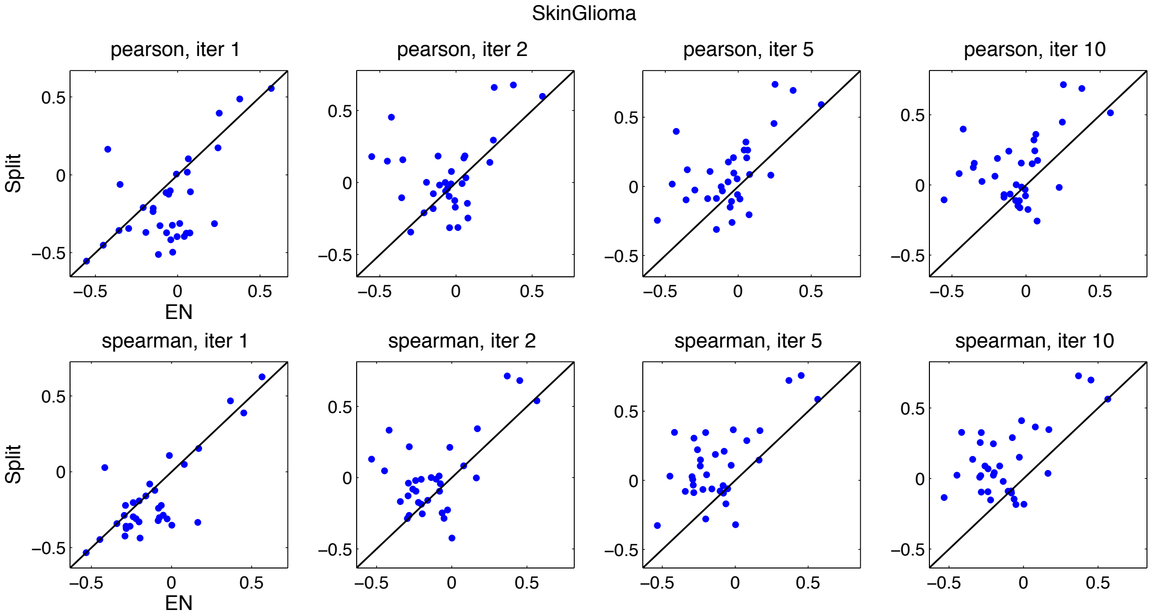

Figure 7-8: Comparison between split-regression and elastic net using CCLE-SkinGlioma data.

Annotations are the same as Figure 7-7.

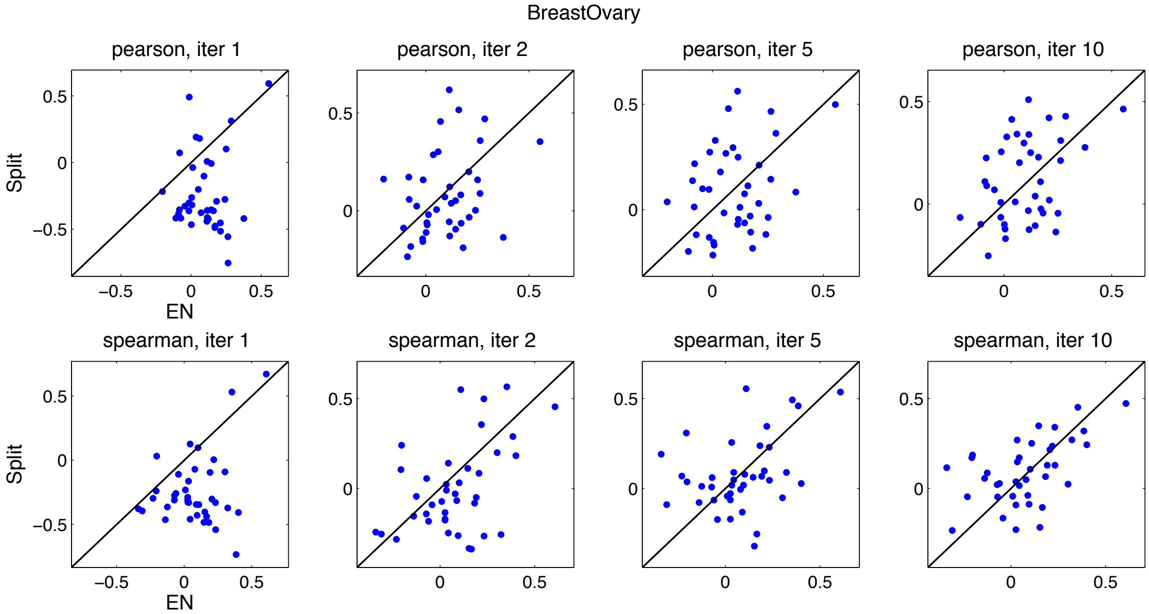

Figure 7-9: Comparison between split-regression and elastic net using CCLE-BreastOvary data.
Annotations are same as Figure 7-7.

The number of features selected by each algorithm differs significantly. Because elastic net is capable to select collinear predictors, it chooses many more features to make predictions than split-regression. It may seem that the threshold (0.3) for relevant frequency is too lenient for elastic net. On the other hand, collinear features compete with each other during bootstrapping runs of split-regression. The algorithm is mostly likely to select one among the collinear features. In order to compensate this, the threshold 0.3 is therefore a reasonable choice for split-regression.

To see how the number of features affects the performance of both models, we control the number of features selected in elastic net to be similar as that of split-regression. The features are chosen according to their relevant frequency in bootstrapping runs. Figure 7-10 shows the comparison of prediction performances when the number of features is controlled to be similar in the two algorithms. We observe that elastic net performs slightly better than before when model size is controlled. This corresponds to a more stringent threshold than 0.3 for elastic net. However, we can still see that many models in CCLE-Blood and CCLE-SkinGlioma still outperform elastic net in prediction.

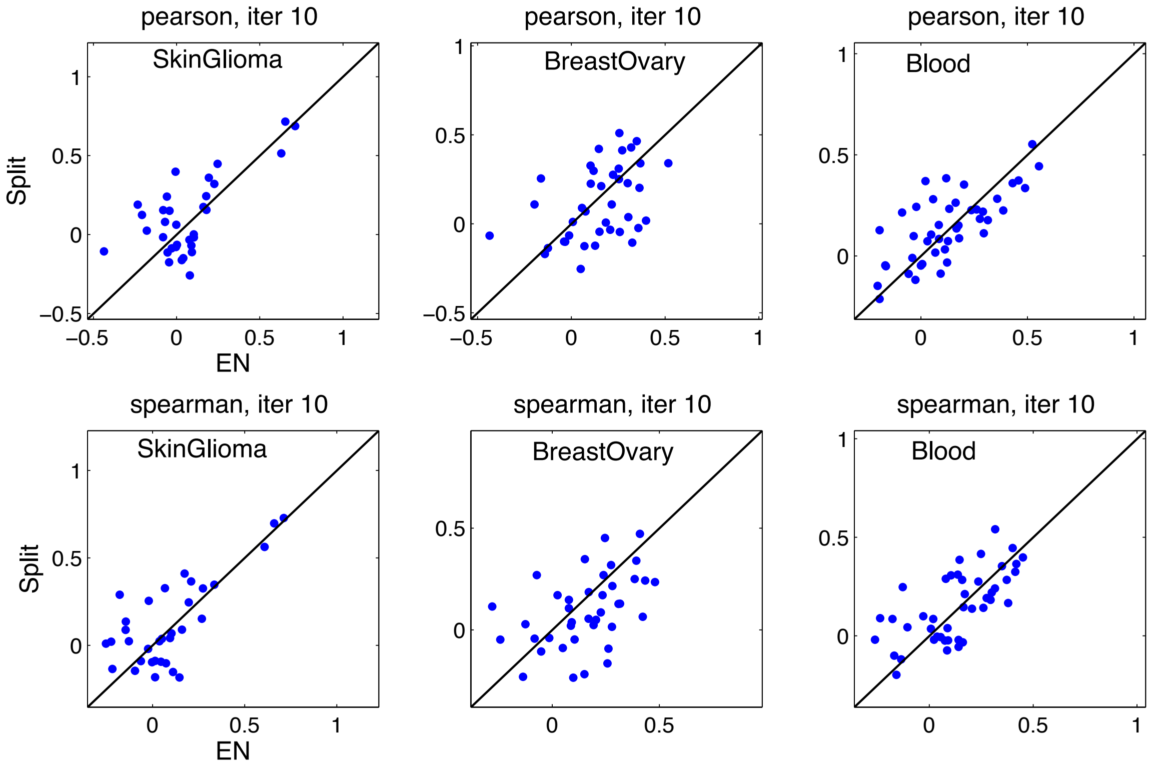

Figure 7-10: Comparison between split-regression and elastic net when model size is controlled.

Elastic net is controlled to have similar number of features as split-regression. Annotations are same as Figure 7-7.

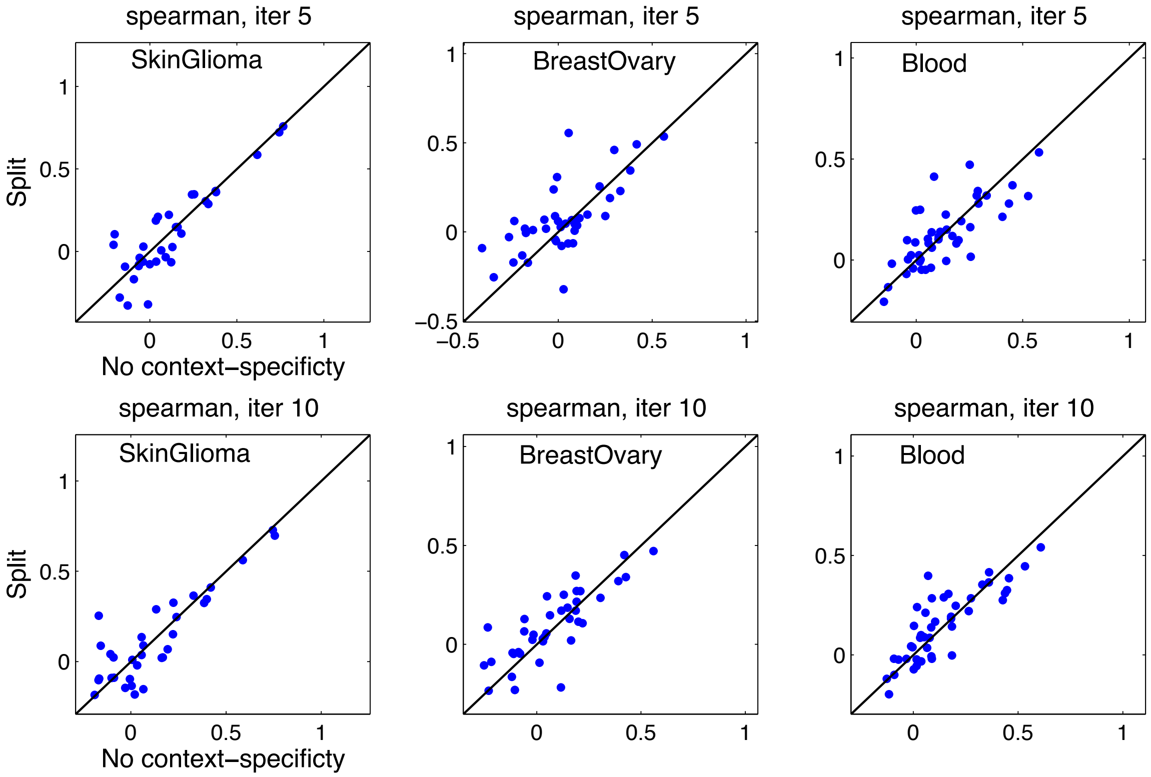

Figure 7-11: Influence of context specificity.
Features from split-regression are obtained and used as shared features for all samples regardless of context. Annotations are similar as Figure 7-7, except x-axis corresponds to models with consideration of context.

Finally, we remove the influence of context from split-regression models. The chosen split, *i.e.* cancer type, is included as a predictor and context-specific features are treated as shared features. Figure 7-11 shows the comparison between models with and without considering context. Predictions of many phenotypes benefit from context-specific models. However, when comparing the 5th to the 10th iterations for CCLE-BreastOvary, we see the prediction accuracy decreases in some cases for the final iteration. This suggests the algorithm include false context-specific features for CCLE-BreastOvary after a few iterations. This is similar to the decrease of precision in our simulation after too many iterations are passed (Figure 7-2).

7.3 Discussion

Both the simulation and results from CCLE data elucidate the importance of context specificity in cancer pharmacogenomics. We propose an algorithm, *split-regression,* to include context specificity in linear regression. Split-regression is a principled approach, based on both information theory and Bayesian statistics. The employed regularization enforces a strong constraint on sparsity. Indeed, we show split-regression is more conservative than elastic net.

The application to CCLE data shows how important context can affect model learning. Using split-regression, we can pool samples of similar cancer types and learn both shared features and cancer-type specific features. This not only sheds lights on the common genomic features conferring phenotypic variation across cancer types, but may also reveal the relevant subtype governing the context specificity. For example, the relevant context is unclear in CCLE-Blood where seven categorizations of samples are defined. Split-regression is designed to uncover the relevant subtype, or split, that dictates context-specific predictive programs for drug sensitivity. We will see examples of this in the next chapter.

Chapter 8 Context Specificity of Drug Sensitivity in Cancer

Identifying biomarkers to predict prognosis or response to treatments can improve health care and enhances our understanding of the biological mechanism underlying the efficacy of drugs. In this chapter, we discuss the predictive features that are associated with a few highlighted drugs and their implication in drug sensitivity between cancer types.

Here we run split-algorithm on the same datasets using all samples. The final model is obtained from the last (10th) iteration and a threshold of 0.3 is applied to the relevant frequency to determine if a predictor is significant. Description of data and features used are described in Section 7.2.1.

8.1 Expression of *JAK3* and *AKT1* Are Predictive of Sensitivity to HDAC Inhibitor Panobinostat in Haematological Malignancy

Panobinostat, a histone deacetylase inhibitor (HDAC), is a potential treatment for haematological malignancy (Deangelo et al., 2013; Lemoine et al., 2012; Lemoine and Younes, 2010; Prince et al., 2009). Recently, a study showed that panobinostat inhibits JAK/STAT pathway activity in Hodgkin lymphoma cell lines, and that the reduced pathway activity may lead to cell death (Lemoine et al., 2012). However, the study also showed that panobinostat treatment leads to activation of mTOR (mammalian target of rapamycin) via reducing phosphorylation levels of AMP-activated protein kinases[2].

Our algorithm suggests copy number of *JAK2* (Janus kinase) as a predictor of sensitivity (IC50) to the drug for all CCLE-Blood samples. Moreover, the algorithm chooses the split “plasma cell myeloma or not” to separate the samples into two groups, where high expression of *JAK3* predicts high sensitivity in myeloma samples; and high expression of *AKT1* (v-akt murine thymoma viral oncogene homolog 1) and *VAPA* (vesicle-associated membrane protein-associated protein A) predict increased sensitivity in non-myeloma samples (Figure 8-1). The selection of *JAK2/3* and *AKT1* suggests the roles of JAK/STAT and PI3K pathways in the response induced by panobinostat.

Interestingly, for the ACT of panobinostat, the algorithm chooses another split “non-Hodgkin lymphoma (B-cell) or not” (NHL-B). Expression of *CDK2* (cyclin-dependent kinase 2) is selected as a predictor only for non-Hodgkin lymphoma (B-cell), whereas expression of *DOT1L* (DOT1-like, histone H3 methyltransferase) is selected for the rest of samples (Figure 8-1).

The predictive power of *DOT1L* suggests an association between histone methylation and inhibition of HDAC. *DOT1L* plays a key role in leukemia associated with *MLL* (mix lineage leukemia) fusions (Bernt et al., 2011; Chi et al., 2010; Slany, 2009). The product of *MLL* fusion genes can recruit DOT1L, which leads to aberrant histone methylation

[2] adenosine monophosphate-activated protein kinase.

marks and drives leukemogensis (Guenther et al., 2008; Krivtsov et al., 2008; Nguyen et al., 2011). Recently a research group used connectivity map (Lamb et al., 2006) to identify panobinostat as a treatment for *MLL*-rearranged infant acute lymphoblastic leukemia (Stumpel et al., 2012). Surprisingly, the HDAC inhibition restored DNA methylation in the promoters of selected proto-oncogenes and down-regulated the expression of both fused-*MLL* and the proto-oncogenes. In our model, high expression of *DOT1L* predicts increasing sensitivity to panobinostat for non-NHL-B samples, suggesting those with high level of *DOT1L* expression have increased histone methylation, which could be restored by panobinostat.

For the drug response to panobinostat, our models imply two possible classifications of samples that govern the underlying key predictors that are associated with the drug mechanism. Each of these models indicates the context of predictability for the drug response in a cancer-type or cell-type specific manner.

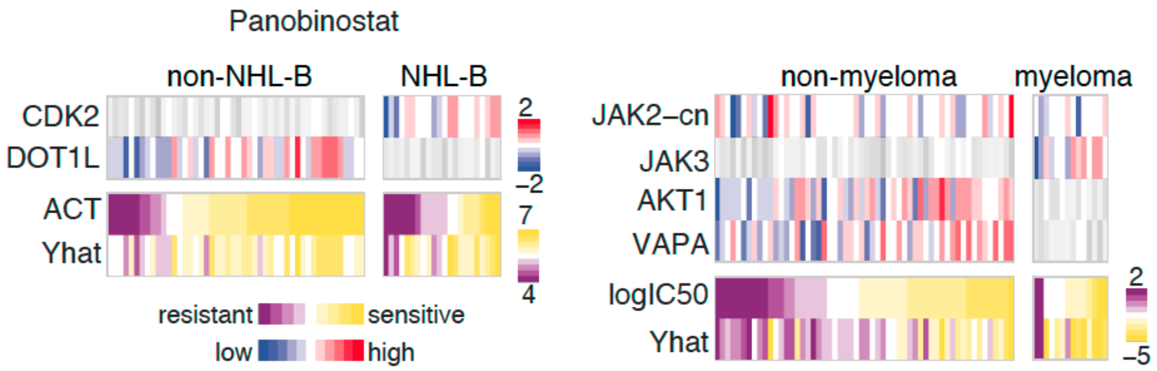

Figure 8-1: Predictive features of sensitivity to panobinostat in CCLE-Blood. Split-regression identifies two different splits of samples for ACT and IC50 phenotypes, respectively. All features are gene expression profiles except *JAK2,* which is a copy number feature. Yhat is the prediction from leave-one-out procedure, where all samples but one are used to obtain regression coefficients. The coefficients are then used to make prediction for the held-out sample. It is repeated for prediction of all samples.

8.2 Gene Expression Suggests Activities of PI3K, MAPK and NF-κB Pathways Are Predictive of Sensitivity to Paclitaxel in Melanoma and Glioma Cell Lines

Paclitaxel targets tubulin and stabilizes the microtubule, leading to defect in cell division. Sensitivity to paclitaxel has been associated with PI3K (Bava et al., 2011; Nguyen et al., 2004), MAPK (Bacus et al., 2001; Boldt et al., 2002) and NF-κB pathways (Mabuchi et al., 2004; Nguyen et al., 2004; Patel et al., 2000). Both melanoma and glioma cell lines in CCLE show a wide range of sensitivity to paclitaxel (ACT ranges from 3 to 7). Split-regression selects both shared and cancer type-specific features for the phenotype ACT of paclitaxel. Expression of both *AKT1* and *WT1* (Wilms tumor 1) are selected as shared predictors (Figure 8-2). Interestingly, expression of *DUSP14* (dual specificity phosphatase) is selected for only glioma samples whereas mutation of *PTEN* (phosphatase and tensin homolog), and expression of *DUSP6* and *USP6* (ubiquitin specific peptidase 6) are selected for melanoma samples. The selection of *AKT1* and *PTEN* suggests the regulation of PI3K/AKT pathway may be predictive of resistance to paclitaxel in melanoma cells. Additionally, transcription of *DUSP6* (dual specificity phosphatase 6) is regulated by ERK (Bermudez et al., 2010; Patterson et al., 2009). Expression of *DUSP6* is therefore a predictor of MAPK activity. Its overexpression has been reported in melanoma with *BRAF*^{V600E} mutations (Bloethner et al., 2005). Recent studies also associate expression of *DUSP6* with sensitivity to the chemotherapy agent cisplatin (Li and Melton, 2012) and the antidiabetic drug metformin (Martin et al., 2012). Our algorithm selects *DUSP6* as a melanoma-specific predictor, suggesting aberrant MAPK activity may confer resistance to paclitaxel in melanoma.

Selection of *DUSP14* as glioma-specific feature is intriguing since *DUSP14* encodes an atypical DUSP (Patterson et al., 2009). *DUSP14* modulates *ERK* activity and negatively regulates proliferation in pancreatic β-cell (Klinger et al., 2008). In our model high expression of *DUSP6* and *DUSP14* is predictive of resistance to paclitaxel in melanoma and glioma, respectively. Both *DUSP6* and *DUSP14* suggest high levels of MAPK activity may be involved in paclitaxel resistance. Indeed, several studies have shown that paclitaxel-induced apoptosis depends on the MAPK pathway. Inhibition of ERK activity could increase sensitivity to paclitaxel in many cancers including colon, lung, and cervical cancers (Bava et al., 2011; Nguyen et al., 2004).

WT1, Wilm's tumor 1, is a transcription factor playing roles in regulation of growth and apoptosis. It has been implicated as both an oncogene and a tumor suppressor in many cancers (Hartkamp et al., 2010; Wagner et al., 2003). The aberrant expression of *WT1* is associated with proliferation in both melanoma (Wagner et al., 2008; Zamora-Avila et al., 2007) and glioma cell lines (Clark et al., 2007; Clark et al., 2010). Intriguingly, *WT1* negatively regulates transcription of *IGF-1R* (insulin-like growth factor I receptor) and activity of IGF-1R signaling can in turn activate MAPK and PI3K pathways (Pollak, 2008). From our model, active IGF-1R signaling is inferred from low expression of *WT1*, which predicts resistance to paclitaxel. Indeed, inhibition of *IGF-1R* has been implicated in sensitivity to paclitaxel in non-small cell lung cancer (Spiliotaki et al., 2011) and breast cancer (Beech et al., 2001).

On the other hand, predictors for the IC50 phenotype of paclitaxel include the expression of *NOTCH1* (shared), *CXCR4* (shared), *PDGFRA, PI3KCA1* and *REL* (melanoma-specific) (Figure 8-2). Close inspection suggests *CXCR4* may be a glioma-specific predictor. It is chosen as a shared or glioma-specific feature with similar frequency (0.35 for shared and 0.33 for glioma). The expression of *CXCR4* (chemokine receptor 4) has been associated with invasiveness in glioma (Ehtesham et al., 2013) and resistance to paclitaxel in stem-cell-like glioblastoma (Liu et al., 2006). The selection of *PDGFRA* and *PI3KCA1* again suggests the association between PI3K and the toxicity of paclitaxel in the two cancers. Moreover, *NOTCH1* and *REL* (member of NF-κB) implicate the role of NF-κB in sensitivity to the drug. High expression of both *NOTCH1* and *REL* are associated with increasing sensitivity in melanoma samples.

Together our model suggests PI3K, MAPK and NF-κB pathway activities are predictive of sensitivity to paclitaxel in melanoma and glioma, with different genes in the pathways modulating the responses in each cancer. It also suggests that complex mechanisms and cross-talk between these pathways could underlie the resistance to paclitaxel toxicity.

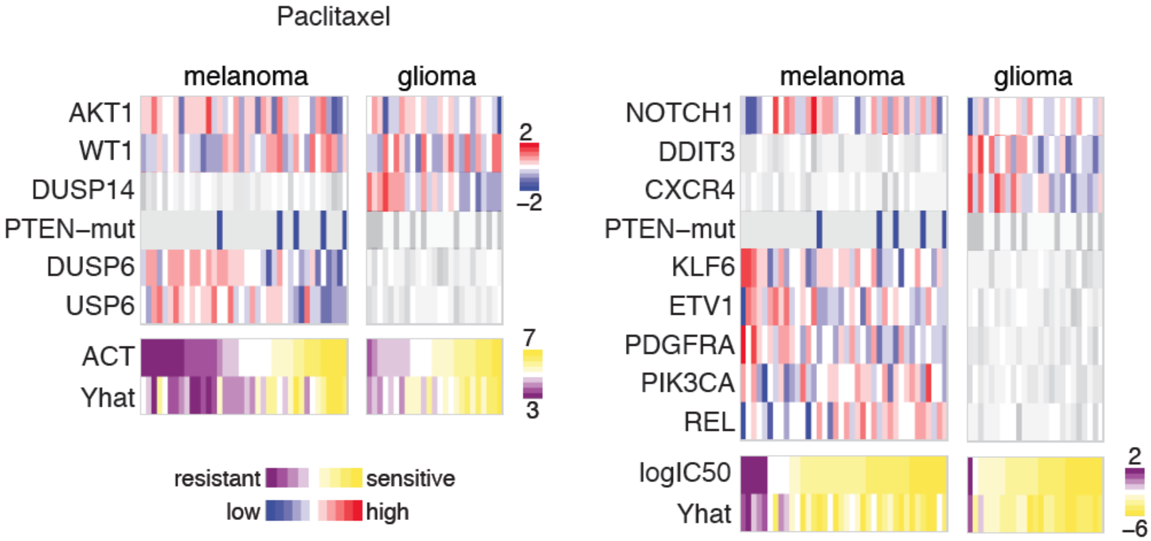

Figure 8-2: Predictive features of sensitivity to paclitaxel in CCLE-SkinGlioma. Similar as Figure 8-1, all features are gene expression profiles except *PTEN* mutation.

8.3 Resistance Predictor *WT1* May Suggest Combined Therapy

Resistance to therapy is a major challenge for cancer treatment. Often resistance to drugs is associated with aberrant pathway activities. As several genes in our models seem to serve as proxies to pathway activity, we ask if the association between expression of genes and drug resistance may suggest combinatorial treatments. The assumption is, if a gene is predictive of hyper-activation of an oncogenic pathway and is associated with resistance to a drug, inhibition of the pathway may overcome the resistance and sensitize individuals to the original therapy.

In CCLE-SkinGlioma dataset, low expression of *WT1* is associated with resistance to paclitaxel, topotecan, and irinotecan for all samples. All three drugs are cytotoxic agents used in chemotherapy. Both topotecan and irinotecan target DNA topoisomerases. Since low expression of *WT1* implies high activation of IGF-1R, inhibition of IGF-1R may overcome the resistance to these chemotherapeutic agents. Indeed, beneficial effects of combining IGF-1R inhibitors with paclitaxel has been reported in non-small cell lung cancers (Goto et al., 2012; Gualberto et al., 2011; Spiliotaki et al., 2011), with topotecan in Wilms tumor cells (Bielen et al., 2012), and with irinotecan in colon cancers (Calzone et al., 2013; Flanigan et al., 2010). For glioma, inhibition of IGF-1R is shown to enhance the sensitivity to etoposide, which also targets topoisomerase (Yin et al., 2005).

8.4 Discussion

Context specificity plays an important role in biology. It allows cells with the same genetic information to differentiate into diverse cell types for various physiological functions. In cancer, context specificity may influence response to treatment. For example, a mutation in BRAF is observed among different tumors, but BRAF-targeted therapy does not work in all cancers (Corcoran et al., 2012; Kopetz et al., 2010; Prahallad et al., 2012). This is because the regulatory programs and signaling pathways have contextual activities that assert a strong influence on phenotype.

Context presents an even greater challenge for understanding tumor biology and treatment when we consider that multiple subtypes are often found in the “same” cancer type. For example, subtypes of breast cancer can have very different response to the same treatment (Heiser et al., 2012). The complication of context gets exacerbated when signaling pathways are rewired in different cancers. This may explain why often a potent therapy works for some patients but fails to alleviate tumor progression in others. These challenges emphasize the need for personalized medicine: treatments are tailored specifically for a patient sample in the context of its genomic profile.

Current efforts to screen drug sensitivity of large collections of cancer cell lines (Barretina et al., 2012; Garnett et al., 2012) provide a model system for personalized medicine. However, since the number of samples of each cancer type is limited, computational analysis either has to neglect context specificity or is underpowered by sample size. Furthermore, for cases where the relevant subtypes are not clear, we cannot separate samples into groups for model learning; a naïve approach would have to ignore contextual effects.

In order to tackle this problem, we propose an algorithm that explicitly accounts for context specificity. The algorithm aims to identify predictive features that are shared among all samples, and features that are context-specific. For data where the relevant subtype is unknown, the algorithm seeks the relevant classification of samples, if any, to provide context specificity. Hence, we can pool samples of similar cancer types together, benefitting from the sample size to uncover shared features and meanwhile searching for context-specific features to predict drug sensitivity.

Our results demonstrate the importance of context. We show that dependence of drug sensitivity on MAPK pathway activity can manifest in expression of distinct genes in melanoma and glioma. In CCLE-Blood data where the relevant subtype for drug sensitivity is unknown, split-regression learns the relevant classifications of samples to reveal the effects of context. Interestingly, the NHL-B (non-Hodgkin lymphoma, B-cell) is the relevant subtype chosen for most phenotypes, followed by NHL (non-Hodgkin lymphoma) and Plasm (plasma cell myeloma). Table 8-1 lists the frequency of each classification chosen as the relevant subtype. The result may suggest different mechanisms and drug responses in the lineages of haematopoietic malignancy.

Table 8-1: Number of phenotypes associated with subtypes.

Sample classification	**Number of phenotypes associated**
NHL-B	15
NHL	7
Plasm	7
AML	6
Haema	5
DLBCL	3
ALL	1

Transfer learning in our algorithm not only boosts the learning power, but may also help reveal relations between drugs and pathways. In CCLE-Blood data, expression of *CRKL* is selected as a predictor to response to 17-AAG (target: HSP90), erlotinib (EGFR), L-685458 (gamma-secretase), LBW242 (XIAP), lapatinib (EGFR), and nilotinib (ABL). CRKL is a substrate of BCR-ABL and plays an important role in leukemia (Hamilton et al., 2006; Sattler and Salgia, 1998). The Pearson's correlation coefficients between the ACT of these drugs range from 0.24 to 0.54, suggesting different degrees of similarity between them. It also implies the responses to these drugs may be mediated by the activities of EGFR, Notch, and BCR-ABL signaling pathways and perhaps the cross-talk between them. Interestingly, treatments combining nilotinib or17-AAG with imatinib (Gleevec, targeting BCR-ABL) have shown some synergistic effects in leukemia patients (Gorre et al., 2002; Hawkins et al., 2005; Radujkovic et al., 2005; Weisberg et al., 2007).

The association between *WT1* and chemotherapeutic agents indicates that IGF-1R pathway activity underlies the drug response to those compounds. Indeed, distributed pathway activities have been associated with drug resistance, even in the case of targeted therapy (Rosenzweig, 2012). It will be interesting how one can use the predictive model to not only suggest the efficacy of a drug, but also recommend potential combinatorial treatment.

Our split-regression algorithm currently only supports one split, that is, only two groups of samples in a model. However, the heterogeneity and similarity between samples may not just be limited to such simplified binary classifications. While we can increase the number of splits in a model, the complexity and search space also increase tremendously. Nevertheless, since the algorithm is based on information theory, implementation with heuristic and greedy search will allow flexible handling of such complex models. It is therefore feasible to extend the algorithm to consider multiple subtypes in a model, if the number of available samples provides sufficient statistical power.

New technologies and data collection give us the prospect of pharmacogenomics, but the complexity of the human genome, inter- and intra-cellular interactions present great challenge. In this thesis, we present two complimentary methods for modeling phenotypic variation with genomic features. Both methods are shown to be powerful to take on difficult problems and reveal interesting biological insight into drug sensitivity. These methods demonstrate the principles of computational modeling for drug sensitivity, highlight the challenges of personalized medicine and indicate potential solutions.
As various functional and genomic data such as siRNA screening (Cheung et al., 2011; Luo et al., 2008) are made available, our methods can be easily extended to integrate information from different data types. We believe such computational modeling will become more common, and will provide the foundation for the future of personalized medicine.

References

Airoldi, E.M., Huttenhower, C., Gresham, D., Lu, C., Caudy, A.A., Dunham, M.J., Broach, J.R., Botstein, D., and Troyanskaya, O.G. (2009). Predicting cellular growth from gene expression signatures. PLoS Comput Biol *5*, e1000257.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X.*, et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503-511.

Allegra, C., Jessup, J., Somerfield, M., Hamilton, S., Hammond, E., Hayes, D., McAllister, P., Morton, R., and Schilsky, R. (2009). American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. J Clin Oncol *27*, 2091-2096.

Altelaar, A., Munoz, J., and Heck, A. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. Nature reviews Genetics *14*, 35-48.

Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. Science *322*, 881-888.

Amado, R., Wolf, M., Peeters, M., Van Cutsem, E., Siena, S., Freeman, D., Juan, T., Sikorski, R., Suggs, S., Radinsky, R.*, et al.* (2008). Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol *26*, 1626-1634.

Bacus, S., Gudkov, A., Lowe, M., Lyass, L., Yung, Y., Komarov, A., Keyomarsi, K., Yarden, Y., and Seger, R. (2001). Taxol-induced apoptosis depends on MAP kinase pathways (ERK and p38) and is independent of p53. Oncogene *20*, 147-155.

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. Nat Rev Genet *7*, 781-791.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A., Kim, S., Wilson, C., Lehár, J., Kryukov, G., Sonkin, D.*, et al.* (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603-607.

Barrett, J., Hansoul, S., Nicolae, D., Cho, J., Duerr, R., Rioux, J., Brant, S., Silverberg, M., Taylor, K., Barmada, M.*, et al.* (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature genetics *40*, 955-962.

Bava, S., Sreekanth, C., Thulasidasan, A., Anto, N., Cheriyan, V., Puliyappadamba, V., Menon, S., Ravichandran, S., and Anto, R. (2011). Akt is upstream and MAPKs are downstream of NF-κB in paclitaxel-induced survival signaling events, which are down-regulated by curcumin contributing to their synergism. The international journal of biochemistry & cell biology *43*, 331-341.

Bean, J., Brennan, C., Shih, J.-Y., Riely, G., Viale, A., Wang, L., Chitale, D., Motoi, N., Szoke, J., Broderick, S.*, et al.* (2007). MET amplification occurs with or without T790M mutations in EGFR mutant lung tumors with acquired resistance to gefitinib or erlotinib. Proceedings of the National Academy of Sciences of the United States of America *104*, 20932-20937.

Beech, D., Parekh, N., and Pang, Y. (2001). Insulin-like growth factor-I receptor antagonism results in increased cytotoxicity of breast cancer cells to doxorubicin and taxol. Oncology reports *8*, 325-329.

Benner, S.A., Cohen, M.A., and Gonnet, G.H. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng *7*, 1323-1332.

Benvenuti, S., Frattini, M., Arena, S., Zanon, C., Cappelletti, V., Coradini, D., Daidone, M., Pilotti, S., Pierotti, M., and Bardelli, A. (2008). PIK3CA cancer mutations display gender and tissue specificity patterns. Human mutation *29*, 284-288.

Bermudez, O., Pagès, G., and Gimond, C. (2010). The dual-specificity MAP kinase phosphatases: critical roles in development and cancer. American journal of physiology Cell physiology *299*, 202.

Bernt, K., Zhu, N., Sinha, A., Vempati, S., Faber, J., Krivtsov, A., Feng, Z., Punt, N., Daigle, A., Bullinger, L.*, et al.* (2011). MLL-rearranged leukemia is dependent on aberrant H3K79 methylation by DOT1L. Cancer cell *20*, 66-78.

Bickel, P.J., and Li, B. (1977). Mathematical statistics. Paper presented at: Test (Citeseer).

Bielen, A., Box, G., Perryman, L., Bjerke, L., Popov, S., Jamin, Y., Jury, A., Valenti, M., Brandon, A., Martins, V.*, et al.* (2012). Dependence of Wilms tumor cells on signaling through insulin-like growth factor 1 in an orthotopic xenograft model targetable by specific receptor inhibition. Proceedings of the National Academy of Sciences of the United States of America *109*, 76.

Bild, A., Yao, G., Chang, J., Wang, Q., Potti, A., Chasse, D., Joshi, M.-B., Harpole, D., Lancaster, J., Berchuck, A.*, et al.* (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature *439*, 353-357.

Bissell, M., and Labarge, M. (2005). Context, tissue plasticity, and cancer: are tumor stem cells also regulated by the microenvironment? Cancer cell *7*, 17-23.

Bloethner, S., Chen, B., Hemminki, K., Müller-Berghaus, J., Ugurel, S., Schadendorf, D., and Kumar, R. (2005). Effect of common B-RAF and N-RAS mutations on global gene expression in melanoma cell lines. Carcinogenesis *26*, 1224-1232.

Boldt, S., Weidle, U., and Kolch, W. (2002). The role of MAPK pathways in the action of chemotherapeutic drugs. Carcinogenesis *23*, 1831-1838.

Bollag, G., Hirth, P., Tsai, J., Zhang, J., Ibrahim, P., Cho, H., Spevak, W., Zhang, C., Zhang, Y., Habets, G*., et al.* (2010). Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. Nature *467*, 596-599.

Bouvard, V., Zaitchouk, T., Vacher, M., Duthu, A., Canivet, M., Choisy-Rossi, C., Nieruchalski, M., and May, E. (2000). Tissue and cell-specific expression of the p53-target genes: bax, fas, mdm2 and waf1/p21, before and following ionising irradiation in mice. Oncogene *19*, 649-660.

Brem, R., and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences of the United States of America *102*, 1572-1577.

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

Calzone, F., Cajulis, E., Chung, Y.-A., Tsai, M.-M., Mitchell, P., Lu, J., Chen, C., Sun, J., Radinsky, R., Kendall, R*., et al.* (2013). Epitope-specific mechanisms of IGF1R inhibition by ganitumab. PLoS ONE *8*.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.

Cancer Genome Atlas Research, N. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

Cardoso, C.M., Custodio, J.B., Almeida, L.M., and Moreno, A.J. (2001). Mechanisms of the deleterious effects of tamoxifen on mitochondrial respiration rate and phosphorylation efficiency. Toxicology and applied pharmacology *176*, 145-152.

Caruana, R. (1998). Multitask learning (Springer).

Chellappa, R., Kandasamy, P., Oh, C.S., Jiang, Y., Vemula, M., and Martin, C.E. (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, B.J., Causton, H.C., Mancenido, D., Goddard, N.L., Perlstein, E.O., and Pe'er, D. (2009). Harnessing gene expression to identify the genetic basis of drug resistance. Mol Syst Biol *5*, 310.

Chen, R., Mias, G., Li-Pook-Than, J., Jiang, L., Lam, H., Chen, R., Miriami, E., Karczewski, K., Hariharan, M., Dewey, F.*, et al.* (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell *148*, 1293-1307.

Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K.*, et al.* (2008). Variations in DNA elucidate molecular networks that cause disease. Nature *452*, 429-435.

Cheung, H., Cowley, G., Weir, B., Boehm, J., Rusin, S., Scott, J., East, A., Ali, L., Lizotte, P., Wong, T.*, et al.* (2011). Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proceedings of the National Academy of Sciences of the United States of America *108*, 12372-12377.

Cheung, V., and Spielman, R. (2002). The genetics of variation in gene expression. Nature genetics *32 Suppl*, 522-525.

Chi, D., Merchant, R., Rand, R., Conrad, A., Garrison, D., Turner, R., Morton, D., and Hoon, D. (1997). Molecular detection of tumor-associated antigens shared by human cutaneous melanomas and gliomas. The American journal of pathology *150*, 2143-2152.

Chi, P., Allis, C., and Wang, G. (2010). Covalent histone modifications--miswritten, misinterpreted and mis-erased in human cancers. Nature reviews Cancer *10*, 457-469.

Clark, A., Dos Santos, W., McCready, J., Chen, M., Van Meter, T., Ware, J., Wolber, S., Fillmore, H., and Broaddus, W. (2007). Wilms tumor 1 expression in malignant gliomas and correlation of +KTS isoforms with p53 status. Journal of neurosurgery *107*, 586-592.

Clark, A., Ware, J., Chen, M., Graf, M., Van Meter, T., Dos Santos, W., Fillmore, H., and Broaddus, W. (2010). Effect of WT1 gene silencing on the tumorigenicity of human glioblastoma multiforme cells. Journal of neurosurgery *112*, 18-25.

Corcoran, R., Ebi, H., Turke, A., Coffee, E., Nishino, M., Cogdill, A., Brown, R., Della Pelle, P., Dias-Santagata, D., Hung, K.*, et al.* (2012). EGFR-mediated re-activation of MAPK signaling contributes to insensitivity of BRAF mutant colorectal cancers to RAF inhibition with vemurafenib. Cancer discovery *2*, 227-235.

Cover, T.M., and Thomas, J.A. (2012). Elements of information theory (Wiley-interscience).

Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I., and Kong, A. (1999). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. Nat Genet *21*, 213-215.

Deangelo, D., Spencer, A., Bhalla, K., Prince, H., Fischer, T., Kindler, T., Giles, F., Scott, J., Parker, K., Liu, A.*, et al.* (2013). Phase Ia/II, two-arm, open-label, dose-escalation study of oral panobinostat administered via two dosing schedules in patients with advanced hematologic malignancies. Leukemia.

Deutschbauer, A.M., and Davis, R.W. (2005). Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat Genet *37*, 1333-1340.

Dhillon, P., Foster, D., and Ungar, L. (2011). Minimum description length penalization for group and multi-task sparse learning. The Journal of Machine Learning ….

Diabetes Genetics Initiative of Broad Institute of, H., Mit, L.U., Novartis Institutes of BioMedical, R., Saxena, R., Voight, B., Lyssenko, V., Burtt, N., de Bakker, P., Chen, H., Roix, J.*, et al.* (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science (New York, NY) *316*, 1331-1336.

Dietel, M., Jöhrens, K., Laffert, M., Hummel, M., Bläker, H., Müller, B., Lehmann, A., Denkert, C., Heppner, F., Koch, A.*, et al.* (2013). Predictive molecular pathology and its role in targeted cancer therapy: a review focussing on clinical relevance. Cancer gene therapy.

Dimas, A., Deutsch, S., Stranger, B., Montgomery, S., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M.*, et al.* (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. Science (New York, NY) *325*, 1246-1250.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics *7*, 1-26.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. Annals of Statistics *32*, 407-451.

Ehtesham, M., Min, E., Issar, N., Kasl, R., Khan, I., and Thompson, R. (2013). The role of the CXCR4 cell surface chemokine receptor in glioma biology. Journal of neuro-oncology.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S.*, et al.* (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423-428.

Engelman, J., Zejnullahu, K., Mitsudomi, T., Song, Y., Hyland, C., Park, J., Lindeman, N., Gale, C.-M., Zhao, X., Christensen, J.*, et al.* (2007). MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science (New York, NY) *316*, 1039-1043.

Evans, G.B., Furneaux, R.H., Gainsford, G.J., and Murphy, M.P. (2000). The synthesis and antibacterial activity of totarol derivatives. Part 3: Modification of ring-B. Bioorganic & medicinal chemistry *8*, 1663-1675.

Fei, P., Bernhard, E., and El-Deiry, W. (2002). Tissue-specific induction of p53 targets in vivo. Cancer research *62*, 7316-7327.

Feng, Z., Hu, W., de Stanchina, E., Teresky, A., Jin, S., Lowe, S., and Levine, A. (2007). The regulation of AMPK beta1, TSC2, and PTEN expression by p53: stress, cell and tissue specificity, and the role of these gene products in modulating the IGF-1-AKT-mTOR pathways. Cancer research *67*, 3043-3053.

Filipits, M., Rudas, M., Jakesz, R., Dubsky, P., Fitzal, F., Singer, C., Dietze, O., Greil, R., Jelen, A., Sevelda, P.*, et al.* (2011). A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res *17*, 6012-6020.

Fisher, S.A., Tremelling, M., Anderson, C.A., Gwilliam, R., Bumpstead, S., Prescott, N.J., Nimmo, E.R., Massey, D., Berzuini, C., Johnson, C.*, et al.* (2008). Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. Nature genetics *40*, 710-712.

Flanigan, S., Pitts, T., Eckhardt, S., Tentler, J., Tan, A., Thorburn, A., and Leong, S. (2010). The insulin-like growth factor I receptor/insulin receptor tyrosine kinase inhibitor PQIP exhibits enhanced antitumor effects in combination with chemotherapy against colorectal cancer models. Clin Cancer Res *16*, 5436-5446.

Frankild, S., and Jensen, L.J. DISEASES: Disease-gene associations mined from literature. (http://diseases.jensenlab.org/).

Fransen, K., Visschedijk, M., van Sommeren, S., Fu, J., Franke, L., Festen, E., Stokkers, P., van Bodegraven, A., Crusius, J., Hommes, D.*, et al.* (2010). Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. Human molecular genetics *19*, 3482-3488.

Frazer, K.A., Murray, S.S., Schork, N.J., and Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. Nat Rev Genet *10*, 241-251.

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

Fu, J., Wolfs, M., Deelen, P., Westra, H.-J., Fehrmann, R., Te Meerman, G., Buurman, W., Rensen, S., Groen, H., Weersma, R.*, et al.* (2012). Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS genetics *8*.

Gabriel, S.B., Salomon, R., Pelet, A., Angrist, M., Amiel, J., Fornage, M., Attie-Bitach, T., Olson, J.M., Hofstra, R., Buys, C.*, et al.* (2002). Segregation at three loci explains familial and population risk in Hirschsprung disease. Nat Genet *31*, 89-93.

Garcia-Rodriguez, L.J., Gay, A.C., and Pon, L.A. (2007). Puf3p, a Pumilio family RNA binding protein, localizes to mitochondria and regulates mitochondrial biogenesis and motility in budding yeast. The Journal of cell biology *176*, 197-207.

Garnett, M., Edelman, E., Heidorn, S., Greenman, C., Dastur, A., Lau, K., Greninger, P., Thompson, I., Luo, X., Soares, J.*, et al.* (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570-575.

Ge, D., Fellay, J., Thompson, A., Simon, J., Shianna, K., Urban, T., Heinzen, E., Qiu, P., Bertelsen, A., Muir, A.*, et al.* (2009). Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature *461*, 399-401.

Gerke, J., Lorenz, K., Ramnarine, S., and Cohen, B. (2010). Gene-environment interactions at nucleotide resolution. PLoS Genet *6*.

Giots, F., Donaton, M.C., and Thevelein, J.M. (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. Molecular microbiology *47*, 1163-1181.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A.*, et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531-537.

Gorre, M., Ellwood-Yen, K., Chiosis, G., Rosen, N., and Sawyers, C. (2002). BCR-ABL point mutants isolated from patients with imatinib mesylate-resistant chronic myeloid leukemia remain sensitive to inhibitors of the BCR-ABL chaperone heat shock protein 90. Blood *100*, 3041-3044.

Goto, Y., Sekine, I., Tanioka, M., Shibata, T., Tanai, C., Asahina, H., Nokihara, H., Yamamoto, N., Kunitoh, H., Ohe, Y.*, et al.* (2012). Figitumumab combined with carboplatin and paclitaxel in treatment-naïve Japanese patients with advanced non-small cell lung cancer. Investigational new drugs *30*, 1548-1556.

Gualberto, A., Hixon, M., Karp, D., Li, D., Green, S., Dolled-Filhart, M., Paz-Ares, L., Novello, S., Blakely, J., Langer, C.*, et al.* (2011). Pre-treatment levels of circulating free IGF-1 identify NSCLC patients who derive clinical benefit from figitumumab. British journal of cancer *104*, 68-74.

Guenther, M., Lawton, L., Rozovskaia, T., Frampton, G., Levine, S., Volkert, T., Croce, C., Nakamura, T., Canaani, E., and Young, R. (2008). Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. Genes & development *22*, 3403-3408.

Hamilton, A., Elrick, L., Myssina, S., Copland, M., Jørgensen, H., Melo, J., and Holyoake, T. (2006). BCR-ABL activity and its response to drugs can be determined in CD34+ CML stem cells by CrkL phosphorylation status using flow cytometry. Leukemia *20*, 1035-1039.

Harashima, T., Anderson, S., Yates, J., and Heitman, J. (2006). The kelch proteins Gpb1 and Gpb2 inhibit Ras activity via association with the yeast RasGAP neurofibromin homologs Ira1 and Ira2. Molecular cell *22*, 819-830.

Hartkamp, J., Carpenter, B., and Roberts, S. (2010). The Wilms' tumor suppressor protein WT1 is processed by the serine protease HtrA2/Omi. Molecular cell *37*, 159-171.

Hastie, T., Tibshirani, R., and Friedman, J.H. (2001). The Elements of Statistical Learning : Data Mining, Inference, and Prediction (Springer Series in Statistics).

Hasumi, K., Shinohara, C., Naganuma, S., and 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.

Hawkins, L., Jayanthan, A., and Narendran, A. (2005). Effects of 17-allylamino-17-demethoxygeldanamycin (17-AAG) on pediatric acute lymphoblastic leukemia (ALL) with respect to Bcr-Abl status and imatinib mesylate sensitivity. Pediatric research *57*, 430-437.

Heinemann, V., Douillard, J., Ducreux, M., and Peeters, M. (2013). Targeted therapy in metastatic colorectal cancer - An example of personalised medicine in action. Cancer treatment reviews.

Heiser, L., Sadanandam, A., Kuo, W.-L., Benz, S., Goldstein, T., Ng, S., Gibb, W., Wang, N., Ziyad, S., Tong, F.*, et al.* (2012). Subtype and pathway specific responses to anticancer compounds in breast cancer. Proceedings of the National Academy of Sciences of the United States of America *109*, 2724-2729.

Hillenmeyer, M.E., Fung, E., Wildenhain, J., Pierce, S.E., Hoon, S., Lee, W., Proctor, M., St Onge, R.P., Tyers, M., Koller, D.*, et al.* (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science *320*, 362-365.

Huang, R., Duan, S., Bleibel, W., Kistner, E., Zhang, W., Clark, T., Chen, T., Schweitzer, A., Blume, J., Cox, N.*, et al.* (2007). A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proceedings of the National Academy of Sciences of the United States of America *104*, 9758-9763.

Hubbell, E., Liu, W.M., and Mei, R. (2002). Robust estimators for expression analysis. Bioinformatics *18*, 1585-1592.

Jelier, R., Semple, J., Garcia-Verdugo, R., and Lehner, B. (2011). Predicting phenotypic variation in yeast from individual genome sequences. Nature genetics *43*, 1270-1274.

Jiang, Y., Vasconcelles, M.J., Wretzel, S., Light, A., Gilooly, L., McDaid, K., Oh, C.S., Martin, C.E., and Goldberg, M.A. (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, C.S., Chellappa, R., and Martin, C.E. (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., and Kumashiro, H. (1992). Effect of haloperidol on cyclic AMP and inositol trisphosphate in rat striatum in vivo. Prostaglandins, leukotrienes, and essential fatty acids *46*, 53-57.

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

Karapetis, C., Khambata-Ford, S., Jonker, D., O'Callaghan, C., Tu, D., Tebbutt, N., Simes, R., Chalchal, H., Shapiro, J., Robitaille, S., *et al.* (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med *359*, 1757-1765.

Khoury, C.M., Yang, Z., Li, X.Y., Vignali, M., Fields, S., and Greenwood, M.T. (2008). A TSC22-like motif defines a novel antiapoptotic protein family. FEMS Yeast Res *8*, 540-563.

Klinger, S., Poussin, C., Debril, M.-B., Dolci, W., Halban, P., and Thorens, B. (2008). Increasing GLP-1-induced beta-cell proliferation by silencing the negative regulators of signaling cAMP response element modulator-alpha and DUSP14. Diabetes *57*, 584-593.

Kopetz, S., Desai, J., Chan, E., Hecht, J., O'Dwyer, P., Lee, R., Nolop, K., and Saltz, L. (2010). PLX4032 in metastatic colorectal cancer patients with mutant BRAF tumors. J Clin Oncol *28*, 3534.

Krivtsov, A., Feng, Z., Lemieux, M., Faber, J., Vempati, S., Sinha, A., Xia, X., Jesneck, J., Bracken, A., Silverman, L.*, et al.* (2008). H3K79 methylation profiles define murine and human MLL-AF4 leukemias. Cancer cell *14*, 355-368.

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

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N.*, et al.* (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science *313*, 1929-1935.

Lee, C.S., Park, S.Y., Ko, H.H., Song, J.H., Shin, Y.K., and Han, E.S. (2005). Inhibition of MPP+-induced mitochondrial damage and cell death by trifluoperazine and W-7 in PC12 cells. Neurochem Int *46*, 169-178.

Lee, I.H., Kim, H.Y., Kim, M., Hahn, J.S., and Paik, S.R. (2008a). Dequalinium-induced cell death of yeast expressing alpha-synuclein-GFP fusion protein. Neurochem Res *33*, 1393-1400.

Lee, S.-I., Pe'er, D., Dudley, A., Church, G., and Koller, D. (2006). Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proceedings of the National Academy of Sciences of the United States of America *103*, 14062-14067.

Lee, S.-Y., and McLeod, H. (2011). Pharmacogenetic tests in cancer chemotherapy: what physicians should know for clinical application. The Journal of pathology *223*, 15-27.

Lee, S.H., van der Werf, J.H., Hayes, B.J., Goddard, M.E., and Visscher, P.M. (2008b). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet *4*, e1000231.

Lee, S.I., Dudley, A.M., Drubin, D., Silver, P.A., Krogan, N.J., Pe'er, D., and Koller, D. (2009). Learning a prior on regulatory potential from eQTL data. PLoS Genet *5*, e1000358.

Lemoine, M., Derenzini, E., Buglio, D., Medeiros, L., Davis, R., Zhang, J., Ji, Y., and Younes, A. (2012). The pan-deacetylase inhibitor panobinostat induces cell death and synergizes with everolimus in Hodgkin lymphoma cell lines. Blood *119*, 4017-4025.

Lemoine, M., and Younes, A. (2010). Histone deacetylase inhibitors in the treatment of lymphoma. Discovery medicine *10*, 462-470.

Li, W., and Melton, D. (2012). Cisplatin regulates the MAPK kinase pathway to induce increased expression of DNA repair gene ERCC1 and increase melanoma chemoresistance. Oncogene *31*, 2412-2422.

Lipson, D., Capelletti, M., Yelensky, R., Otto, G., Parker, A., Jarosz, M., Curran, J., Balasubramanian, S., Bloom, T., Brennan, K.*, et al.* (2012). Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. Nature medicine *18*, 382-384.

Litvin, O., Causton, H.C., Chen, B.J., and Pe'er, D. (2009). Modularity and interactions in the genetics of gene expression. Proc Natl Acad Sci U S A *106*, 6441-6446.

Liu, G., Yuan, X., Zeng, Z., Tunici, P., Ng, H., Abdulkadir, I., Lu, L., Irvin, D., Black, K., and Yu, J. (2006). Analysis of gene expression and chemoresistance of CD133+ cancer stem cells in glioblastoma. Molecular cancer *5*, 67.

Luo, B., Cheung, H., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J., Beroukhim, R., Weir, B.*, et al.* (2008). Highly parallel identification of essential genes in cancer cells. Proceedings of the National Academy of Sciences of the United States of America *105*, 20380-20385.

Mabuchi, S., Ohmichi, M., Nishio, Y., Hayasaka, T., Kimura, A., Ohta, T., Kawagoe, J., Takahashi, K., Yada-Hashimoto, N., Seino-Noda, H.*, et al.* (2004). Inhibition of inhibitor of nuclear factor-kappaB phosphorylation increases the efficacy of paclitaxel in in vitro and in vivo ovarian cancer models. Clin Cancer Res *10*, 7645-7654.

Maller, J., George, S., Purcell, S., Fagerness, J., Altshuler, D., Daly, M.J., and Seddon, J.M. (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.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A.*, et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753.

Martin, C.E., Oh, C.S., and Jiang, Y. (2007). Regulation of long chain unsaturated fatty acid synthesis in yeast. Biochimica et biophysica acta *1771*, 271-285.

Martin, M., Hayward, R., Viros, A., and Marais, R. (2012). Metformin accelerates the growth of BRAF V600E-driven melanoma by upregulating VEGF-A. Cancer discovery *2*, 344-355.

Mathew, C. (2008). New links to the pathogenesis of Crohn disease provided by genome-wide association scans. Nature reviews Genetics *9*, 9-14.

MathWorks, T. MATLAB.

McLeod, H. (2013). Cancer pharmacogenomics: early promise, but concerted effort needed. Science (New York, NY) *339*, 1563-1566.

Mehrabian, M., Allayee, H., Stockton, J., Lum, P., Drake, T., Castellani, L., Suh, M., Armour, C., Edwards, S., Lamb, J.*, et al.* (2005). Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. Nature genetics *37*, 1224-1233.

Metzker, M. (2010). Sequencing technologies - the next generation. Nature reviews Genetics *11*, 31-46.

Monteiro, A. (2003). BRCA1: the enigma of tissue-specific tumor development. Trends in genetics : TIG *19*, 312-315.

Mouillon, J.M., and Persson, B.L. (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.

Natarajan, B.K. (1995). Sparse approximate solutions to linear systems. SIAM journal on computing *24*, 227-234.

Nguyen, A., Taranova, O., He, J., and Zhang, Y. (2011). DOT1L, the H3K79 methyltransferase, is required for MLL-AF9-mediated leukemogenesis. Blood *117*, 6912-6922.

Nguyen, D., Chen, G., Reddy, R., Tsai, W., Schrump, W., Cole, G., and Schrump, D. (2004). Potentiation of paclitaxel cytotoxicity in lung and esophageal cancer cells by pharmacologic inhibition of the phosphoinositide 3-kinase/protein kinase B (Akt)-mediated signaling pathway. The Journal of thoracic and cardiovascular surgery *127*, 365-375.

Nica, A., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K.*, et al.* (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS genetics *7*.

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

Nulton-Persson, A.C., and Szweda, L.I. (2001). Modulation of mitochondrial function by hydrogen peroxide. J Biol Chem *276*, 23357-23361.

Patel, N., Nozaki, S., Shortle, N., Bhat-Nakshatri, P., Newton, T., Rice, S., Gelfanov, V., Boswell, S., Goulet, R., Sledge, G.*, et al.* (2000). Paclitaxel sensitivity of breast cancer cells with constitutively active NF-kappaB is enhanced by IkappaBalpha super-repressor and parthenolide. Oncogene *19*, 4159-4169.

Patterson, K., Brummer, T., O'Brien, P., and Daly, R. (2009). Dual-specificity phosphatases: critical regulators with diverse cellular targets. The Biochemical journal *418*, 475-489.

Pearl, J. (2000). Causality : models, reasoning, and inference (Cambridge: Cambridge University Press).

Peeters, T., Louwet, W., Gelade, R., Nauwelaers, D., Thevelein, J.M., and Versele, M. (2006). Kelch-repeat proteins interacting with the G(alpha) protein Gpa2 bypass adenylate cyclase for direct regulation of protein kinase A in yeast. Proceedings of the National Academy of Sciences of the United States of America *103*, 13034-13039.

Perlstein, E.O., Ruderfer, D.M., Ramachandran, G., Haggarty, S.J., Kruglyak, L., and Schreiber, S.L. (2006). Revealing complex traits with small molecules and naturally recombinant yeast strains. Chemistry & biology *13*, 319-327.

Perlstein, E.O., Ruderfer, D.M., Roberts, D.C., Schreiber, S.L., and Kruglyak, L. (2007). Genetic basis of individual differences in the response of small-molecule drugs in yeast. Nature Genetics *39*, 496-502.

Pfaffl, M. (2001). A new mathematical model for relative quantification in real-time RT-PCR. Nucleic acids research *29*.

Pollak, M. (2008). Insulin and insulin-like growth factor signalling in neoplasia. Nature reviews Cancer *8*, 915-928.

Prahallad, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., Beijersbergen, R., Bardelli, A., and Bernards, R. (2012). Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature *483*, 100-103.

Price, A., Helgason, A., Thorleifsson, G., McCarroll, S., Kong, A., and Stefansson, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS genetics *7*.

Prickett, T., and Samuels, Y. (2012). Molecular pathways: dysregulated glutamatergic signaling pathways in cancer. Clin Cancer Res *18*, 4240-4246.

Prince, H., Bishton, M., and Harrison, S. (2009). Clinical studies of histone deacetylase inhibitors. Clin Cancer Res *15*, 3958-3969.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748-752.

Radujkovic, A., Schad, M., Topaly, J., Veldwijk, M., Laufs, S., Schultheis, B., Jauch, A., Melo, J., Fruehauf, S., and Zeller, W. (2005). Synergistic activity of imatinib and 17-AAG in imatinib-resistant CML cells overexpressing BCR-ABL--Inhibition of P-glycoprotein function by 17-AAG. Leukemia *19*, 1198-1206.

Raina, R., Ng, A.Y., and Koller, D. (2006). Constructing informative priors using transfer learning. Paper presented at: Proceedings of the 23rd international conference on Machine learning (ACM).

Rissanen, J. (1978). Modeling by shortest data description. Automatica *14*, 465-471.

Rissanen, J. (1999). Hypothesis selection and testing by the MDL principle. The Computer Journal *42*, 260-269.

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

Rosenzweig, S. (2012). Acquired resistance to drugs targeting receptor tyrosine kinases. Biochemical pharmacology *83*, 1041-1048.

Safiulina, D., Veksler, V., Zharkovsky, A., and Kaasik, A. (2006). Loss of mitochondrial membrane potential is associated with increase in mitochondrial volume: physiological role in neurones. Journal of cellular physiology *206*, 347-353.

Saint-Georges, Y., Garcia, M., Delaveau, T., Jourdren, L., Le Crom, S., Lemoine, S., Tanty, V., Devaux, F., and 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, M.D., and Garcia-Perez, A.I. (2007). Dequalinium induces cell death in human leukemia cells by early mitochondrial alterations which enhance ROS production. Leuk Res *31*, 969-978.

Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., Kawaguchi, T., Tsunoda, T., Watanabe, M., Takeda, A.*, et al.* (2009). Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nat Genet *41*, 1303-1307.

Sattler, M., and Salgia, R. (1998). Role of the adapter protein CRKL in signal transduction of normal hematopoietic and BCR/ABL-transformed cells. Leukemia *12*, 637-644.

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C.*, et al.* (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet *37*, 710-717.

Scott, L., Mohlke, K., Bonnycastle, L., Willer, C., Li, Y., Duren, W., Erdos, M., Stringham, H., Chines, P., Jackson, A.*, et al.* (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science (New York, NY) *316*, 1341-1345.

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. Nature genetics *36*, 1090-1098.

Simón-Sánchez, J., Schulte, C., Bras, J., Sharma, M., Gibbs, J., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S., Hernandez, D.*, et al.* (2009). Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nature genetics *41*, 1308-1312.

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

Slany, R. (2009). The molecular biology of mixed lineage leukemia. Haematologica *94*, 984-993.

Smith, E.N., and Kruglyak, L. (2008). Gene-environment interaction in yeast gene expression. PLoS Biol *6*, e83.

Spiliotaki, M., Markomanolaki, H., Mela, M., Mavroudis, D., Georgoulias, V., and Agelaki, S. (2011). Targeting the insulin-like growth factor I receptor inhibits proliferation and VEGF production of non-small cell lung cancer cells and enhances paclitaxel-mediated anti-tumor effect. Lung cancer (Amsterdam, Netherlands) *73*, 158-165.

Storici, F., Durham, C.L., Gordenin, D.A., and Resnick, M.A. (2003). Chromosomal site-specific double-strand breaks are efficiently targeted for repair by oligonucleotides in yeast. Proc Natl Acad Sci U S A *100*, 14994-14999.

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

Stumpel, D., Schneider, P., Seslija, L., Osaki, H., Williams, O., Pieters, R., and Stam, R. (2012). Connectivity mapping identifies HDAC inhibitors for the treatment of t(4;11)-positive infant acute lymphoblastic leukemia. Leukemia *26*, 682-692.

Suppiah, V., Moldovan, M., Ahlenstiel, G., Berg, T., Weltman, M., Abate, M., Bassendine, M., Spengler, U., Dore, G., Powell, E.*, et al.* (2009). IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nature genetics *41*, 1100-1104.

Tagkopoulos, I., Liu, Y.C., and Tavazoie, S. (2008). Predictive behavior within microbial genetic networks. Science *320*, 1313-1317.

Takeuchi, F., McGinnis, R., Bourgeois, S., Barnes, C., Eriksson, N., Soranzo, N., Whittaker, P., Ranganath, V., Kumanduri, V., McLaren, W*., et al.* (2009). A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS genetics *5*.

Tanaka, Y., Nishida, N., Sugiyama, M., Kurosaki, M., Matsuura, K., Sakamoto, N., Nakagawa, M., Korenaga, M., Hino, K., Hige, S*., et al.* (2009). Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. Nature genetics *41*, 1105-1109.

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

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. J R Statist Soc B *58*, 267-288.

Tuquet, C., Dupont, J., Mesneau, A., and Roussaux, J. (2000). Effects of tamoxifen on the electron transport chain of isolated rat liver mitochondria. Cell biology and toxicology *16*, 207-219.

Turalba, A., Leite-Morris, K., and Kaplan, G. (2004). Antipsychotics regulate cyclic AMP-dependent protein kinase and phosphorylated cyclic AMP response element-binding protein in striatal and cortical brain regions in mice. Neuroscience letters *357*, 53-57.

van de Vijver, M., He, Y., van't Veer, L., Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M*., et al.* (2002). A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med *347*, 1999-2009.

van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T*., et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530-536.

Vance, D., Goldberg, I., Mitsuhashi, O., and Bloch, K. (1972). Inhibition of fatty acid synthetases by the antibiotic cerulenin. Biochemical and biophysical research communications *48*, 649-656.

Wagner, K.-D., Wagner, N., Wellmann, S., Schley, G., Bondke, A., Theres, H., and Scholz, H. (2003). Oxygen-regulated expression of the Wilms' tumor suppressor Wt1 involves hypoxia-inducible factor-1 (HIF-1). FASEB journal : official publication of the Federation of American Societies for Experimental Biology *17*, 1364-1366.

Wagner, N., Panelos, J., Massi, D., and Wagner, K.-D. (2008). The Wilms' tumor suppressor WT1 is associated with melanoma proliferation. Pflügers Archiv : European journal of physiology *455*, 839-847.

Wang, K., Zhang, H., Kugathasan, S., Annese, V., Bradfield, J., Russell, R., Sleiman, P., Imielinski, M., Glessner, J., Hou, C.*, et al.* (2009). Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. American journal of human genetics *84*, 399-405.

Wang, L., McLeod, H., and Weinshilboum, R. (2011). Genomics and drug response. N Engl J Med *364*, 1144-1153.

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

Waring, P., Khan, T., and Sjaarda, A. (1997). Apoptosis induced by gliotoxin is preceded by phosphorylation of histone H3 and enhanced sensitivity of chromatin to nuclease digestion. The Journal of biological chemistry *272*, 17929-17936.

Weedon, M.N., and Frayling, T.M. (2008). Reaching new heights: insights into the genetics of human stature. Trends Genet *24*, 595-603.

Weinstein, J. (2006). Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines. Molecular cancer therapeutics *5*, 2601-2605.

Weisberg, E., Catley, L., Wright, R., Moreno, D., Banerji, L., Ray, A., Manley, P., Mestan, J., Fabbro, D., Jiang, J.*, et al.* (2007). Beneficial effects of combining nilotinib and imatinib in preclinical models of BCR-ABL+ leukemias. Blood *109*, 2112-2120.

Wellcome Trust Case Control, C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661-678.

Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M.*, et al.* (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nature genetics *40*, 161-169.

Wolff, E., Chihara, Y., Pan, F., Weisenberger, D., Siegmund, K., Sugano, K., Kawashima, K., Laird, P., Jones, P., and Liang, G. (2010). Unique DNA methylation patterns distinguish noninvasive and invasive urothelial cancers and establish an epigenetic field defect in premalignant tissue. Cancer research *70*, 8169-8178.

Wykoff, D.D., Rizvi, A.H., Raser, J.M., Margolin, B., and O'Shea, E.K. (2007). Positive feedback regulates switching of phosphate transporters in S. cerevisiae. Mol Cell *27*, 1005-1013.

Xie, M.W., Jin, F., Hwang, H., Hwang, S., Anand, V., Duncan, M.C., and 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.

Yin, D., Tamaki, N., Parent, A., and Zhang, J. (2005). Insulin-like growth factor-I decreased etoposide-induced apoptosis in glioma cells by increasing bcl-2 expression and decreasing CPP32 activity. Neurological research *27*, 27-35.

Yip, K.W., Mao, X., Au, P.Y., Hedley, D.W., Chow, S., Dalili, S., Mocanu, J.D., Bastianutto, C., Schimmer, A., and Liu, F.F. (2006). Benzethonium chloride: a novel anticancer agent identified by using a cell-based small-molecule screen. Clin Cancer Res *12*, 5557-5569.

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

Zamora-Avila, D., Franco-Molina, M., Trejo-Avila, L., Rodríguez-Padilla, C., Reséndez-Pérez, D., and Zapata-Benavides, P. (2007). RNAi silencing of the WT1 gene inhibits cell proliferation and induces apoptosis in the B16F10 murine melanoma cell line. Melanoma research *17*, 341-348.

Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R., Rayner, N.W., Freathy, R.M.*, et al.* (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science *316*, 1336-1341.

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

Zuk, O., Hechter, E., Sunyaev, S., and Lander, E. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences of the United States of America *109*, 1193-1198.