The Dissertation Committee for Jonathan Hubert Young
certifies that this is the approved version of the following dissertation:

Computational Discovery of Genetic Targets and Interactions: Applications to Lung Cancer

Committee:

Edward M. Marcotte, Supervisor

Oscar Gonzalez

Inderjit Dhillon

Ron Elber

Claus Wilke

Computational Discovery of Genetic Targets and Interactions: Applications to Lung Cancer

by

Jonathan Hubert Young, B.S.; M.S.C.A.M.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2016

Dedicated to my parents.

Acknowledgments

Through the years, I have met a number of individuals from whom I have learned much and received invaluable support. First, I would like to acknowledge my committee: Professors Oscar Gonzalez, Inderjit Dhillon, Ron Elber and Claus Wilke. They have provided me with constructive input regarding my research projects, and their comments proved to be especially helpful when I was in the process of publishing my first paper. The financial support provided by the Computational and Applied Mathematics fellowship here at The University of Texas at Austin, and from the Cancer Prevention and Research Institute of Texas (CPRIT) is greatly appreciated. It has filled me with great pride to be a part of the amazing university here.

To the members of the Marcotte lab, thank you for contributing to a welcoming and encouraging environment. It has been reassuring to know that I could turn to you for new ideas or direction throughout my time as a graduate student. While there are too many to name here, I would particularly like to thank Kevin Drew for being generous with his time whenever I needed help, and Jon Laurent and Dan Boutz for their friendship and research advice.

I have been fortunate to have as my PhD advisor Edward Marcotte, who has influenced me beyond my development as a computational scientist, from the way I think about and approach problems in and outside of science.

In terms of scientific interests, he introduced me to 3-D printing and has been most instrumental in developing my appreciation and passion for computer programming and machine learning. The lessons he has taught me in forming and executing a research project will serve me well in areas outside of biology. My time in his lab reminds of a quote from Richard Feynman: "I was born not knowing and have had only a little time to change that here and there." I thank Edward for his patience, guidance and support.

Finally, there are those whose impact on me beyond graduate school has been immeasurable. I am indebted to them and will forever be grateful. To Corey Bryant, Ben Gilman and Gabriel Wu, I would not have made it without you. To John Uecker, M.D. and Reginald Baptiste, M.D., thank you so much for taking me under your wing and providing me with encouragement, inspiration, and mentorship. And to my parents, your unwavering and unconditional support has meant the world to me. Thank you for everything.

Computational Discovery of Genetic Targets and Interactions: Applications to Lung Cancer

Publication No. __________________

Jonathan Hubert Young, Ph.D.
The University of Texas at Austin, 2016

Supervisor: Edward M. Marcotte

We present new modes of computational drug discovery in each of the three key themes of target identification, mechanism, and therapy regimen design. In identifying candidate targets for therapeutic intervention, we develop novel applications of unsupervised clustering of whole genome RNAi screening in prioritizing biological systems whose inhibition differentially sensitizes diseased cells apart from a normal population. When applied to lung cancer, our approach identified protein complexes for which various tumor subtypes are especially dependent. Consequently, each complex represents a candidate drug target specifically intended for a particular patient sub-population. The cellular functions impacted by the protein complexes include splicing, translation, and protein folding. We obtained experimental validation for the predicted sensitivity of a lung adenocarcinoma cell line to Wnt inhibition.

For our second theme, we focus on genetic interactions as a mechanism underlying sensitivity to target inhibition. Experimental characterization of

such interactions has relied on brute-force assessment of gene pairs. To alleviate the experimental burden, our hypothesis is that functionally related genes tend to share common genetic interaction partners. We thereby examine a method that recognizes functional network clusters to generate high-confidence predictions of different types of genetic interactions across yeast, fly and human. Our predictions are leave-one-out cross-validated on known interactions. Moreover, by using yeast as a model, we simulatr the degree to which further human genetic interactions need to be screened in order to understand their distribution in biological systems.

Finally, we employ yeast as a model organism to assess the feasibility of designing synergistic or antagonistic drug pairs based on genetic interactions between their targets. The hypothesis is that if the target genes of one chemical compound are close to those of a second compound in a genetic interaction network, then synergistic or antagonistic growth effects will occur. Proximity between sets in a gene network are quantified through graph metrics, and predictions of synergy and antagonism are validated by literature-curated gold standards. Ultimately, integrating knowledge of druggable targets, how gene perturbations interact with the genetic background of an individual, and design of personalized therapeutic regimens will provide a general framework to treat a variety of diseases.

Table of Contents

Acknowledgments v

Abstract vii

Chapter 1. Introduction 1

Chapter 2. Computational Discovery of Pathway-Level Genetic Vulnerabilities in Non-Small-Cell Lung Cancer 4

2.1 Abstract 4

2.2 Introduction 5

2.3 Methods 8

2.3.1 Experimental datasets and procedures 8

2.3.2 Application of k-means clustering 10

2.3.3 Biclustering 13

2.3.4 Alternative methods for determination of gene set sensitivity 14

2.4 Results 14

2.4.1 Subgroup-specific NSCLC vulnerabilities are found among protein complexes 17

2.4.2 Biclustering finds genetic vulnerabilities without reliance on annotated gene sets 23

2.4.3 Small-molecule screen confirms predicted cell line sensitivity 28

2.4.4 Alternative approaches to measuring complex sensitivity do not prioritize bimodality 29

2.5 Discussion 32

2.6 Summary 36

2.7 Funding 36

2.8 Acknowledgments 37

Chapter 3. Predictability of Genetic Interactions from Functional Gene Modules 38

3.1 Abstract 38

3.2 Introduction 39

3.3 Materials and Methods 41

3.3.1 Statistical Analysis 42

3.3.2 Data Availability 42

3.4 Results 43

3.4.1 The human functional gene network is predictive for phenotype-based genetic interactions 45

3.4.2 Fly phenotypic enhancement and suppression interactions are predicted from functional net clusters 47

3.4.3 High-confidence predictability is found in human, fly and yeast 49

3.4.4 Trends of interaction enrichment within gene modules vary with interaction type 51

3.5 Discussion 54

3.6 Acknowledgments 56

Chapter 4. Development of Drug Combination Regimens: Yeast as a Model System 57

4.1 A History of Yeast in Drug Discovery 57

4.2 Predicting Synergistic and Antagonistic Drug Pairs in Yeast . 60

4.2.1 Abstract 60

4.2.2 Introduction 61

4.2.3 Methods 63

4.2.3.1 Negative and positive genetic interaction network 63

4.2.3.2 Chemical compound targets and gold standards for synergy and antagonism 63

4.2.3.3 Distances in networks 64

4.2.3.4 Software Availability 65

4.2.4 Results 65

4.2.4.1 Close proximity between drug target genes in the genetic interaction network does not strongly predict synergy or antagonism 65

4.2.4.2 Genetic interaction strength is moderately correlated with network distance 67

4.2.5 Discussion 68

Chapter 5. Conclusion **71**

Appendices **74**

Appendix A. NSCLC Genetic Vulnerabilities **75**

Appendix B. Genetic Interactions Predicted from YeastNet **88**

Bibliography **92**

Chapter 1

Introduction

From various perspectives, drug discovery and development is an intricate and arduous process. The total cost of a drug, which includes pre-clinical investigations through clinical trials, varies according to estimates, but an agreeable figure is on the order of hundreds of millions of U.S. dollars [1]. On average, over seven years are spent to bring a drug to market [40]. It is clear that advancements within the drug development pipeline to lessen the time and cost burden would alleviate the burden of disease on patients and healthcare systems.

On the other hand in recent years, a wealth of molecular and genetic data on organisms ranging from bacteria to multicellular eukaryotes including humans has been generated with the maturity of high-throughput experiments. With computing resources requiring relatively low barriers to entry in terms of cost and accessibility, it has become natural to employ computational techniques on readily available data to not only construct a clearer understanding of biology but also to generate applications, in particular to drug discovery and development. Traditionally, drug discovery involves construction or identification of a druggable target or lead biologic or chemical compound. Following this process is drug development, which involves turning an agent into an actual marketable drug through clinical trials and regulatory approval.

In this work, we address the development of computational techniques to address three key themes in the drug discovery process. Chapter 2 discusses the first theme of therapeutic target identification, in which we investigate how to choose genetic vulnerabilities that sensitize diseased cells but not normal healthy cell populations. An crucial goal to achieve here is specificity - to inhibit only those cells causing disease and distinguish non-responders versus responders. Responders are patients for whom an intended therapeutic intervention is successful, while non-responders are refractory to treatment. The particular application we address in this theme is lung cancer, in which we wish to computationally establish novel candidate drug targets specific to the various subtypes of the disease.

The second theme in Chapter 3 involves uncovering the mechanism underlying how targeting a gene or biological system brings about the desired therapeutic effect. From cancers to infectious diseases, a commonly desired outcome is to selectively inhibit tumors or infected cells while leaving normal cells unaffected. While such diseases can be extraordinarily diverse in their pathophysiology, characterization a generally applicable mechanism for selective inhibition would benefit any drug discovery effort. The particular mechanism addressed here is genetic interactions, which involves how the behavior of a gene depends on the action of other genes. Because experimental determination of genetic interactions is technically intensive and laborious, our goal will be establishing a computational prediction algorithm to either guide the experimental testing or reliably predict novel such interactions.

Our final theme presented in Chapter 4 involves general guiding principles for developing therapeutic regimens. Using yeast as a model organism, we attempt to design effective drug combinations by exploiting genetic interactions as therapeutic targets. Yeast serves as an ideal platform for proof-of-concept due to the multitude of chemical and genetic screening data available, owing to its experimental tractability relative to mammalian cells. Traditionally, there have been two main strategies in drug discovery. One such strategy is phenotypic screening, in which biologics or small-molecules of interest are selected based on generation of desired phenotypes in high-throughput assays. A second approach is target-based screening relies on the discovery of a druggable point of intervention for a disease; agents are then developed to specifically target the vulnerability. A recent survey of history of phenotypic and target-based screening concluded that the majority of FDA-approved drugs originated from the phenotypic approach [77]. Yet many effective therapies were discovered through target-based screening; a notable example is imatinib for chronic myelogenous leukemia [21]. In addressing the three-part theme of target, mechanism, and therapy, we believe that aspects of the target-based approach are amenable to computational advances.

Chapter 2

Computational Discovery of Pathway-Level Genetic Vulnerabilities in Non-Small-Cell Lung Cancer[1]

2.1 Abstract

Motivation: Novel approaches are needed for discovery of targeted therapies for non-small-cell lung cancer (NSCLC) that are specific to certain patients. Whole genome RNAi screening of lung cancer cell lines provides an ideal source for determining candidate drug targets.

Results: Unsupervised learning algorithms uncovered patterns of differential vulnerability across lung cancer cell lines to loss of functionally related genes. Such genetic vulnerabilities represent candidate targets for therapy and are found to be involved in splicing, translation and protein folding. In particular, many NSCLC cell lines were especially sensitive to the loss of components of the LSm2-8 protein complex or the CCT/TRiC chaperonin. Different vulnerabilities were also found for different cell line subgroups. Furthermore, the predicted vulnerability of a single adenocarcinoma cell line to loss of the Wnt

[1] Published as Young JH, Peyton M, Kim HS, McMillan E, Minna JD, White MA, Marcotte EM, Computational Discovery of Pathway-Level Genetic Vulnerabilities in Non-Small-Cell Lung Cancer. *Bioinformatics*, 32(9): 1373-1379 (2016). J.H.Y. designed the study, performed data analysis, and wrote the paper.

pathway was experimentally validated with screening of small-molecule Wnt inhibitors against an extensive cell line panel.

Availability and implementation: The clustering algorithm is implemented in Python and is freely available at `https://bitbucket.org/youngjh/nsclc_paper`.

Contact: marcotte@icmb.utexas.edu or jon.young@utexas.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

2.2 Introduction

Non-small-cell lung cancer (NSCLC) remains a significant healthcare burden despite recent progress in drug discovery and development. Recent FDA-approved targeted therapies are only intended for appropriate subpopulations of patients. The drug Xalkori (crizotinib) is highly effective, but only for ˜4% of lung cancer patients [69]. Similarly, Iressa (gefitinib) and other EGFR inhibitors target mutations found only in a portion of patients while the majority have the wild-type version [44]. Compared with cytotoxic chemotherapy, targeted therapy has the advantage of greater specificity. However, discovery and development of such agents requires the identification of druggable targets. Inhibitors of certain characteristic mutations, such as KRAS G12C and G12D, are still under extensive development for clinical use [17, 36]. The heterogeneity of NSCLC is another barrier confronting drug discovery. A number of different subtypes exist, and identifying the appropriate patient subpopulation for therapy is crucial. Therefore, the problem becomes one of identifying druggable targets in NSCLC that will guide discovery of small-molecule compounds or antibodies against these targets, and also to identify the patient subpopulations to which the targets apply. We attempt to tackle the former issue through computational analysis of a high-throughput whole genome RNA interference (RNAi) screen against a panel of NSCLC cell lines.

When identifying drug targets, one approach is to identify genes whose knockdown selectively leads to death of cancer cells but not matched normal cells. Such genes represent genetic vulnerabilities and potential drug targets. Several studies have been able to use whole genome RNAi to identify genetic vulnerabilities for cancer drug target discovery. A screen against all genes in two lung cancer cell lines identified proteasome members as candidate targets and discovered that small-molecule proteasome inhibition synergized with radiotherapy in a mouse xenograft model [18]. Another study applied a whole genome shRNA screen on lung cancer cell lines to discover genes that were part of the Wnt pathway whose knockout potentiates EGFR inhibition [11].

For drug discovery, it is desirable to investigate many cell lines. A major effort, termed Project Achilles, involved RNAi knockdown of more than 11000 human genes using shRNA libraries in over 100 cancer cell lines. Ovarian cancer cell lines were found to be especially dependent upon ˜50 genes [12]. A follow-up study sought to uncover essential cancer genes based on the hypothesis that certain genes that are not themselves oncogenes but show copy-number loss could be cancer vulnerabilities. A scoring scheme was developed to prioritize genes that were essential to cancer cell lines and also exhibited partial copy number loss [53]. Recently, independent of Project Achilles, extensive high-throughput chemical and genetic screens were employed to explore new avenues of treating NSCLC. The study found molecular signatures of FLIP and COPI addiction and indolotriazine sensitivity that indicate genetic vulnerabilities present in patient populations [41]. The genetic screens leading to these results included siRNA screening of a number of lung cancer cell lines. This screening was ultimately conducted on a whole genome scale, which motivated this study.

Here, we propose a novel computational approach to prioritize candidate drug targets for NSCLC by subdividing cell lines into different groups and identifying genetic vulnerabilities targeted to each group. In particular, we aim to attain a binary partitioning of cell lines into either sensitive or resistant to targeting of a particular genetic vulnerability. We are interested only in genetic vulnerabilities that sensitize a subgroup of cell lines rather than all cell lines because due to the genetic heterogeneity of lung cancer, an effective universal treatment for all NSCLC types is not thought to exist. Applications of unsupervised learning algorithms were developed that identify biological processes and protein complexes to which NSCLC cell lines are differentially sensitive upon siRNA knockdown. The top-scoring results represent lung cancer genetic vulnerabilities and candidate therapy targets.

2.3 Methods

2.3.1 Experimental datasets and procedures

Our study centers on a cell line panel consisting of 12 patient-derived NSCLC cell lines and one immortalized normal epithelial cell line (Table 2.1). Included among the cell lines are subtypes commonly observed in patients: adenocarcinoma, squamous-cell and large-cell carcinoma. As described previously [41], a whole genome knockdown screen with Ambion and Dharmacon siRNA libraries in the 96-well plate format was conducted against the cell line panel. For each gene, either three siRNAs (Ambion) or four siRNAs (Dharmacon) were pooled, and cell line viability was measured using the CellTiter-Glo (Promega) assay. Raw data were row and column median normalized. Using siMacro [72], a robust Z score was calculated from the screening data to reflect the viability of each cell line to knockdown of a single gene. A robust Z score is defined as

$$\frac{\text{Cell viability} - \text{median}}{\text{Median absolute deviation}}$$

and is less sensitive to outliers than a traditional Z score. Both the median and median absolute deviation were calculated over data grouped by experimental batch.

It was determined that robust Z scores less than -3.0 reflected non-viability. Scores were combined from both Ambion and Dharmacon libraries by taking the minimum of the scores. Thus, it was assumed that disagreement

H1155	Large cell neuroendocrine
HCC366	Adenosquamous
H1819	Adenocarcinoma
HCC44	Adenocarcinoma
HCC4017	Large cell carcinoma
H1993	Adenocarcinoma
H460	Large cell carcinoma
H2073	Adenocarcinoma
H2009	Adenocarcinoma
H2122	Adenocarcinoma
H1395	Adenocarcinoma
HCC95	Squamous cell carcinoma
HBEC30	Normal bronchiole epithelial

Table 2.1: Thirteen cell lines on which our whole genome RNAi screen and computational analyses were conducted.

between the results of the two libraries were more likely to be due to false-negatives. The siRNA screen Z scores were further simplified by binarizing as follows. All robust Z scores less than -3.0 were set equal to 1; otherwise the score was set equal to 0. In essence, a binarized score of 1 represents a hit or sensitivity of a cell line to the corresponding gene knockdown.

A larger pool of NSCLC cell lines encompassing the cell line panel described above was screened with the tankyrase inhibitors IWR-1-endo (Calbiochem) and XAV 939 (Tocris) in an 8-point 4-fold dilution series (top dose = 100 M) in 96-well plates. Cells were plated 24 h prior to the addition of drug, incubated for 4 days, and assayed using MTS (CellTiter 96 Aqueous One Solution Cell Proliferation Assay) according to the manufacturer's instructions (Promega). Cell number per well was determined empirically and ranged from 500 to 4000 per well, inversely proportional to doubling times (typically 2000/well). Dose response curves were generated and IC50s calculated using in-house software, DIVISA. All cells were grown in RPMI-1640 (Sigma) supplemented with 5% FBS and incubated at 37°C in a humidified atmosphere with 5% CO_2. Cell lines were authenticated using the Power-Plex 1.2 kit (Promega) and confirmed to match the DNA fingerprint library maintained by ATCC and the Minna/Gazdar laboratory and confirmed to be free of mycoplasma by e-Myco kit (Boca Scientific).[2]

RNAi screens of cancer cell lines from Project Achilles [16] were utilized as an external comparison dataset for our study. Results from shRNA knockdown of 5711 genes on 19 NSCLC cell lines were extracted from Project Achilles v2.4.3. NaN values were imputed by replacement with row medians. No thresholding of the data was carried out so viability was assessed on a continuous spectrum. We followed the Project Achilles convention of identifying lower gene knockdown values with greater essentiality and higher values with reduced essentiality. A number of genes were associated with multiple knockdown values for each cell line; these data were kept as is.

2.3.2 Application of *k*-means clustering

The gene sets examined for genetic vulnerabilities were protein complexes chosen from CORUM [64] and literature sources (`http://metazoa.med.utoronto.ca`) [31]. The full RNAi data were represented as a $m \times n$

[2] Paragraph contributed by Michael Peyton.

matrix M where m is the number of genes, n is the number of cell lines, and

$$M_{ij} = \begin{cases} 0, & \text{if cell line } j \text{ survives knockdown of gene } i \\ 1, & \text{otherwise} \end{cases}$$

Extracting the RNAi sensitivity profiles for genes in a protein complex yields a $r \times n$ submatrix P of M where r is the number of genes in the complex. Thus, every protein complex is represented as a matrix of ones and zeros.

For each protein complex, we measured the degree of bimodal response to gene knockdowns as follows. Denoting by P the matrix for a protein complex as above, we computed a vector v by calculating the column means of P: $v_j = (1/r)\sum_{i=1}^{r} P_{ij}$. Then k-means clustering with $k = 2$ was applied to v, and the difference between the resulting centroids was the score assigned to each protein complex. By calculating the column means, we normalized for the size of the complex. We considered two different implementations of the k-means clustering algorithm. First, we used the standard implementation found in the Python library scikit-learn, which runs the algorithm 10 times with different centroid seeds, choosing the result on the basis of the within-cluster sum-of-squares [59]. The technique from k-means++ was followed for centroid initialization [3]. Second, an alternative implementation was *Ckmeans.1d.dp*, which employs dynamic programming to guarantee optimal solutions for the one-dimensional case [81].

A permutation test to determine statistical significance was performed in the following manner. For each protein complex, the entire RNAi data for all genes were permuted, followed by repeating the k-means clustering on the same complex. The permutations were repeated 1000 times to generate a distribution of scores from the randomized data. Then the score from the actual complex was compared with the distribution to calculate a P-value. For every protein complex, this entire process of generating a distribution of scores from permuted data to compare against the complex's actual score to yield a P-value was repeated. Finally, multiple hypothesis correction at 10% FDR was carried out using the q-value statistical package [76]. One alternative to the permutation test in assessing statistical significance is Fisher's exact test. Specifically, for each protein complex, a contingency table was tabulated according to the number of viable and non-viable gene knockdowns, and which cell line cluster (according to the 2-means clustering) those knockdowns fell within. The Bonferroni correction at 10% FDR was applied to the P-values from Fisher's exact test.

A procedure to benchmark the performance of the 2-means clustering method was based on a leave-one-out strategy. For each protein complex, a single gene member was randomly withheld. The remaining gene members formed a training set, on which the 2-means clustering was calculated. The clustering resulted in assignments of cell lines to either a knockdown-sensitive or knockdown-resistant cluster. Using these assignments, we tested whether the average number of RNAi hits in the sensitive cell lines for the withheld gene was greater than that of the resistant lines. To assess for statistical significance in the training set, multiple hypothesis correction was performed using a permutation test as described above. A receiver operating characteristic (ROC) curve was plotted from the test set consisting of the withheld genes, given that the corresponding training samples were statistically significant. This benchmarking procedure was repeated multiple times and a mean ROC curve was generated by vertical averaging.

2.3.3 Biclustering

Independent of the 2-means clustering approach, a second method was employed to detect NSCLC genetic vulnerabilities without reliance on annotated gene sets. The entire RNAi knockdown dataset was represented as a matrix as described above, where each row is a gene knockdown and each column is a cell line. The Large Average Submatrix (LAS) biclustering algorithm was applied to this matrix to uncover biclusters in which the rows (genes) exhibit similar behavior across a set of columns (cell lines) [67]. In particular, the desired biclusters have the property of being large in average value relative to other submatrices of similar size and represent biological systems to which certain NSCLC cell lines are especially dependent. The genes corresponding to each bicluster were then used to query the Database for Annotation, Visualization and Integrated Discovery (DAVID) for functional enrichment [35, 34]. We also searched each bicluster for enrichment of protein complexes by calculating the hypergeometric probability of obtaining at least the observed number of overlap between a complex and the bicluster genes. Statistical significance of complex enrichment was controlled at 5% FDR by the BenjaminiHochberg procedure [6].

2.3.4 Alternative methods for determination of gene set sensitivity

We also considered alternative measures of gene set sensitivity. For every cell line within each gene set, the probability of observing the number of 'hit' genes (gene knockdowns producing non-viability) was computed according to a hypergeometric distribution. The resulting probabilities for each line were then multiplied together to obtain an overall score for the gene set. Statistical significance of the scores was found using a permutation test, as was done for the 2-means clustering.

2.4 Results

From a whole genome RNAi screen of NSCLC cell lines, we identified candidate drug targets in the form of genetic vulnerabilities specific to cell line subgroups. A couple of factors were considered when determining genetic vulnerabilities. First, given the heterogeneity of NSCLC, gene deletions that are almost universally toxic across the cell line panel would likely be toxic to other normal human cells as well. In addition, it is desirable to specifically target lung cancer cells but not normal cells, yet only one cell line in the panel was from a non-cancerous normal lineage. This suggests that in the interest of specificity, the number of cell lines constituting a genetic vulnerability should not be too large. On the other hand, a vulnerability consisting of a single cell line may be less likely to generalize to an appreciable number of patients.

Therefore, the challenge stems from identifying groups of genes to which some, but not all, NSCLC cell lines are especially dependent. Ideally, these cell lines would represent a particular patient subpopulation. Another challenge was to avoid a combinatorially intractable problem of having to examine all possible combinations of cell lines against all possible combinations of genes. Novel applications of unsupervised learning algorithms were developed to overcome these challenges to prioritize potential NSCLC targets from RNAi sensitivities.

The general workflow of this study is outlined in Figure 2.1. From the whole genome RNAi screen on 12 NSCLC cell lines and one normal epithelial line, we extracted knockdown sensitivity profiles for selected gene sets. Each gene set was clustered, scored and ranked by statistical significance. The clustering score measures the degree to which the cell lines segregate into sensitive and resistant groups upon knockdown of genes in the set. Gene sets with a clear segregation of sensitive and resistant lines are termed bimodal. It is imperative that our scoring scheme prioritizes such bimodal sets over other patterns of RNAi sensitivity that would be less desirable as a candidate drug target. For example, gene sets that are all toxic or half-toxic are undesirable due to predicted toxicity beyond those in our cell line panel. In addition, gene sets that are largely resistant or having a random pattern of sensitivity clearly would not be desired as well.

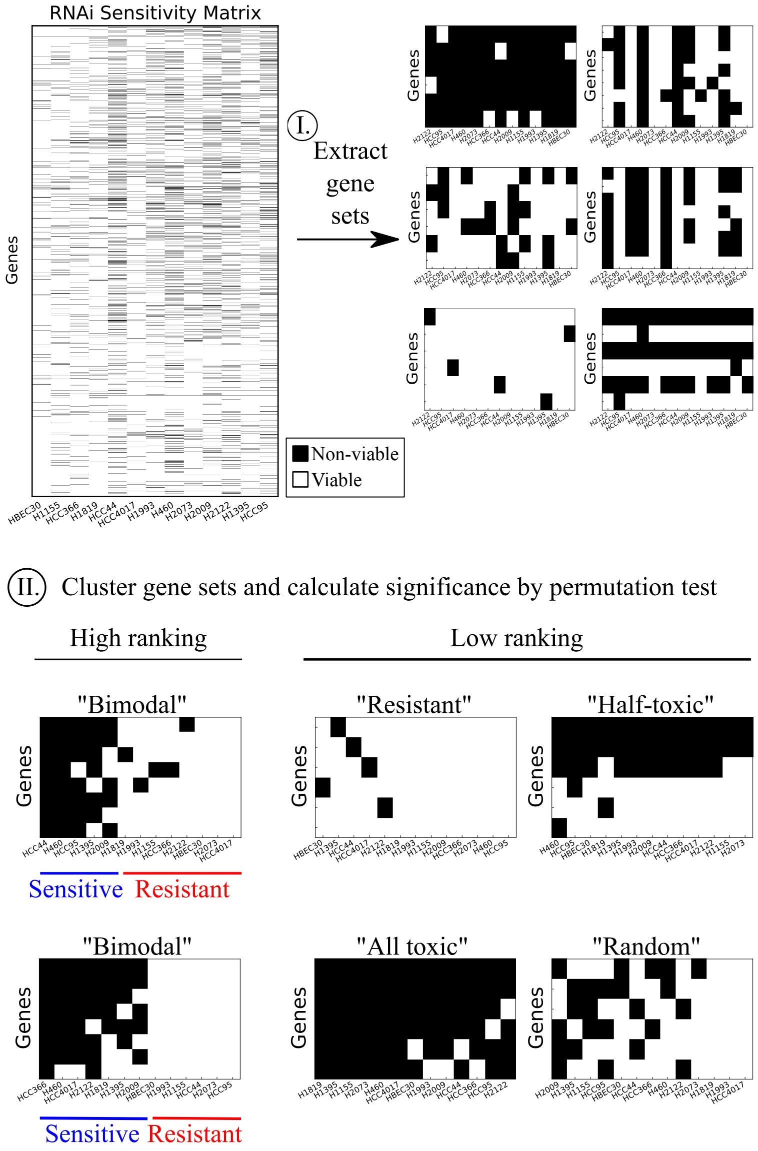


Figure 2.1: Gene sets with bimodal sensitivity represent NSCLC vulnerabilities. RNAi sensitivity profiles were extracted for selected gene sets (six examples shown). A ranking scheme was designed to prioritize gene sets whose knockdown leads to a bimodal response of cell lines.

2.4.1 Subgroup-specific NSCLC vulnerabilities are found among protein complexes

The selected gene sets we chose to examine were 2820 protein complexes. As calculated by the 2-means clustering approach, 35 had statistically significant scores at 10% FDR from a permutation test (Appendix 1). Fisher's exact test also determined 33 of those 35 complexes to be highly statistically significant (Figure 2.2). We found that the standard k-means algorithm and the 1-D optimal k-means method, *Ckmeans.1d.dp*, yielded identical results although *Ckmeans.1d.dp* demonstrated marked runtime speedup. Finally, simulations of a synthetic dataset showed that the permutation test for statistical significance was not biased toward larger or smaller complexes (Figure 2.3).

Overall, the 2-means clustering method found strong genetic vulnerabilities including components of splicing and translation (Figure 2.4). The top-ranking gene sets are protein complexes that all exhibit the desired bimodal behavior, in which one particular group of cell lines is far more vulnerable to loss of components in the complex than the other cell lines. Notably, the HBEC30 normal cell line did not generally show sensitivity to knockdown of any of the top-ranking vulnerabilities.

RNA splicing is a major category of lung cancer vulnerabilities, as evidenced by the RNAi sensitivity patterns of the LSm2-8, 17S U2 snRNP and CDC5L complexes. Components of the translation machinery, represented by the eIF3 complex and ribosome small subunit, constitute another major class of vulnerabilities. Two notable candidate drug targets are the proteasome and

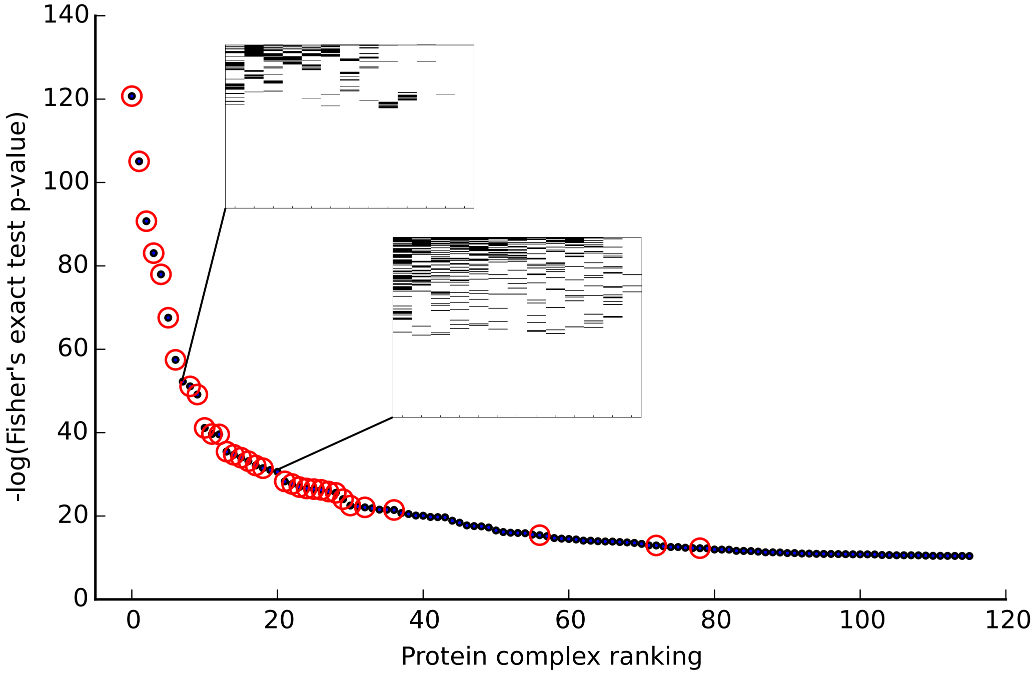


Figure 2.2: Circled in red are 33 of the 35 protein complexes from the permutation test to indicate where they rank according to Fisher's exact test, which discovered 116 significant complexes at 10% FDR. Also shown are two complexes not prioritized by the permutation test.

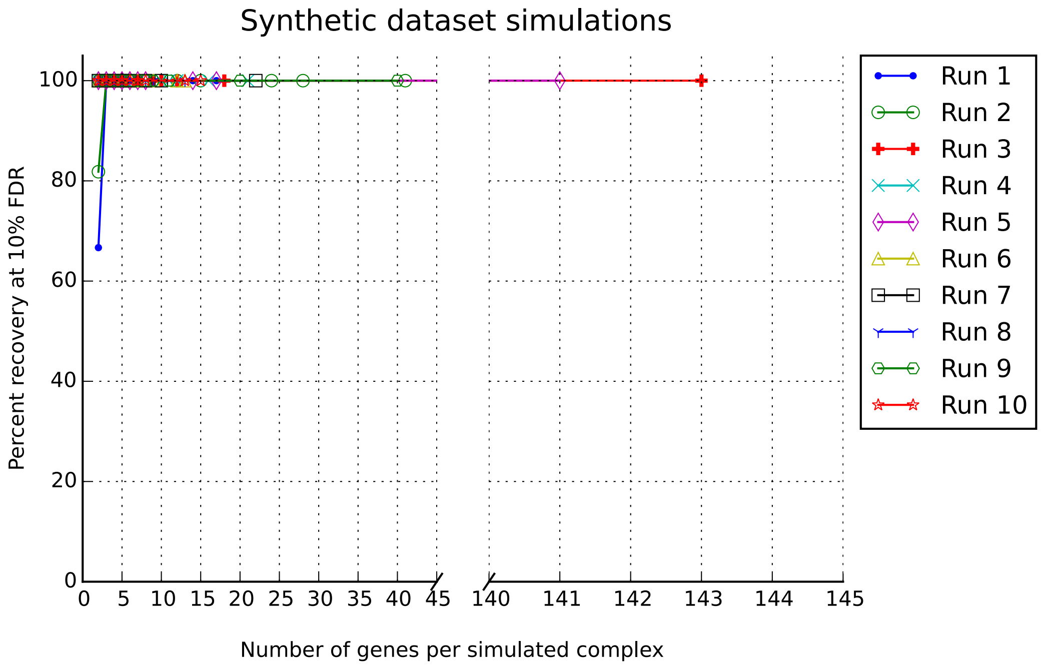


Figure 2.3: **Simulations of 2-means clustering and permutation testing on a synthetic dataset.** To evaluate whether permutation testing for statistical significance of 2-means clustering scores is biased towards larger or smaller complexes, simulations on synthetically generated data were conducted. A 90% random sparse matrix of ones and zeros with the same dimensions as the actual RNAi data (24866 rows and 13 columns) was constructed. Each simulation run consisted of randomly choosing 200 submatrices with the row size of each submatrix drawn from the same size distribution as that of the actual protein complexes. Therefore, each submatrix simulates a protein complex, with the number of genes in the simulated protein complex corresponding to the number of rows in the submatrix. Of those 200 submatrices, 30 were constructed to be bimodal by setting between 4 and 7 columns equal to 1. The 2-means clustering and permutation test were run on all 200 submatrices. The entire procedure just described constitutes one simulation run. For every simulation run, the percentage of bimodal submatrices that were called significant at 10% FDR was calculated for the various complex sizes. Shown is a plot over 10 simulation runs; the permutation scheme approach is not biased toward larger or smaller complexes, particularly for complexes with more than 2 genes.

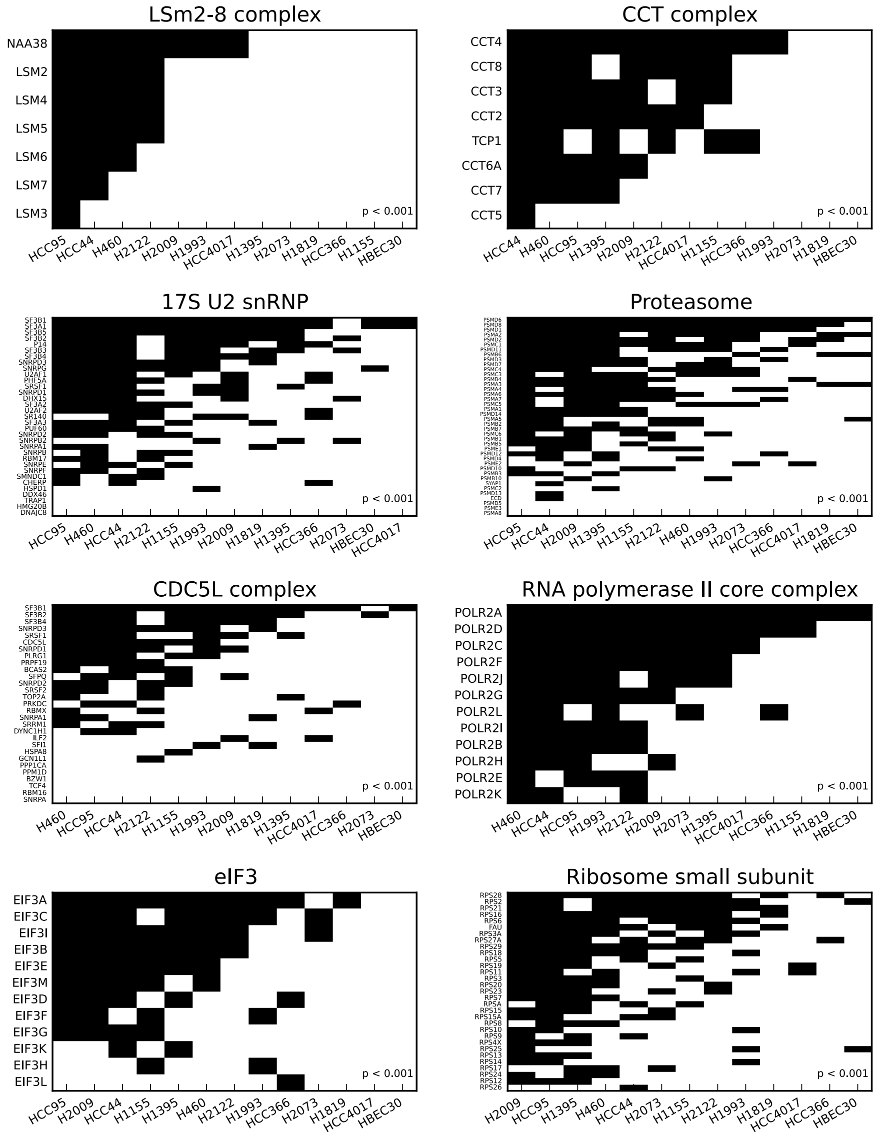


Figure 2.4: k-means clustering uncovers differential essentiality of NSCLC cell lines to protein complexes. The clustering partitions the cell line panel into two groups: sensitive and insensitive to loss of components of the complex. Shown are eight of 35 protein complexes that yielded statistically significant scores (10% FDR)

the CCT/TRiC chaperonin complex, which gives one of the cleanest signals in terms of clustering the cell lines into sensitive and insensitive groups.

Moreover, different cell line subgroups exhibit different sensitivities. For example, HCC95, HCC44, H460 and H2122 are especially vulnerable to loss of the LSm2-8 complex, while a slightly broader cell line set is highly dependent upon the CCT complex. A few cell lines, particularly HCC95, are frequently sensitive to many of the genetic vulnerabilities. For some of the genetic vulnerabilities, the cell lines affected do not belong to a single histological subtype. Rather, they encompass at least one of the three main NSCLC subtypes of adenocarcinoma, squamous-cell and large-cell carcinoma.

In assessing the performance of our 2-means clustering by a leave-one-out strategy, only protein complexes above a certain size were considered. For complexes with at least five members, the 2-means clustering achieves a mean AUC of 0.62, with the average being computed over five iterations of withholding a random gene from each complex. When considering protein complexes containing at least eight members, a mean AUC of 0.66 is attained over eight iterations (Figure 2.5).

We applied our 2-means clustering method to data from Project Achilles on sensitivity of 19 NSCLC cell lines to shRNA knockdown of 5711 genes. Due to the lower coverage of genes compared to our whole genome knockdown dataset, the 2-means clustering is able to be applied to only a portion of many of the protein complexes. According to a permutation test, we were unable to find any statistically significant protein complexes at the same FDR 10% level

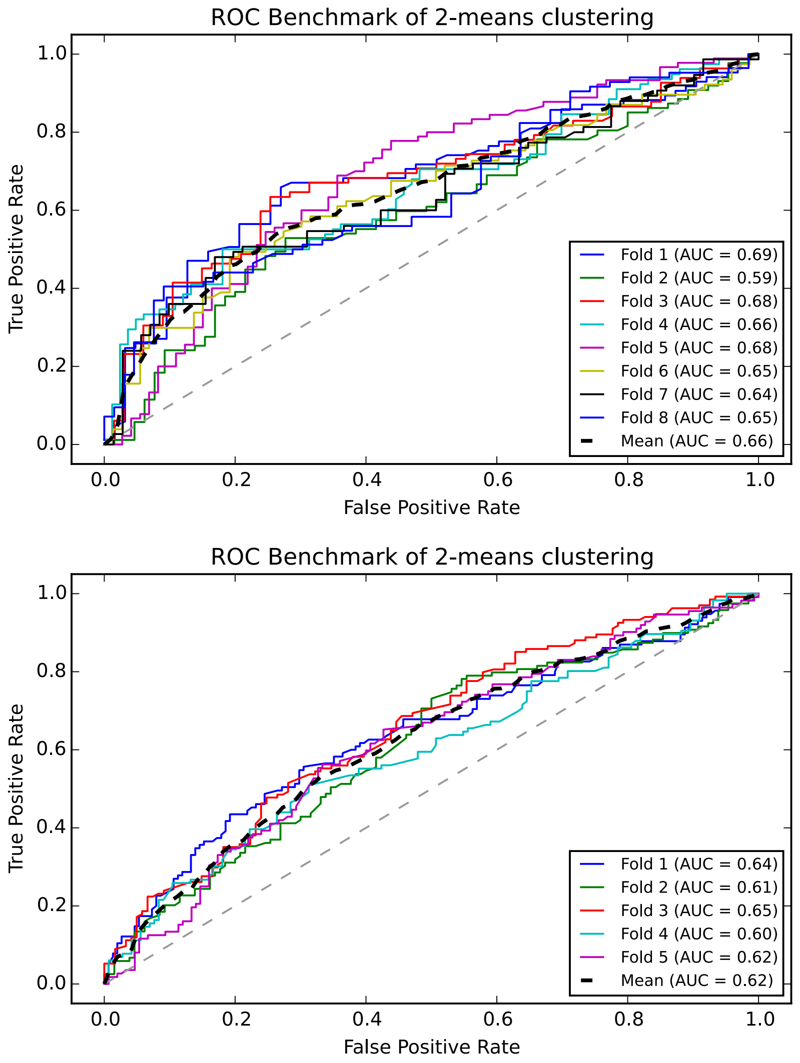


Figure 2.5: **Receiver operating characteristic (ROC) benchmarking of 2-means clustering.** The top plot was obtained from 8 iterations of a leave-one-out benchmarking scheme for protein complexes with at least 8 members. Each iteration, a gene member was randomly withheld from each complex. 2-means clustering of the remaining members predicted average sensitivity vs. resistance of cell lines in the withheld test gene. The bottom plot shows the analogous results for 5 iterations when restricting to complexes with at least 5 members.

previously used. The top scoring result is a ribosomal complex, followed by two proteasome complexes. A majority of the significant protein complexes found from our own RNAi dataset also maintain the general pattern of partitioning into sensitive and resistant cell lines in the Project Achilles experiment (Appendix 1). Because no statistical significance was found, we did not carry out the benchmarking procedure as above.

2.4.2 Biclustering finds genetic vulnerabilities without reliance on annotated gene sets

LAS biclustering was employed as an independent and complementary approach to identifying candidate drug targets. The 2-means clustering approach relies on annotated gene sets, namely protein complexes, to address the challenge of selecting gene groups to interrogate for bimodal response to RNAi knockdown. On the other hand, biclustering offered an alternative strategy to tackle this challenge as it could find genetic vulnerabilities without regard to any prior annotation. The LAS algorithm found 22 statistically significant biclusters with Bonferroni-corrected P-values $< 10^{-5}$. Of the top 10 highest ranking biclusters, the first represents a nearly universally toxic set - all of the lung cancer cell lines are vulnerable to loss of almost any of the genes. The next best-ranking results are those which are toxic only to a single cell line. The lower-ranked statistically significant biclusters tend to represent vulnerabilities for a broader set of cell lines (Table 2.2). Functional enrichment was not found for three of the top 10 results. In addition, searching each bicluster for enrichment of protein complexes yielded heavy enrichment for the spliceosome. For most of the biclusters, the functions of the enriched protein complexes match those found from the DAVID enrichment (Table 2.3).

Bicluster rank	Size (genes × lines)	Lines affected	Enriched annotations
1	1591 × 12	all but HBEC30	Translation, splicing, kinetochores, mitosis
2	756 × 1	HBEC30	No functional enrichment
3	1060 × 1	HCC4017	Translation, splicing, nuclear lumen
4	1141 × 1	HCC366	Nuclear proteins, proteasome non-ATPase subunits
5	1219 × 1	H1819	Translation, splicing
6	813 × 1	H1155	Nucleolar, cytoskeletal proteins
7	1154 × 1	H2073	Wnt pathway
8	859 × 2	H460, H2122	No functional enrichment
9	920 × 1	H1395	Translation
10	1254 × 1	H1993	No functional enrichment
11	1629 × 1	HCC95	Translation, splicing, lysosomal ATPase
12	450 × 2	HCC44, H2009	No functional enrichment

Table 2.2, Continued on next page

Bicluster rank	Size (genes × lines)	Lines affected	Enriched annotations
13	104 × 9	9 lines	Ribosome, splicing, nuclear lamin, proteasome
14	159 × 2	H460, H2009	No functional enrichment
15	119 × 2	H2009, H2122	No functional enrichment
16	216 × 2	HCC44, H460	No functional enrichment
17	177 × 2	HCC44, H2122	No functional enrichment
18	66 × 3	H1155, H2073, H1395	No functional enrichment
19	74 × 4	H1155, H2009, H1395, HCC95	Ribosome, proteasome, COPI transport
20	120 × 3	H1993, H2073, H1395	No functional enrichment
21	14 × 5	5 lines	No functional enrichment
22	20 × 7	7 lines	Splicing

Table 2.2: All statistically significant biclusters (Bonferroni-corrected $P < 10^{-5}$) discovered from Large Average Submatrix (LAS) biclustering. Functional enrichment as discovered through DAVID also shown.

Bicluster rank	Number of genes in bicluster	Number of enriched protein complexes	Significant enrichments (P-value)
1	1591	71	Spliceosome (2.1×10^{-43}); 40S ribosomal subunit (2.3×10^{-16})
2	756	4	20S proteasome (5.5×10^{-5})

Table 2.3, Continued on next page

Bicluster rank	Number of genes in bicluster	Number of enriched protein complexes	Significant enrichments (P-value)
3	1060	30	Spliceosome (1.5×10^{-15}); Nop56p-associated pre-rRNA complex (5.4×10^{-11}); NRD complex (8.0×10^{-6})
4	1141	21	NuA4/Tip60-HAT complex B (2.1×10^{-5}); PA700-20S-PA28 complex (1.7×10^{-6})
5	1219	24	Spliceosome (2.9×10^{-13})
6	813	2	109-member metabolic and motor protein complex (1.1×10^{-4})
7	1154	2	PID complex (1.4×10^{-4})
8	859	10	Spliceosome (1.5×10^{-11})
9	920	5	140-member ribosomal protein complex (1.8×10^{-9})
10	1254	0	No enrichment
11	1629	25	Spliceosome (1.8×10^{-17})

Table 2.3, Continued on next page

Bicluster rank	Number of genes in bicluster	Number of enriched protein complexes	Significant enrichments (P-value)
12	450	1	SAR1A-TPD52 complex (8.9×10^{-5})
13	104	50	Spliceosome (8.6×10^{-13})
14	159	2	Nop56p-associated pre-rRNA complex (5.2×10^{-7})
15	119	0	No enrichment
16	216	18	Spliceosome (4.4×10^{-6})
17	177	0	No enrichment
18	66	0	No enrichment
19	74	23	PA700-20S-PA28 complex (2.8×10^{-18}); 40S ribosomal subunit (5.1×10^{-7})
20	120	0	No enrichment
21	14	0	No enrichment
22	20	3	C complex spliceosome (4.2×10^{-8})

Table 2.3: **Protein complex enrichment of bicluster genes discovered from LAS biclustering.** Enrichment was computed from the hypergeometric probability of obtaining at least the observed amount of overlap between protein complex and bicluster genes. All enriched protein complexes are statistically significant at 5% FDR as determined by the Benjamini-Hochberg procedure. For brevity, statistically significant complexes with biological functions also found to be enriched with DAVID are shown.

Many of the protein complexes from 2-means clustering also participate in biological processes found in LAS biclustering (Table 2.2). In particular, there appears to be frequent enrichment for translation and splicing, which are the functions of the ribosome small subunit, and the LSm2-8, CDC5L and 17S U2 snRNP complexes. No functional enrichment was found for the bicluster genes affecting HBEC30, which was also often resistant to knockdown of the protein complexes prioritized by 2-means clustering. Interestingly in 2-means clustering, HCC4017, HCC366, and H1819 were mostly among the groups of resistant cell lines although in biclustering, their genes were enriched in translation, splicing and proteasome components. Upon closer examination, the genes responsible for this enrichment are different from those comprising the translation, splicing and proteasome protein complexes. Apparently, biclustering complements the 2-means clustering in uncovering certain genetic vulnerabilities not found by the latter. We also note that the Wnt pathway, which is enriched in the 7th-ranked bicluster, was not discovered by the 2-means clustering as many of its genes either did not appear in our protein complex set or were only present individually in single complexes.

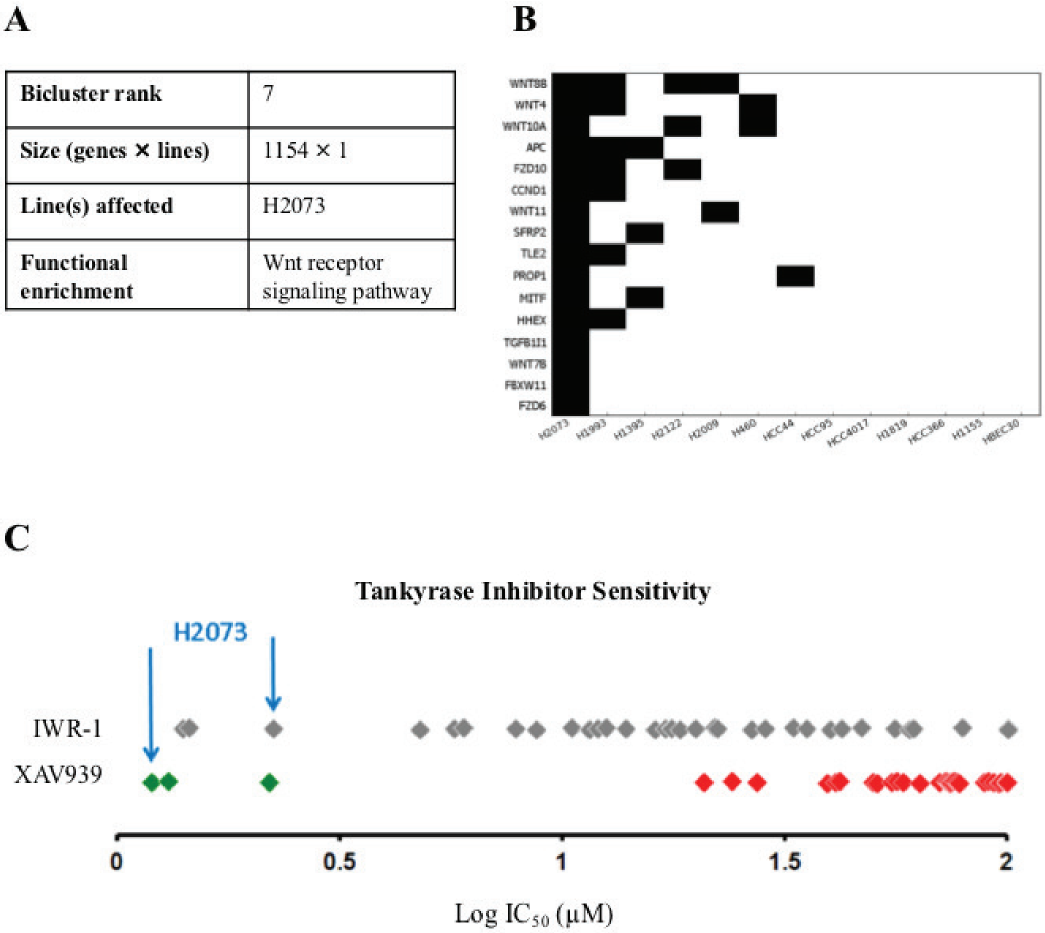
2.4.3 Small-molecule screen confirms predicted cell line sensitivity

One notable bicluster showed enrichment for the Wnt pathway (Figure 2.6A and B). The H2073 adenocarcinoma cell was highly vulnerable to loss of Wnt pathway members. This suggests that small-molecule compounds targeting Wnt should reproduce the RNAi gene knockdown sensitivity pattern when tested on the cell line panel. Two Wnt inhibitors, IWR-1 and XAV939, were screened on a larger group of cell lines encompassing the panel, and the results confirmed the predicted sensitivity (Figure 2.6C). The two compounds had a selective deleterious effect on H2073 while essentially sparing the other cell lines. Each cell line denoted by a diamond was colored according to a normal mixture model that predicted the number of groups. If there were two groups, green and red were used for sensitive and resistant, respectively. Otherwise, the diamonds were colored gray.

2.4.4 Alternative approaches to measuring complex sensitivity do not prioritize bimodality

We evaluated several other methods to measure complex sensitivity. These approaches depended on annotated gene sets, as opposed to biclustering, which has no such constraints. Cell line viability results from our whole genome screen are not approximately normally distributed (Figure 2.7), which precludes the use of a simple z-test comparing the complex members' scores to the background distribution. Even if the data were normally distributed (as the Project Achilles data is), this method would not distinguish bimodal from half-toxic complexes (Figure 2.8), and in fact would prioritize universally toxic complexes.

We also considered using the hypergeometric distribution to assess the significance of multiple occurrences of sensitivity within a protein complex. From a permutation test, we found 544 statistically significant protein com-



Bicluster rank	7
Size (genes × lines)	1154 × 1
Line(s) affected	H2073
Functional enrichment	Wnt receptor signaling pathway

Figure 2.6: Biclustering finds strong vulnerability of H2073 to loss of Wnt signaling. (A) The seventh-ranked bicluster, containing 1154 genes, is enriched for the Wnt pathway. (B) The sensitivity profile of the gene set comprising the functional enrichment shows sensitivity of H2073 to knockdown of any of the genes in the set. (C) Screening of IWR-1 and XAV939 against an expanded panel of NSCLC cell lines (each denoted by a diamond) indeed shows that H2073 is markedly vulnerable to chemical inhibition of the Wnt pathway. Figure courtesy of Michael Peyton.

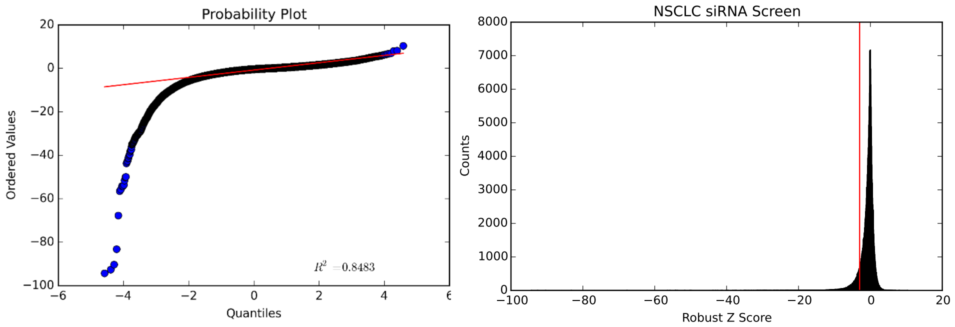


Figure 2.7: Q-Q plot (left) shows that the robust Z scores from the whole genome siRNA screen are not normally distributed. The score distribution (right) is skewed; scores to the left of the red line indicate non-viability.

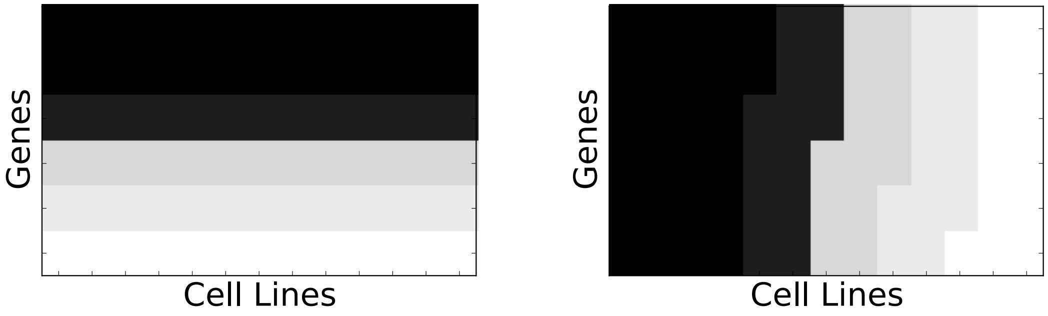


Figure 2.8: Even if the data were normally distributed, a z-test comparing complex sensitivity to the background distribution would not distinguish a half-toxic complex (left) from a bimodal complex (right).

plexes at FDR 10% (Figure 2.9). With the large number of protein complexes being statistically significant, we felt that this method was less discriminative than the 2-means clustering approach in prioritizing complexes.

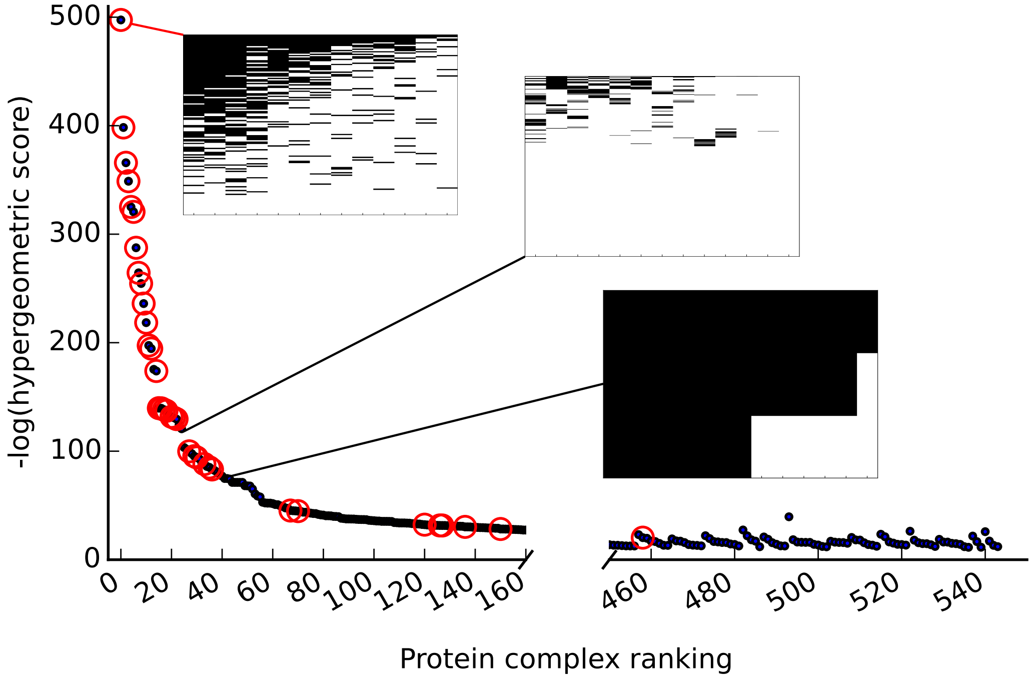


Figure 2.9: Circled in red are the 35 protein complexes from 2-means clustering to indicate where they rank according to an alternative hypergeometric scoring method, which discovered 544 significant complexes. Both methods used a permutation test and q-value to control the FDR at 10%. Also shown are two complexes not prioritized by 2-means clustering.

2.5 Discussion

Collectively, the protein complexes we discovered to be NSCLC genetic vulnerabilities span various cellular processes including splicing, translation and protein folding. It is natural to ask how they fit in with currently established cancer therapies and whether known drugs could be repurposed for these complexes. Clearly, they contrast with hormonal therapy or the usual mitotic targets of cytotoxic chemotherapy. It turns out that some of the strongest genetic vulnerabilities are known targets of small-molecules.

Arsenic trioxide (As_2O_3) targets the TRiC/CCT complex [54] and has been used to treat acute promyelocytic leukemia in patients who did not respond well to other types of chemotherapy [70, 74, 73]. As_2O_3 can perhaps be repurposed for NSCLC, particularly for patients whose tumors bear similarity to the sensitive cell line subgroups identified from the 2-means clustering analysis. Several studies have evaluated the effect of As_2O_3 in human lung primary fibroblasts and in the lung cancer cell lines A549 and H460 [48, 56]. Collectively, they suggest that H460 is markedly more sensitive to As_2O_3 than lung fibroblasts, consistent with the CCT complex vulnerability we observed.

We also identified the proteasome as a candidate NSCLC drug target and recently, proteasome inhibitors have been investigated as anti-cancer agents. One such inhibitor is Velcade (bortezomib), which has been FDA-approved for multiple myeloma [42]. Bortezomib has shown to be effective in combination with other chemotherapy agents for NSCLC [19] and has been evaluated in clinical trials for NSCLC as well [7, 39, 61]. This also suggests that newer and more specific proteasome inhibitors, such as Kyprolis (carfilzomib) could be efficacious for patients with NSCLC.

In addition, translation and splicing emerged as strong genetic vulnerabilities from the 2-means analysis. Previously, translation has been proposed as a potential target in cancer [28]. Moreover, eIF3 is known to be overexpressed in lung cancers [60], and ectopic expression of five eIF3 subunits has been shown to transform immortalized fibroblasts into malignant cells [88]. Notably, in our study we found that knockdown of four of those five subunits strongly sensitizes six of the 12 NSCLC cell lines in our panel, while an immortalized epithelial line is comparatively unaffected (Figure 2.4). The splicing apparatus has been suggested as a cancer target as well [27, 80]. Of the splicing-associated protein complexes discovered from the 2-means analysis (Figure 2.4), the SF3b component of U2 snRNP is known to be targeted by a number of small-molecule compounds. Both the pladienolides and meayamycin target SF3b, and the latter has been shown to be more deleterious in human lung cancer cells than normal lung fibroblasts [2, 10].

Some of the NSCLC genetic vulnerabilities that were found by our computational analysis include protein complexes that may appear to be entirely essential. It is perhaps surprising that certain cell lines are largely resistant to knockdown of many of these genes. One explanation may simply be a result of the strict thresholding of the RNAi data to produce binary readings of cell line viability, which could be affected by the $<100\%$ sensitivity of the assay. Another explanation may be provided by essential gene "moonlighting" and "flipping" of protein complex essentiality between distantly related species [65]. It was shown that certain protein complexes almost completely flip essentiality between *Saccharomyces cerevisiae* and *Schizosaccharomyces* *pombe.* A similar phenomenon may be occurring among our NSCLC cell line panel. Although the cell lines are not necessarily distantly related, they likely differ sufficiently due to different mutational compositions. Different yeast species flip protein complex essentiality as a result of adaptations to differing needs and environments, a phenomenon likely common to cancer cells as well. Moreover, particular NSCLC cell lines are largely resistant to loss of most, but not all, members of certain protein complexes. Those genes that are mostly essential in both sensitive and insensitive cell line subgroups could exhibit “moonlighting” behavior by having multiple functions in both essential and nonessential processes.

The NSCLC genetic vulnerabilities uncovered by the computational analysis described here extends an earlier study [41] in uncovering additional potential targets for therapy that were not previously reported. While our study shares the general aim of identifying genetic vulnerabilities, we exclusively focus on identification of biological systems, such as protein complexes, that certain lung cancers are especially dependent upon. Our results also present a complementary viewpoint to the Project Achilles effort in analyzing whole genome RNAi knockdown of cancer cells. One goal from Project Achilles was to discover genes that simultaneously had partial copy number loss and were essential to cancer cells [53]. Interestingly, the results from that analysis found single gene vulnerabilities in splicing, translation and the proteasome as well. One such vulnerability was LSM4, which is a part of the LSm2-8 complex. A key difference is that the NSCLC vulnerabilities presented here are from the viewpoint of looking not only at single genes but biological systems such as protein complexes. In contrast with previous analyses, we obtain cell line subgroups that may represent particular patient populations along with candidate targets for each of those subgroups.

2.6 Summary

Novel candidate drug targets were found from computational analysis of a whole genome RNAi knockdown screen in NSCLC cell lines. The targets are protein complexes specific for particular lung cancer cell lines and function in splicing, translation and protein folding. Results of previous studies support further investigation of these protein complexes as avenues of therapeutic intervention in NSCLC. Moreover, the candidate targets provide an opportunity for drug repurposing, which could lead to reduced time in the drug development pipeline. Our results simultaneously establish lung cancer cell line subgroups and potentially novel druggable targets that are specific to each subgroup. This study contributes to a deeper understanding of therapeutically relevant events at the molecular scale in NSCLC.

2.7 Funding

This work was supported by grants from the Cancer Prevention and Research Institute of Texas (CPRIT). E.M.M. acknowledges funding from the National Institutes of Health; National Science Foundation; and the Welch Foundation [F1515].

Conflict of Interest: none declared.

2.8 Acknowledgments

We thank members of the University of Texas Southwestern Medical Center and Baylor College of Medicine Lung Cancer Consortium for sharing whole genome RNAi screening data.

Chapter 3

Predictability of Genetic Interactions from Functional Gene Modules[1]

3.1 Abstract

Characterizing genetic interactions is crucial to understanding cellular and organismal response to gene-level perturbations. Such knowledge can inform the selection of candidate disease therapy targets. Yet experimentally determining whether genes interact is technically non-trivial and time-consuming. High-fidelity prediction of different classes of genetic interactions in multiple organisms would substantially alleviate this experimental burden. Under the hypothesis that functionally-related genes tend to share common genetic interaction partners, we evaluate a computational approach to predict genetic interactions in *Homo sapiens*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae*. By leveraging knowledge of functional relationships between genes, we cross-validate predictions on known genetic interactions and observe high-predictive power of multiple classes of genetic interactions in all three organisms. Additionally, our method suggests high-confidence candidate

[1] Published as Young JH and Marcotte EM. Predictability of Genetic Interactions from Functional Gene Modules. *bioRxiv* (2016): 049627 and submitted to *G3 (Bethesda)*. J.H.Y. designed the study, performed data analysis, and wrote the paper.

interaction pairs that can be directly experimentally tested. A web application is provided for users to query genes for predicted novel genetic interaction partners. Finally, by subsampling the known yeast genetic interaction network, we found that novel genetic interactions are predictable even when knowledge of currently known interactions is minimal.

3.2 Introduction

Determining the genetic interactions in an organism provides a basis for understanding how the role of a gene is influenced by the action of any other gene. By definition, two or more genes interact when combining variants of each gene produces a significantly pronounced phenotype when compared to the phenotypes of individual variants [52, 5]. The applications of exploiting such interactions extend to drug target discovery. Strategies such as targeting genes that interact with cancer-specific mutations have been proposed and reviewed extensively [4, 23] and have led to clinical trials [24]. Because experimental determination of genetic interactions involves examining all possible pairs from a group of genes, practical difficulties arise when a comprehensive interaction map of an entire organism is desired. Multicellular organisms present the challenge of various differentiated cell types, each having potentially differing genetic interactions. Moreover, there are different kinds of genetic interactions, ranging from those based on growth effects to other phenotypic effects. There exists a need to either reduce the search space for testing genetic interactions or to reliably predict them. Here, we evaluate a computational approach to predict and validate different types genetic interactions across multiple organisms.

Previous studies to predict genetic interactions leveraged existing sources of biological information. Integration of biological features in yeast (i.e. gene co-expression, protein interaction and function) and their associated network topological properties guided the training of probabilistic decision trees to predict synthetic sick or lethal (SSL) interactions [84]. In a similar vein, an ensemble classifier was trained on a set of 152 genetic interaction-independent features to predict SSL in yeast [55]. Compiling multiple biological features has also been extended to more than one organism. By considering the orthologous gene pairs among yeast, fly and worm, features such as functional annotation were used to train a logistic regression model to predict a genome-wide map of genetic interactions [89]. Alternatively, studies have also explored network-based approaches for genetic interaction prediction. Novel SSL interactions were predicted by way of a diffusion kernel on a network of known SSL gene pairs [62]. Interrogating functional gene networks that were constructed from integration of biological data from literature have proven useful in predicting modifier genes in yeast and worm [46]. Many of these approaches have focused on a single genetic interaction type in a single organism.

Here, we examine an algorithm to predict multiple types of genetic interactions across diverse organisms based on the hypothesis that genes strongly participating in shared functions also share common genetic interaction partners. Our approach relies on a functional gene network for a given organism and knowledge of known genetic interactions of a particular type. We tested our approach on three organisms - human (*Homo sapiens*), fly (*Drosophila melanogaster*), and yeast (*Saccharomyces cerevisiae*) - and found predictability across different types of genetic interactions. We also investigated how some interactions are enriched in yeast and human gene modules, specifically protein complexes, and the degree to which genetic interactions need to experimentally determined before enrichment can be found.

3.3 Materials and Methods

For various classes of genetic interactions in human, fly, and yeast, a list of genes and each of their known genetic interaction partners were assembled. A gene and its known interaction partners are collectively referred to as a "seed set." Receiver operating characteristic (ROC) analysis was performed to quantify whether the interaction partners of any given gene are clustered in the organism's functional gene network. Specifically, for every group of interaction partners of a gene, a score vector consists of entries that are sums of functional network edge weights between each gene in the network to the interaction partners. Because there are no self-edges in the network, leave-one-out cross-validation is carried out on the known interaction partners. An accompanying label vector indicates whether each gene in the network is indeed an interaction partner. The two vectors yield a ROC curve and the corresponding area under the curve (AUC). A seed set's AUC is the measure of how tightly connected the interaction partners are in the functional network and therefore how predictive the seed set is for novel interactions [46]. None of the known genetic interactions used for prediction were contained in the functional gene network.

Enrichment of genetic interactions within yeast and human protein complexes was calculated with a binomial model defined as $P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$, where the background probability p equals the proportion of all possible gene pairs that are genetically interacting. The number of trials n is the number of possible gene pairs in the complex, and k equals the number of interacting pairs in the protein complex.

3.3.1 Statistical Analysis

If k is the number of genetic interactions within a protein complex, then the corresponding p-value is $P(X \geq k)$ according to a binomial model as previously described, with control of FDR at 5% through the Benjamini-Hochberg procedure [6]. Seed sets with AUC $\geq$ 0.9 were considered highly predictive of novel genetic interactions.

3.3.2 Data Availability

All genetic interactions were downloaded from version 3.4.130 of BIOGRID [75]. Organism-specific functional gene networks were downloaded for human [45], fly [71], and yeast [47]. Previous studies served as sources of protein complexes for yeast and human [30, 64]. Python code using the Matplotlib [37], scikit-learn [59], and *mygene* [86] libraries is available at `https://bitbucket.org/youngjh/genetic_interact`. All network visualizations were produced in Cytoscape [68]. A supplementary web page at `http://marcottelab.org/Genetic_Interact/` allows users to query a gene of interest. If the gene has known genetic interaction partners that are predictive, then the functional network cluster is displayed. Raw data files listing the seed sets with AUC ≥ 0.7 are also available.

3.4 Results

We sought to determine whether clusters of functionally related genes, for example genes A-E in Figure 3.1, are predictive of genetic interactions. In this example, genes A and C-E are known to share genetic interactions with gene X, and our hypothesis would suggest gene B as a novel interaction partner of X. Our method identifies predictive clusters by leave-one-out cross-validation and receiver operating characteristic (ROC) analysis; when applied to the network in Figure 3.1, each of genes A and C-E are individually withheld as known interaction partners one at a time and predicted back with high recall. Subsequently, gene B is a novel high-confidence predicted interaction partner of X. The approach described here was evaluated for several classes of phenotypic and growth-based genetic interactions in human, fly and yeast.

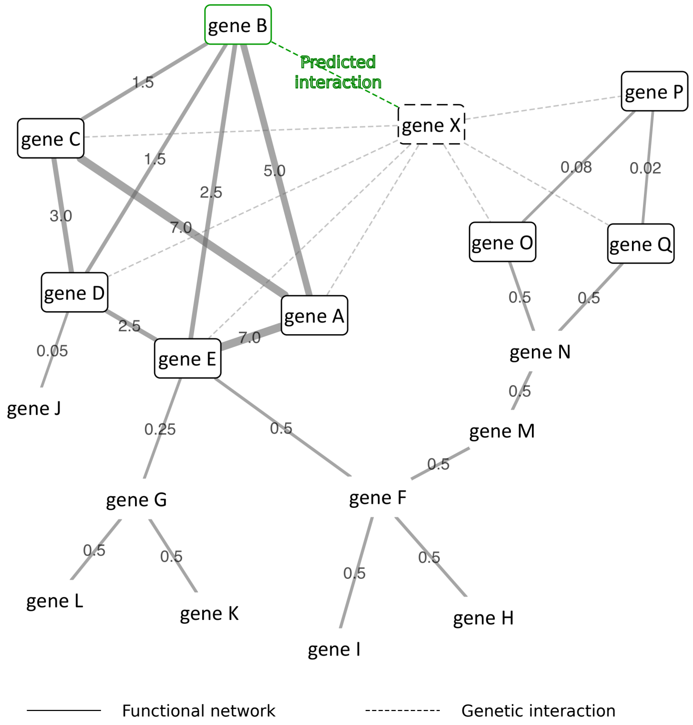


Figure 3.1: **Genetic Interaction Prediction.** Dashed edges indicate known genetic interactions. Solid edges connect genes that participate in the same biological process, with log-likelihood (LLS) scores as edge weights reflecting the degree of confidence in the genes' shared functionality. Genes *A, C-E* are genetic interaction partners of gene X and members of a functional net cluster; then the remaining cluster member, gene B, is a predicted interaction partner of gene X as well. Candidate clusters are evaluated by first assigning scores to each gene in the network by summing the edge weights, as shown in the first row of the matrix. $LLS_{g,A}$ denotes the log-likelihood score between genes g and A. The second row is populated with binary labels indicating whether the gene is a known interaction partner of X. In this fashion, a ROC curve is constructed to yield an AUC.

3.4.1 The human functional gene network is predictive for phenotype-based genetic interactions

As shown in Figure 3.2A, our method demonstrated high performance in predicting phenotypic enhancing and suppressing human gene pairs. In these interactions, a double mutant has an enhanced or suppressed phenotype (other than growth) in comparison to either of the single mutants. The plots for phenotypic enhancement and suppression in Figure 3.2A display the performance of seed sets, each of which are defined as a group of known phenotypic enhancing or suppressing partners of a particular gene. There are 238 phenotypic enhancement seed sets, of which 30 have AUC ≥ 0.9. Similarly, 36 of 215 phenotypic suppression seed sets have AUC ≥ 0.9. The AUC is the area under the receiver operating characteristic (ROC) curve that measures how well the known interaction partners rank in our leave-one-out cross-validation scheme. Those that are not predictive are the ones with AUC $= 0.5$, indicating that their predictability is no better than random. For the most part, seed sets are either at least moderately predictive, or not at all.

Shown in Figure 3.2B are illustrative seed sets with high predictability that form well-defined clusters in the human functional gene network, HumanNet. For clarity, only functional network edges with log-likelihood scores (LLS) above 3.0 are shown. Furthermore, HumanNet genes are shown only if they connect to at least 2 of the known genetic interaction partners. The seed set consisting of the SNW domain containing 1 in phenotypic enhancement with members of the SMAD family and nuclear receptor coactivators yielded

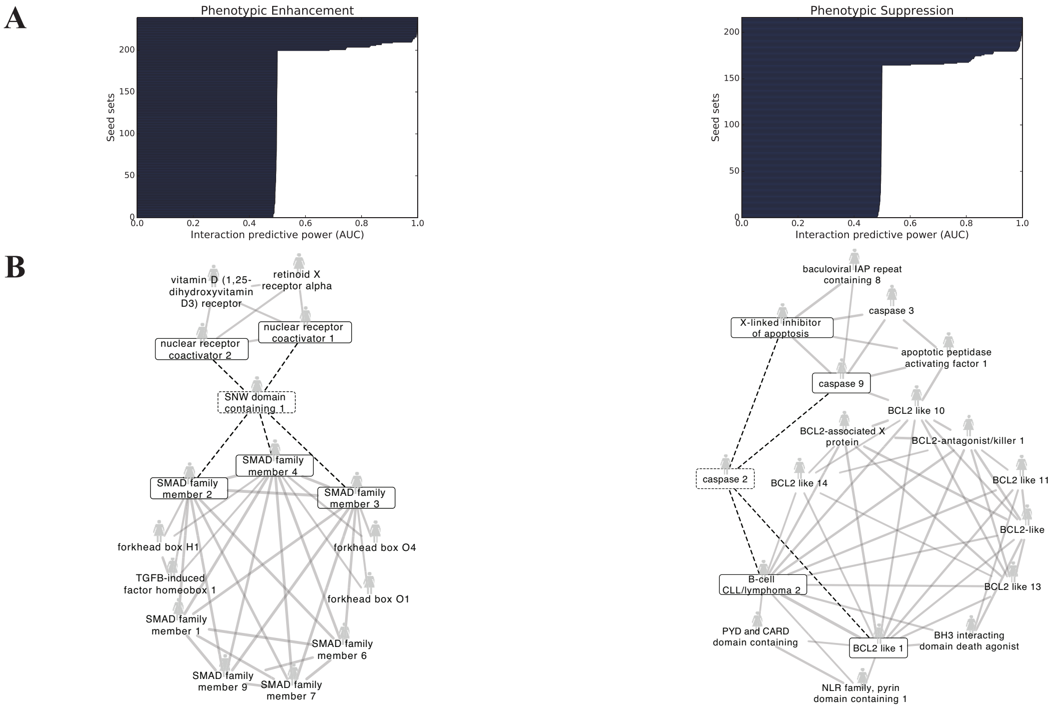


Figure 3.2: **Predictive functional net clusters yield novel phenotypic enhancing and suppressing human gene pairs.** (A) Each horizontal bar represents the set of known genetic interaction partners of a specific human gene; each of these sets is referred to as a "seed set." High AUC scores indicate that the interaction partners participate together in a cluster in HumanNet, the human functional gene network. Therefore, other members of the cluster are predicted as novel interaction partners. (B) Shown are two examples of well-defined HumanNet clusters that are highly predictive for phenotypic enhancement (left) and suppression (right), with the known interactions from the seed set denoted by the boxed genes and dashed edges.

an AUC of 0.91. The prediction is that the SNW domain containing 1 also phenotypically enhances with other members of the SMAD family along with members of the forkhead box. In the phenotypic suppression case, we find that known phenotypic suppressors of caspase 2 are tightly functionally linked with members of the BCL2-like family, among other genes. With a resulting AUC of 0.90, these BCL2-like genes are expected to participate in phenotypic suppression with caspase 2.

3.4.2 Fly phenotypic enhancement and suppression interactions are predicted from functional net clusters

Similar to the human case, the fly functional network FlyNet is particularly predictive of phenotypic enhancement and suppression, as shown in Figure 3.3. A larger proportion of the seed sets are predictive than in the human case. For phenotypic enhancement, 322 out of 754 seed sets had AUC ≥ 0.9, and 398 phenotypic suppression seed sets (out of 818) met the same threshold. Figure 3.3B shows a well-defined gene cluster (AUC = 0.94) containing phenotypic enhancement interaction partners of seven up. From this cluster, genes involved in the sevenless signaling and the Drosophila epidermal growth factor receptor signal transduction pathways achieved high recall, and neighbor genes also involved in the same signaling pathways are expected to phenotypically enhance seven up. Turning to phenotypic suppression, several Enhancer of split genes are tightly clustered (AUC = 0.98) with known phenotypic suppressors of hairy that include the *achaete-scute* complex, thereby implicating them as additional, novel phenotypic suppressors of hairy.

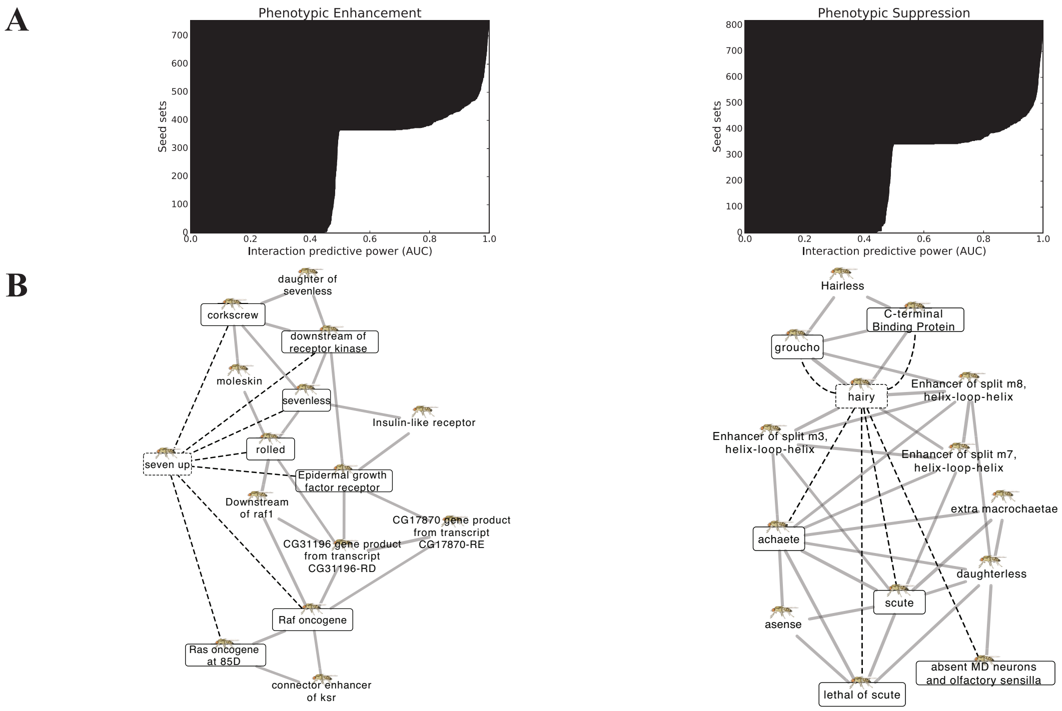


Figure 3.3: **FlyNet predictability for phenotypic enhancing and suppressing genetic interactions.** (A) Each horizontal bar represents a single fly gene that is known to interact with a number of other genes. (B) Predictive seed gene sets are shown for phenotypic enhancement (left) and suppression (right).

3.4.3 High-confidence predictability is found in human, fly and yeast

The full range of various genetic interaction classes that were analyzed from BIOGRID are listed in Table 3.1. Genetic interactions were generally based on phenotypic effects or growth and lethality measurements. Each entry in Table 3.1 lists the number of predictive seed sets having AUC ≥ 0.9 of out the total examined. In human, our method performed well primarily for phenotypic enhancement and suppression as described above, but did not offer predictability for the dosage lethality and synthetic growth defect and rescue interactions determined to date. For fly, most of the known interactions fall into the phenotypic enhancement and suppression categories, for which high predictability was observed. Although a moderate number of fly dosage rescue interactions are known, no predictive seed sets were found. In both human and fly, several classes of interactions have not been extensively determined and thus were untested in our prediction scheme.

Our method also performed well in most of the interaction categories for *S. cerevisiae* (Table 3.1, Appendix 2). Notably, negative and positive genetic interactions fared poorly as few predictive seed sets were identified, even though most of the experimentally determined interactions in yeast fall into these categories.

	H. sapiens	*D. melanogaster*	*S. cerevisiae*
Dosage Growth Defect	Not tested	Not tested	${}^{176}/_{1146}$
Dosage Lethality	${}^{2}/_{108}$	Not tested	${}^{116}/_{689}$
Dosage Rescue	${}^{5}/_{65}$	${}^{0}/_{144}$	${}^{203}/_{1358}$
Phenotypic Enhancement	${}^{30}/_{238}$	${}^{322}/_{754}$	${}^{287}/_{1958}$
Phenotypic Suppression	${}^{36}/_{215}$	${}^{398}/_{818}$	${}^{223}/_{1751}$
Synthetic Growth Defect	${}^{4}/_{445}$	${}^{1}/_{5}$	${}^{576}/_{3417}$
Synthetic Rescue	${}^{2}/_{131}$	${}^{5}/_{26}$	${}^{218}/_{2089}$
Synthetic Lethality	Not tested	Not tested	${}^{221}/_{2706}$
Negative Genetic	Not tested	Not tested	${}^{65}/_{4618}$
Positive Genetic	Not tested	Not tested	${}^{55}/_{3586}$

For each fraction, the numerator indicates the number of seed sets with $\mathrm{AUC} \geq 0.9$ and the denominator equals the total number of seed sets tested.

Table 3.1: Predictive power of functional networks across different genetic interactions.

3.4.4 Trends of interaction enrichment within gene modules vary with interaction type

With genetic interactions predicted across multiple organisms, it was natural to investigate their evolutionary conservation. In particular, if a protein complex were enriched in genetic interactions, then perhaps a homologous protein complex would also exhibit similar enrichment. We found enrichment of various types of interactions within yeast protein complexes, but none thus far for human. Therefore, instead the problem shifted to identifying the degree to which genetic interactions must be determined in order to find enrichment, and therefore predictability. Using yeast as a test case, simulations successively withheld increasing proportions of genetic interactions, with enrichment within yeast protein complexes computed at each point. The interaction types considered were negative and positive genetic, and synthetic growth defect and lethality. As shown in Figure 3.4, when withholding genes with a genetic interaction degree (the number of interacting partners of a certain gene) of more than 5, corresponding to withholding >90% of synthetic growth defect and >80% of synthetic lethality pairs, then an immediate drop-off in enrichment resulted. No such behavior was observed for negative and positive genetic interactions, for which enrichment linearly decreased as a function of the withheld proportion. Similarly, when removing interacting pairs at random, there was a steady decrease in the number of significantly enriched complexes among all types. Finally, when withholding pairs under a degree cutoff, there was also no point beyond which enrichment failed to be found (3.5).

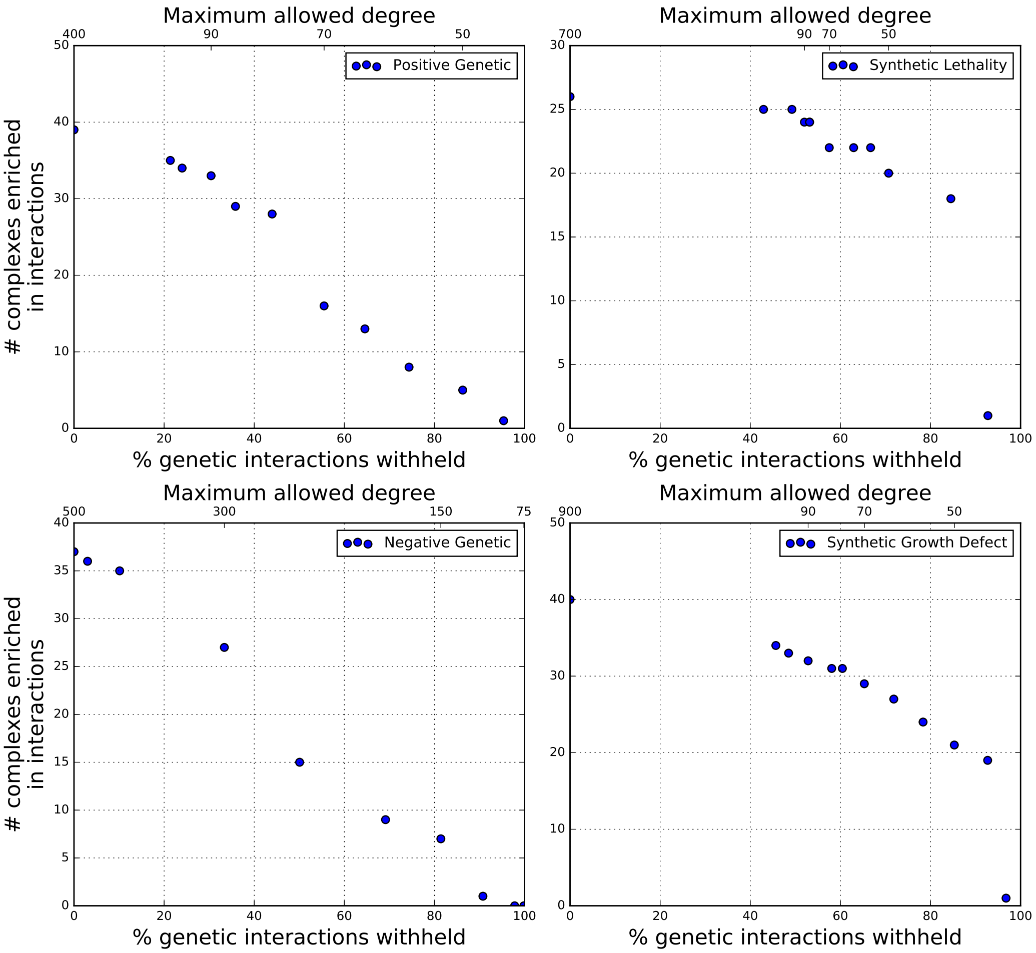


Figure 3.4: **Predictability of genetic interactions can be found even when known interactions are sparse.** By successively withholding known yeast genetic interactions according to each gene's interaction degree (e.g. number of interaction partners), enrichment and therefore predictability is still detectable when information of known interactions is minimal. This effect is especially pronounced for synthetic growth defect and lethality, provided genes possess sufficiently high interaction degree.

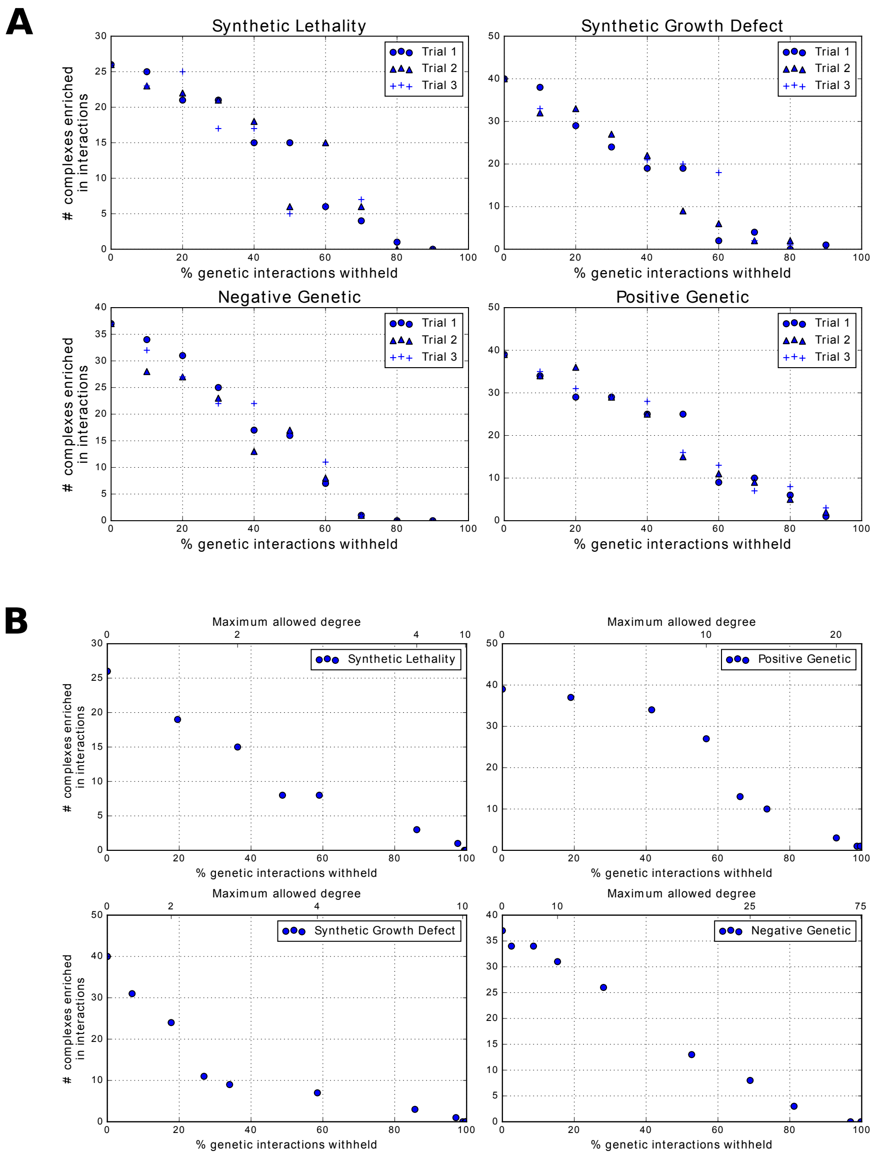


Figure 3.5: **Alternative subsampling methods find predictability even when known genetic interactions are sparse.** (A) In edge-based subsampling, synthetic growth defect and lethality, and negative and positive genetic interacting gene pairs were withheld uniformly at random over three trials. (B) Considering the same genetic interactions as in (A), except that at each point, genes under a set interaction degree cutoff are removed.

3.5 Discussion

Our results demonstrate that various classes of genetic interactions in different organisms can be successfully predicted based on the hypothesis that functional gene clusters tend to share genetic interaction partners. For *S. cerevisiae* in particular, predictability was obtained whether the genetic interaction type was based on growth effects or non-growth phenotype-based measurements (i.e. phenotypic suppression). Interestingly, our method did not yield predictability for negative and positive genetic interactions, which happen to be the interaction types for which most of the pairs have been tested [15]. While the range of predictable genetic interaction classes for human and fly were limited to phenotypic enhancement and suppression, we believe that this is probably due to the sparsity of known genetic interactions for these organisms. In this study, the source of known genetic interactions, BIOGRID, had over 150000 yeast gene pairs but only ˜2800 pairs for fly and ˜1500 for human. As shown in Table 3.1, many types of genetic interactions could simply not be tested for fly and human.

This sparseness of experimentally-determined genetic interactions, especially in human, led to the lack of enrichment in gene modules such as protein complexes. In our simulations of withholding genetic interacting pairs, we expected that regardless of the interaction type, there would be a point after which no enrichment would be found. Thus, it was surprising that negative and positive genetic interactions exhibited a linear decrease in enrichment, regardless of how the pairs were withheld (by degree or at random). On the other hand, the enrichment signal in synthetic growth defect and lethality is sensitive to the interaction degree, as there was a steep drop-off when most of the interaction pairs were withheld. In the negative and positive genetic networks, there appears to sufficient genetic interaction density such that even when high numbers of interacting pairs are withheld, enrichment under a binomial model can still be found. By extrapolating to the human case, a modest increase in the number of screened human gene pairs is likely to dramatically increase the ability to predict additional genetic interactions, especially for synthetic growth defect and lethality where the genes have multiple interaction partners.

Similar to previous genetic interaction prediction approaches [62, 89], our algorithm requires knowledge of known experimentally determined genetic interactions. While other studies proceed without such requirements, the assimilation of a host of biologically annotated features are still necessary for their prediction method [55, 84]. In contrast to the aforementioned studies, our methodology systematically examined more than one class of genetic interaction and was successfully applied to multiple eukaryotic organisms, thereby generalizing results from a previous study by Lee et al. [46]. Since the detection of tightly connected sets of nodes in a network is central to our method, further avenues for exploration perhaps include investigating methods such as graph clustering [22] or community detection algorithms [25], though these algorithms lack built-in validation. It would also be interesting to explore using tissue-specific gene networks instead of a single integrated functional gene network for more targeted predictions [26].

As one major goal of any genetic interaction prediction is to at least narrow down the search space for experimentally testing genetically interacting pairs, our predictions are specifically testable experimentally, perhaps through CRISPR-Cas9 for human cells [83]. We also contribute to available prediction methodologies for suggesting genetic interactions as candidate therapeutic targets. Ultimately, we demonstrate the power of leveraging knowledge of known genetic interactions and integrated biological information in functional gene networks to predict novel genetic interactions from single-cell to multicellular organisms.

3.6 Acknowledgments

E.M.M. acknowledges funding from the National Institutes of Health, the National Science Foundation, the Cancer Prevention and Research Institute of Texas, and the Welch Foundation (F1515). The authors thank Kevin Drew for assistance with web server setup.

Chapter 4

Development of Drug Combination Regimens: Yeast as a Model System

4.1 A History of Yeast in Drug Discovery[1]

In the field of drug discovery, yeast provide a useful high-throughput platform both to select candidate drug compounds and to identify drug targets. In perhaps the simplest case, screening an overexpression or deletion yeast strain collection can identify strains that are overly sensitive or resistant to drug treatment. For example, screening a set of kinase-directed compounds against a yeast overexpression library revealed several compounds targeting the PKC1-MAPK1 pathway. One compound was found to directly target yeast Pkc1 [50]. In a complementary approach, Lum et al. [51] assayed a heterozygous yeast deletion mutant library looking for haploinsufficiency in response to a set of known therapeutics and successfully identified protein targets for several compounds.

Other approaches to drug discovery have been made possible by combining chemogenomic screens of the yeast deletion library [32] and high through-

[1] Adapted from Laurent JM, Young JH, Kachroo AH, Marcotte EM. Efforts to make and apply humanized yeast. *Briefings in Functional Genomics*, 15(2):155-163 (2016). Portions of the paper written by J.H.Y comprise this section.

put quantification of yeast genetic interactions [15]. Identifying mutants that are sensitive to a drug of interest relative to wild-type suggests that the drug's target may be a genetic interaction partner of the deleted gene. Bioinformatic querying of genetic interaction and chemical sensitivity databases has yielded both drug targets and off-target effects for multiple compounds, e.g. tamoxifen and benomyl [57, 58]. Experimental screens of chemical-genetic interactions have also been fruitful. For example, significant genetic interactions between yeast SOD1 and the DNA damage and checkpoint repair (DDCR) pathway guided the discovery of a small-molecule inhibitor of DDCR in yeast. Sod1δ strains showed sensitivity to several compounds in a screen of over 3000 small molecules. One compound allowed partial rescue of yeast growth inhibition in the presence of DNA-damaging agents, suggesting DDCR as the target for the compound, which was confirmed in human colorectal cancer cell lines [79].

Genetic interaction is also evident when overexpression of one gene inhibits growth in the deletion background of another gene. In some cancers, Mad2, a critical cell-cycle checkpoint control protein, is overexpressed and screening for genes whose deletion causes reduced growth in Mad2-overexpressing yeast identified candidate target genes [8]. Thirteen of the identified yeast genes had human orthologs, and knockdown of one of these (PPP2R1A) caused lethality in human cells (HeLa) that had MAD2 overexpressed. Interestingly, PPP2R1A is a regulatory subunit of protein phosphatase 2 (Ppa2), the target of cantharidin, which was found to inhibit the MAD2-overexpressing osteosarcoma cell line OS-17.

Finally, alternative high-throughput techniques for drug target identification that do not involve genetic interaction screening have also been developed. The molecularly barcoded yeast open reading frame (MoBY-ORF) collection comprises a library of ˜5000 yeast genes cloned in expression plasmids flanked by upstream and downstream barcodes that enable plasmid identification in pooled growth assays, greatly lowering the amount of drug necessary. In the case of a drug-resistant mutant, the MoBY-ORF collection is transformed into mutant strains and assayed for renewed sensitivity to the drug, followed by amplification and identification of the responsible gene via microarray [33]. The method was validated by identifying targets of rapamycin and several antifungals.

In previous chapters, we established methods for identifying candidate therapeutic targets and a mechanism, namely genetic interactions, by which inhibiting those targets sensitize the desired population. With yeast serving as our model system, we now examine a strategy to integrate genetic interactions as targets to design drug combination regimens.

4.2 Predicting Synergistic and Antagonistic Drug Pairs in Yeast[2]

4.2.1 Abstract

Although drug combinations have proven efficacious in a variety of diseases, the design of such regimens often involves extensive experimental screening due to the myriad choice of drugs and doses. To address these challenges, we utilize the budding yeast *Saccharomyces cerevisiae* as a model organism to evaluate whether drug synergy or antagonism is mediated through genetic interactions between their target genes. Specifically, we hypothesize that if the inhibition targets of one chemical compound are in close proximity to those of a second compound in a genetic interaction network, then the compound pair will exhibit synergy or antagonism. Graph metrics are employed to make precise the notion of proximity in a network. Knowledge of genetic interactions and small-molecule targets are compiled through literature sources and curated databases, with predictions validated according to experimentally determined gold standards. Finally, we test whether genetic interactions propagate through networks according to a "guilt-by-association" framework. Our results suggest that close proximity between the target genes of one drug and those of another drug does not strongly predict synergy or antagonism. In addition, we find that the extent to which the growth of a double gene mutant deviates from expectation is moderately anti-correlated with their distance in

[2] Published as Young JH and Marcotte EM. Predicting Synergistic and Antagonistic Drug Pairs in Yeast. *bioRxiv*, (2016): 050567. J.H.Y. designed the study, performed data analysis, and wrote the paper.

a genetic interaction network.

4.2.2 Introduction

Drug combinations have an established history in treating disease, dating to the MOPP regimen for Hodgkin's lymphoma in the 1960s to highly active antiretroviral therapy (HAART) for HIV in the 1990s [29, 20]. In combating antibiotic resistance, combination regimens have proven effective and are actively under continued development [85]. Yet in designing combination therapies, it is not immediately clear which drugs and doses to group together; there are simply a myriad of possible choices and the combinatorial space quickly grows unwieldy. As a result, any computational technique to either guide readily testable candidates or reliably predict the effect of drug combinations would be desirable. In this study, using the budding yeast *Saccharomyces cerevisiae* as a test platform, we determine whether the effect of drug pairs can be predicted from genetic interactions between their target genes.

The effect of a drug combination can be classified as synergistic, antagonistic, or additive. Two drugs are synergistic if they cause a significantly greater growth defect than expected, based on the effect of each drug individually. Antagonism is similar, although the effect is far more pronounced growth than expected. Drug additivity implies that no interaction exists between the agents, and the resulting phenotype is the sum of each drug's individual effect. There is more than one choice of a null model that defines the "expected effect" - commonly used models include Loewe additivity and Bliss independence [49, 9, 87].

Previous studies to uncover genetic interactions as a mechanism underlying drug combinations have involved exhaustive screening of a number of small-molecule chemical compounds. An examination of 200 compound pairs administered in *Saccharomyces cerevisiae* found 38 of them to be synergistic, but genetic interactions were determined to be responsible for only 14 of those 38 [14]. Another study screened all possible pairs of 128 compounds from a chemically diverse library to experimentally deduce synergy and antagonism, thereby establishing a validation set. Moreover, a model based directly on chemical-genetic and genetic interactions had low predictive power for synergy or antagonism, but combining naive Bayes and random forests trained on additional features led to successful predictions [82].

We hypothesized that the proximity in a genetic interaction network between one drug's target genes and another drug's targets controls the degree to which the drug pair is synergistic or antagonistic. In particular, rather than considering only direct interactions between genes, our approach factored in whether a gene is within a neighborhood of (though not necessarily adjacent to) some other gene in the network. We leveraged knowledge of known small-molecule inhibition targets in *S. cerevisiae* from the Search Tool for Interactions of Chemicals (STITCH) database [43] and experimentally determined negative and positive genetic interactions [15]. Finally, predictions of synergy or antagonism were validated against gold standards assembled from the literature.

4.2.3 Methods

4.2.3.1 Negative and positive genetic interaction network

Negative and positive genetic interactions were compiled from a high-throughput yeast synthetic genetic array (SGA) screening dataset [15]. The intermediate cutoff for the genetic interaction score ϵ was chosen as the threshold for interacting versus non-interacting gene pairs. For the purposes of data processing, the suffixes "_tsq" and "_damp" were removed from gene symbols. Both unweighted and weighted versions of each of the negative and positive genetic interaction networks were assembled. Nodes in the networks correspond to genes and two genes are connected by an edge if they interact according to the intermediate cutoff ϵ. Because a larger magnitude of ϵ indicated stronger genetic interaction, in the weighted networks the edge weights were assigned by reversing the ϵ values. For instance, the strongest genetically interacting pair was assigned an edge weight with the smallest ϵ instead. All edge weights were set to be non-negative.

4.2.3.2 Chemical compound targets and gold standards for synergy and antagonism

Two literature sources were used as the gold standard to validate chemical synergy predictions [14, 82]. The inhibition targets in *S. cerevisiae* of chemical compounds (identified by CID) were assembled from STITCH version 4 [43]. Only chemical names were available from the Cokol et al. dataset;

these were converted to CIDs with PubChemPy `https://pypi.python.org/pypi/PubChemPy`. SMILES strings from the Wildenhain et al. dataset were also converted to CIDs. Prediction performance was assessed with receiver operating characteristic (ROC) analysis as implemented in scikit-learn [59].

4.2.3.3 Distances in networks

Distances between all pairs of nodes in unweighted and weighted versions of both the negative and positive genetic network were computed using Dijkstra's algorithm as implemented in NetworkX [66]. The distance between two sets A and B of nodes in an unweighted network was calculated using the earth mover's metric (EMD) [63]. Here in the 1-dimensional special case, the EMD reduces to differences between cumulative distribution functions [13]. For the purpose of measuring the distance between two sets of nodes, $\text{EMD}(A, B) = \sum_{i \in \mathbb{N}_0} |F_{X_{\text{ref}}}(i) - F_X(i)|$, where $F_{X_{\text{ref}}}$ and F_X are the cumulative distribution functions (CDFs) of a reference distribution X_{ref} and a random variable X. The reference distribution is intended to represent the scenario where every node in A is adjacent to some node of B. In an unweighted network, the reference probability mass function (pmf) of $F_{X_{\text{ref}}}$ is defined as

$$P(X_{\text{ref}} = k) \coloneqq \begin{cases} 1 & \text{if } k = 1 \\ 0 & \text{if } k \neq 1, k \in \mathbb{N} \end{cases}$$

and the pmf for X is constructed from the frequencies of all possible node pair distances between A and B as found from Dijkstra's algorithm.

In a weighted network, we have

$$
\begin{aligned}
\mathrm{EMD}(A,B) &= \int_0^{+\infty} |F_{X_{\mathrm{ref}}}(t) - F_X(t)|\, dt \\
&= \sum_{i=1} (x_i - x_{i-1})(F_{X_{\mathrm{ref}}}(x_{i-1}) - F_X(x_{i-1})) \\
&= \sum_{i=1} (x_i - x_{i-1})(1 - F_X(x_{i-1})) \\
&= (x_1 - x_0) + \sum_{i=2} (x_i - x_{i-1})(1 - F_X(x_{i-1}))
\end{aligned}
$$

where by choosing x_0 to be the minimum edge weight $P(X_{\mathrm{ref}} = x_0) \coloneqq 1$ and 0 elsewhere, and $P(X)$ is non-zero only for the node pair distances $x_1, x_2, \ldots$ with $x_0 \leq x_1 < x_2 < \cdots$.

4.2.3.4 Software Availability

Computational analyses were performed with Python version 3.4; scripts and Jupyter notebooks are available under the BSD license at

`https://bitbucket.org/youngjh/yeast_chem_synergy`.

All plots were created with Matplotlib and Seaborn [37].

4.2.4 Results

4.2.4.1 Close proximity between drug target genes in the genetic interaction network does not strongly predict synergy or antagonism

We hypothesized that if two chemical compounds are synergistic, then the inhibition target genes of one compound would be close to those of the second compound in a negative genetic interaction network. Similarly, antagonistic compound pairs would have their respective targets near one another in a positive genetic network. Proximity between target genes were assessed in both unweighted and weighted genetic interaction networks. An experimental screen in *S. cerevisiae* provided the gold standard benchmark for testing the synergy hypothesis [82]. In this dataset, all possible pairs of 128 chemical compounds were screened, but only 7 compounds had inhibition target genes found in both the negative genetic network and the Search Tool for Interactions of Chemicals (STITCH) database. Thus, there were 21 possible pairs available for validation, three of which exhibited synergy from the screening results. None of the antagonistic compounds in this dataset contained targets listed in STITCH. As shown in Table 4.1, close proximity of the target genes were only weakly predictive of synergy, according to the area under the curve (AUC) from the receiver operating characteristic (ROC) analysis. The AUC from the unweighted network was reasonably consistent with that from the weighted network.

For the antagonism case, the gold standard was constructed from another experimental screen [14]. Of the 200 pairs screened from 33 compounds, only 10 pairs had compounds whose inhibition targets were both listed in STITCH and in the positive genetic network. Of these 10, 8 were experimentally determined to show antagonism. None of the synergistic compounds in this dataset contained targets listed in STITCH. No evidence was found to suggest that close proximity of target genes was predictive of antagonism (Ta-

	Synergy	Antagonism
Unweighted network	0.61	0.41
Weighted network	0.57	0.19

Synergy and antagonism prediction performance assessed by AUC, the area under the receiver operating characteristic (ROC) curve.

Table 4.1: Chemical compound pairs were scored and ranked for synergy or antagonism by the distance between their inhibition targets in a genetic interaction network. The predictions were validated through receiver operating characteristic (ROC) analysis with true interactions labeled according to gold standards for synergy and antagonism. In the synergy case, target gene proximity is only marginally more predictive than random for chemical synergy or antagonism.

ble 4.1). In fact, the results suggest that the farther apart one set of target inhibition genes is from those of a second compound in the positive genetic network, the more likely the compound pairs are to be antagonistic. Strikingly, in contrast to the synergy case above, the AUC value from the unweighted network was quite far apart from that of the weighted network.

4.2.4.2 Genetic interaction strength is moderately correlated with network distance

One assumption underlying our hypothesis was that any two genes that were not adjacent in the genetic interaction network but within a sufficiently small neighborhood of one another would still express some degree of interaction. Conversely, if the genes were located very far apart, they would essentially not interact at all. To examine the validity of this assumption, we sought to determine the correlation, if any, between a gene pair's distance in the network and its corresponding strength of genetic interaction. The interaction strength was simply the magnitude of the genetic interaction score $|\epsilon|$ from the raw results of the synthetic genetic array (SGA) screening [15]. The network distance of a gene pair was once again the distance computed from Dijkstra's algorithm as described above, such that smaller distances implied stronger interaction and consequently larger $|\epsilon|$. Therefore, we expected to observe negative correlations for both negative and positive genetic. As shown in Figure 4.1, indeed the Spearman's rank correlation correlation is in fact moderately negative and statistically significant.

4.2.5 Discussion

Our results suggest that there is no evidence to support the claim that synergy or antagonism arises when the target genes of one chemical compound are close in a genetic interaction network to those of another compound. We confirmed previous results that such drug interactions are not directly mediated through genetic interactions [14, 82], and also showed that neighborhoods of genetic interactions are neither a contributing factor as well. In the process, we presented an application of distance measures satisfying the mathematical definition of a metric to quantify proximity between sets of nodes in gene networks. Prediction performance was measured through AUC due to its robustness to unbalanced data in positive versus negative class labels [38].

It is particularly notable that the gold standard for synergy produced results different than those from the antagonism gold standard. One potential

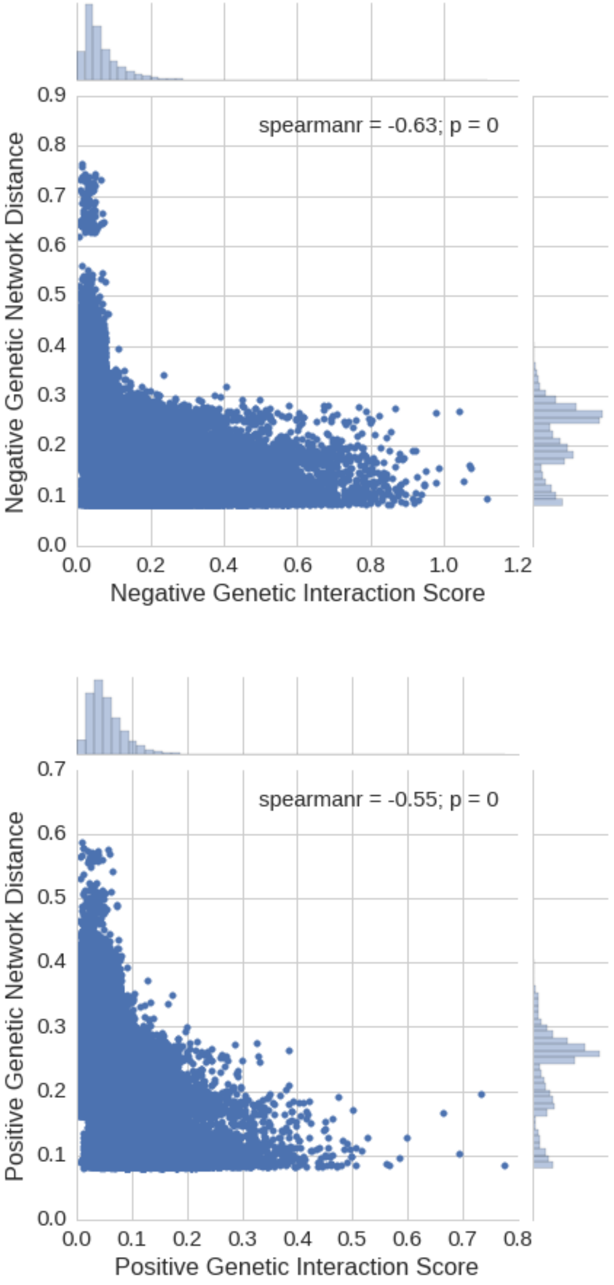


Figure 4.1: The magnitude of the genetic interaction score ϵ is moderately anti-correlated with gene network distance. Thus, the greater the growth deviation from expectation of a double mutant, the closer the two genes are in the genetic interaction network.

contributing factor is that the benchmark derived from Cokol et al. used the Loewe additivity model [49, 78] to determine synergy and antagonism, while Wildenhain et al. instead utilized Bliss independence [9]. The Bliss theory is closer to the multiplicative fitness model employed in calling negative and positive genetic interactions, which was defined as $\epsilon_{ij} = f_{ij} - f_i f_j$ with f_{ij} equal to the double mutant fitness and f_i, f_j as the single mutant fitness scores [15].

The moderate correlation between genetic interaction strength and network distance goes some way towards supporting the results from the synergy gold standard, where AUCs of 0.57 and 0.61 were attained. In any case, the weak correlation implies that genetic interactions cannot be reliably identified through "guilt-by-assocation" in the network. It should be noted that both datasets used to benchmark prediction performance were highly imbalanced, thus reflecting the need for even more data on which chemical compounds are synergistic or antagonistic, and which genes are inhibited by the compounds of interest. Yet despite the relatively limited data available to construct gold standards, our results and those of others indicate that a more nuanced mechanism beyond genetic interactions of target genes is responsible for explaining effects of chemical compound interactions.

Chapter 5

Conclusion

At the outset, we sought to develop computational methodologies to address barriers in each of the areas of target identification, mechanism, and therapy design in drug discovery and development. In choosing lung cancer as an application, it was important that inhibiting identified targets would sensitize some, but not all, of the cancer cell lines in our sample. As a result, specificity and relative non-toxicity to normal cells would be expected, given the genetic heterogeneity of lung cancer. One barrier to overcome was the combinatorially large space from which sets of functionally related genes were to be chosen as pathway-level vulnerabilities. A method based on k-means clustering successfully identified candidate targets with functions ranging from splicing, translation, protein folding and ribosome biogenesis. In fact, certain agents targeting some of the targets have been through clinical trials. Biclustering was also evaluated for candidate target selection and implicated many of the same biological functions prioritized by the k-means-based approach. We are able to obtain experimental validation for one of the highly ranked predictions from biclustering.

In our second aim, we concentrated on genetic interaction as one mechanism underlying disease vulnerability to gene perturbation. In order for such a mechanism to be generally applicable, knowledge of which genes interact without resorting to exhaustive experimental testing would be especially valuable. As such, we successfully applied a computational algorithm to produce high-confidence predictions for various classes of genetic interactions in multiple organisms. High recall was obtained when the predictions were cross-validated on known interactions curated from databases. Our results supported the hypothesis that functional network gene clusters - which are genes that participate in similar molecular or cellular functions - tend to share genetic interaction partners. The algorithm described proved to be valid for genetic interactions observed as growth defects and other phenotypic effects, although negative and positive genetic interactions in yeast proved to be an exception. We concluded with simulations of yeast genetic interaction networks to infer the manner in which homologous genetic interactions populate human biological systems such as protein complexes.

Finally, we focused on yeast as a model organism to integrate the themes of target identification and genetic interaction for design of therapeutic regimens. We began by reviewing the history of fungal species, particularly *Saccharomyces cerevisiae*, in biological assays relevant to human disease and contributing to drug discovery and development. Subsequently, a model for computationally predicting synergistic and antagonistic drug pairs in yeast was evaluated. The model was based on the hypothesis that close proximity in a genetic interaction network between the target genes of one drug and those of another drug would lead to synergy or antagonism. Graph metrics were chosen to quantify proximity in networks. As no evidence was found to support our hypothesis, we deduced that genetic interactions did not propagate through networks through "guilt-by-association." In total, we have applied and developed computational tools to advance various steps in the targeting, mechanism and therapy of the drug discovery pipeline.

Appendices

Appendix A

NSCLC Genetic Vulnerabilities

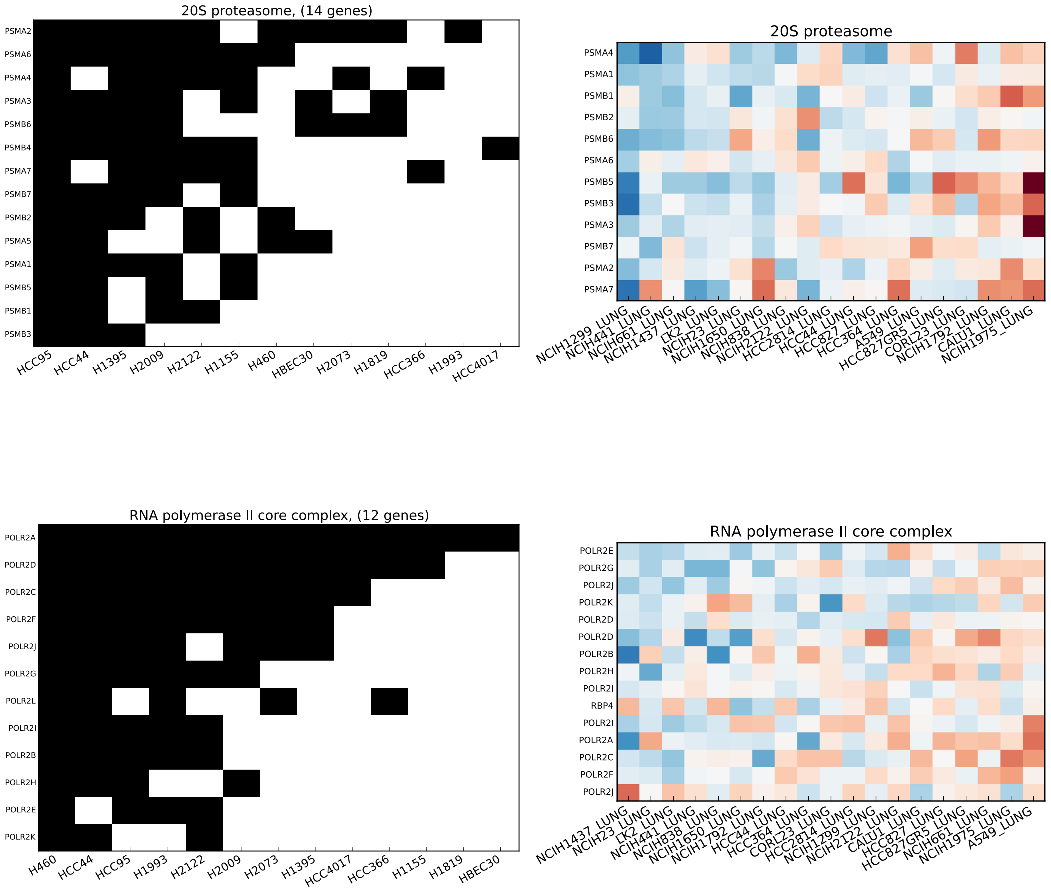


Figure A.1, continued on next page

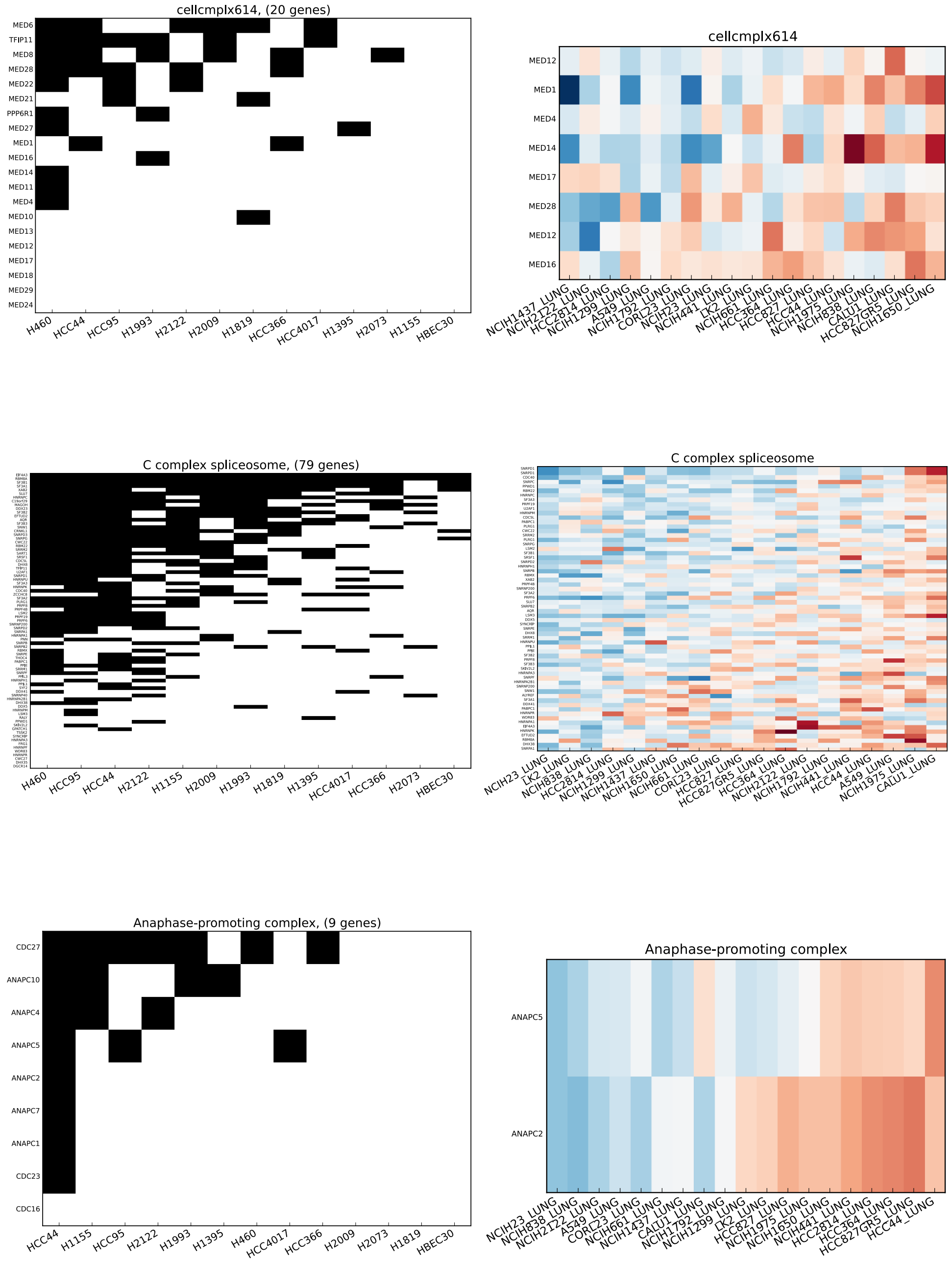


Figure A.1, continued on next page

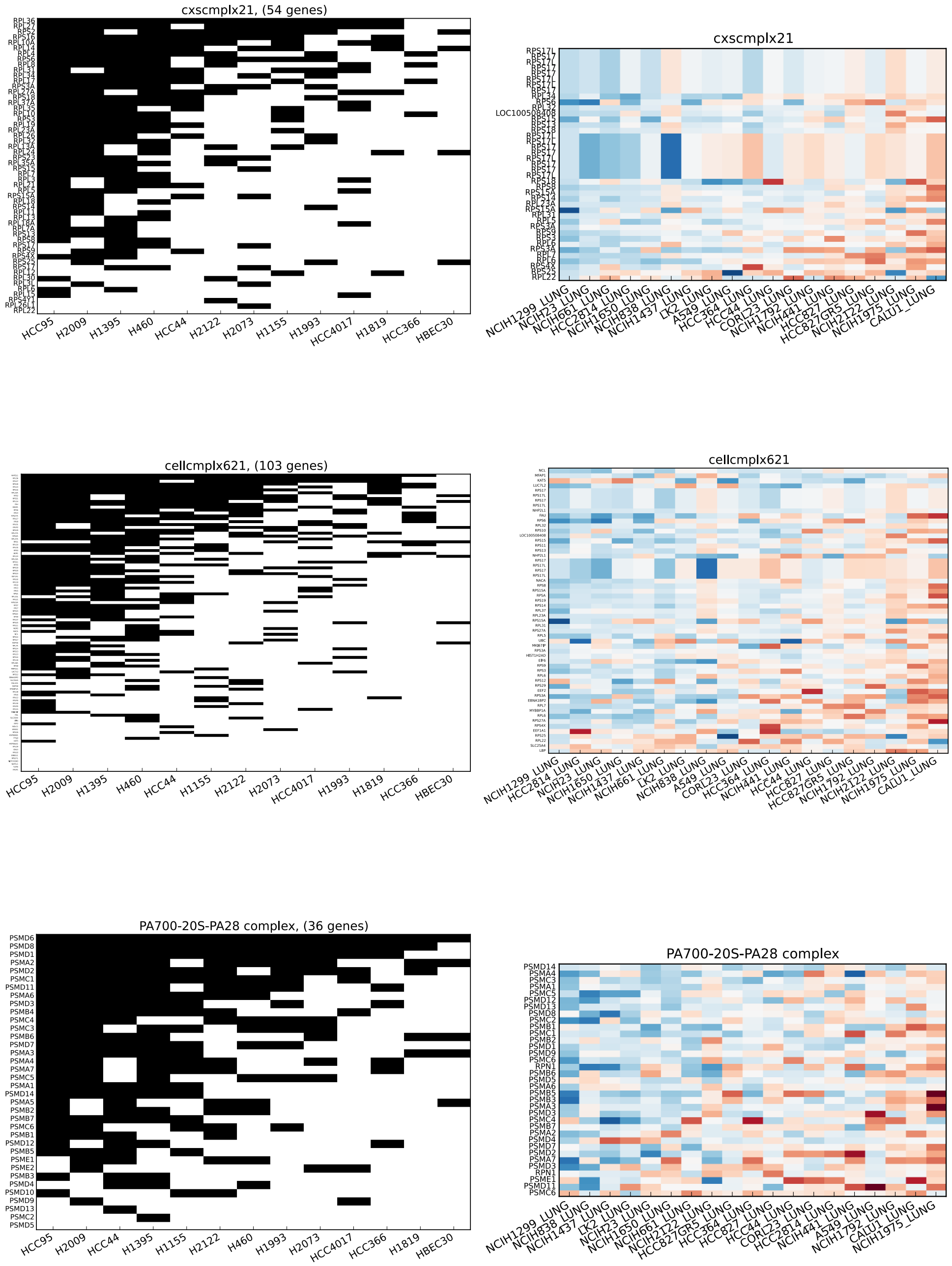


Figure A.1, continued on next page

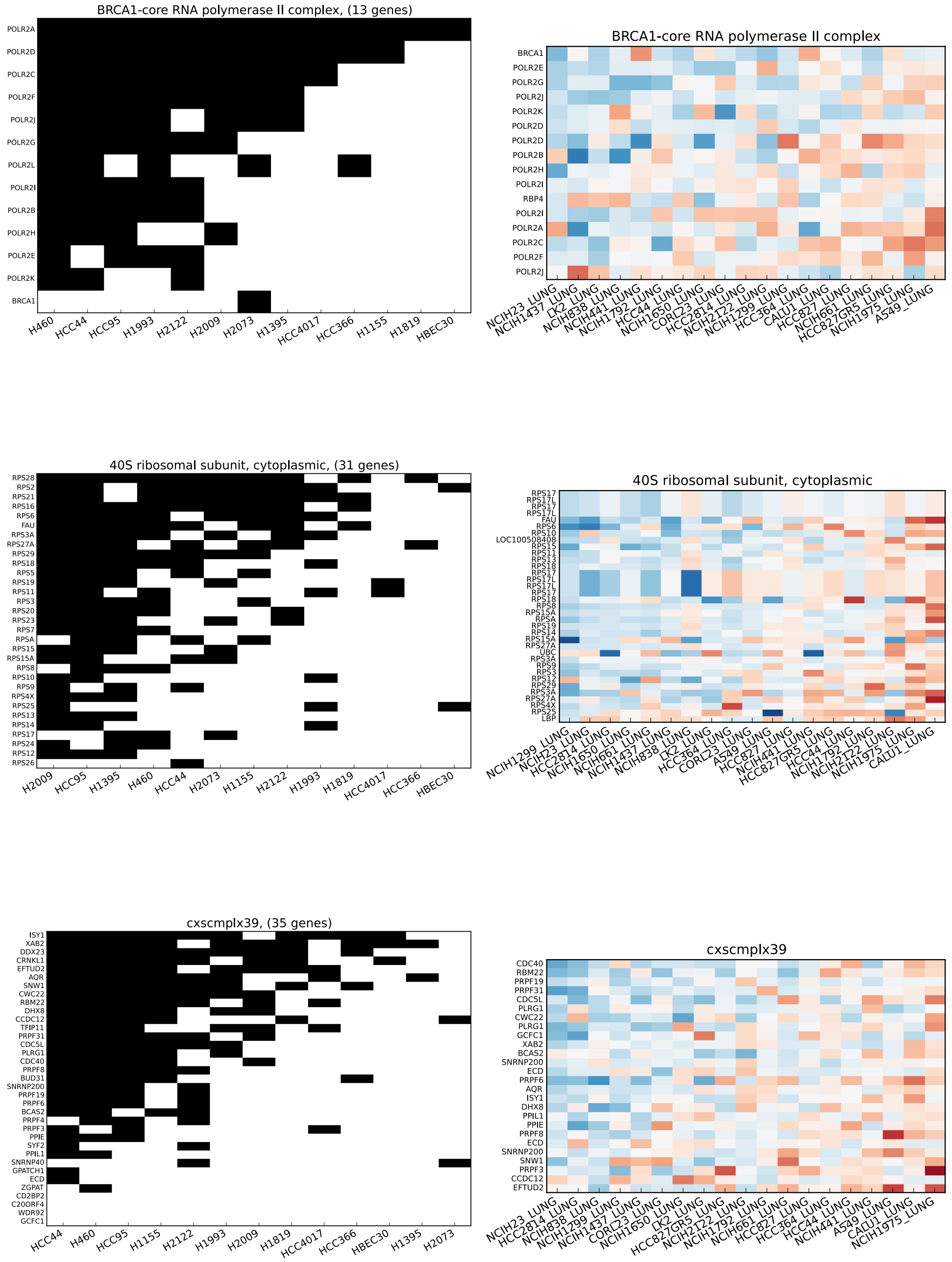


Figure A.1, continued on next page

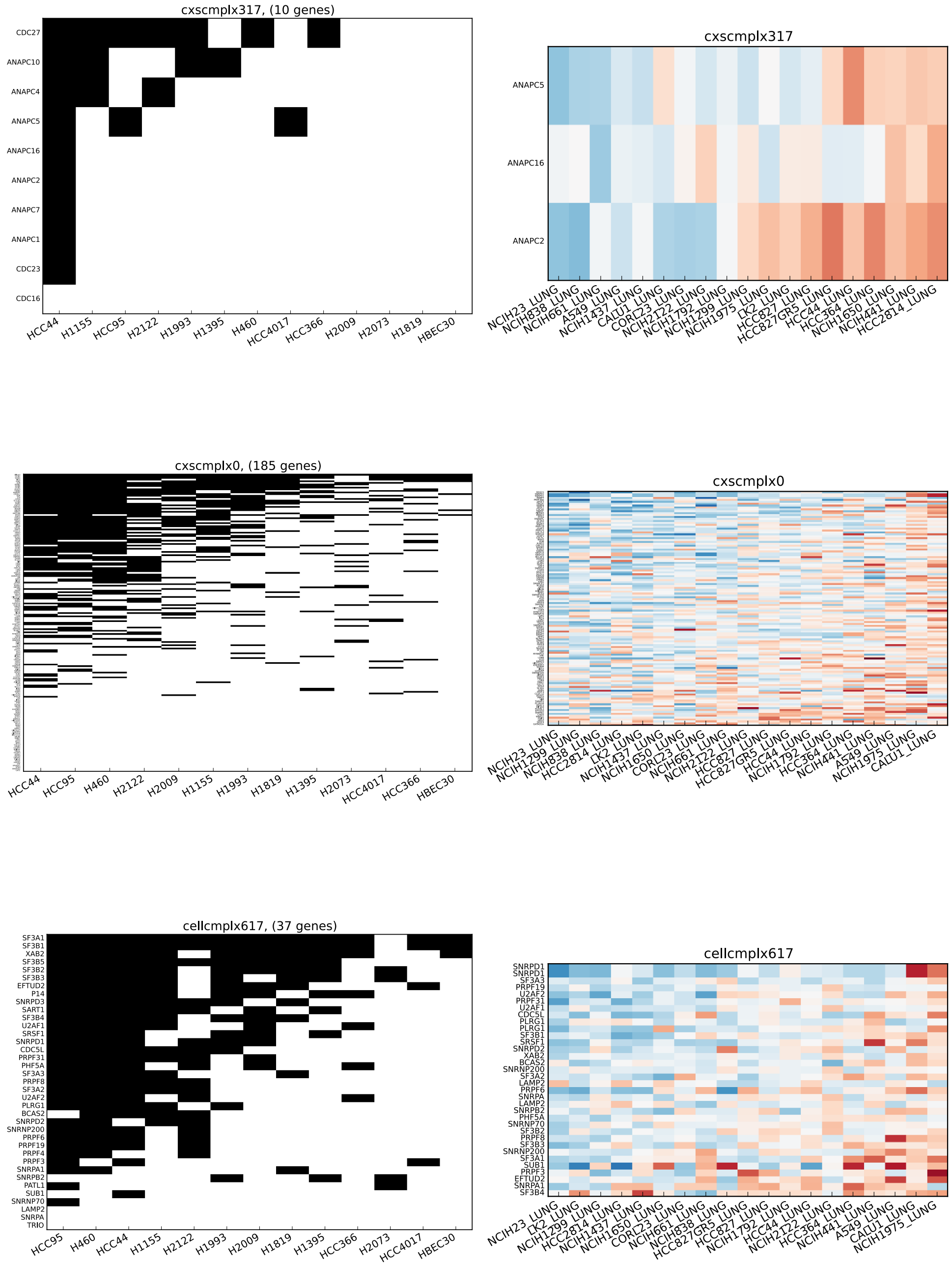


Figure A.1, continued on next page

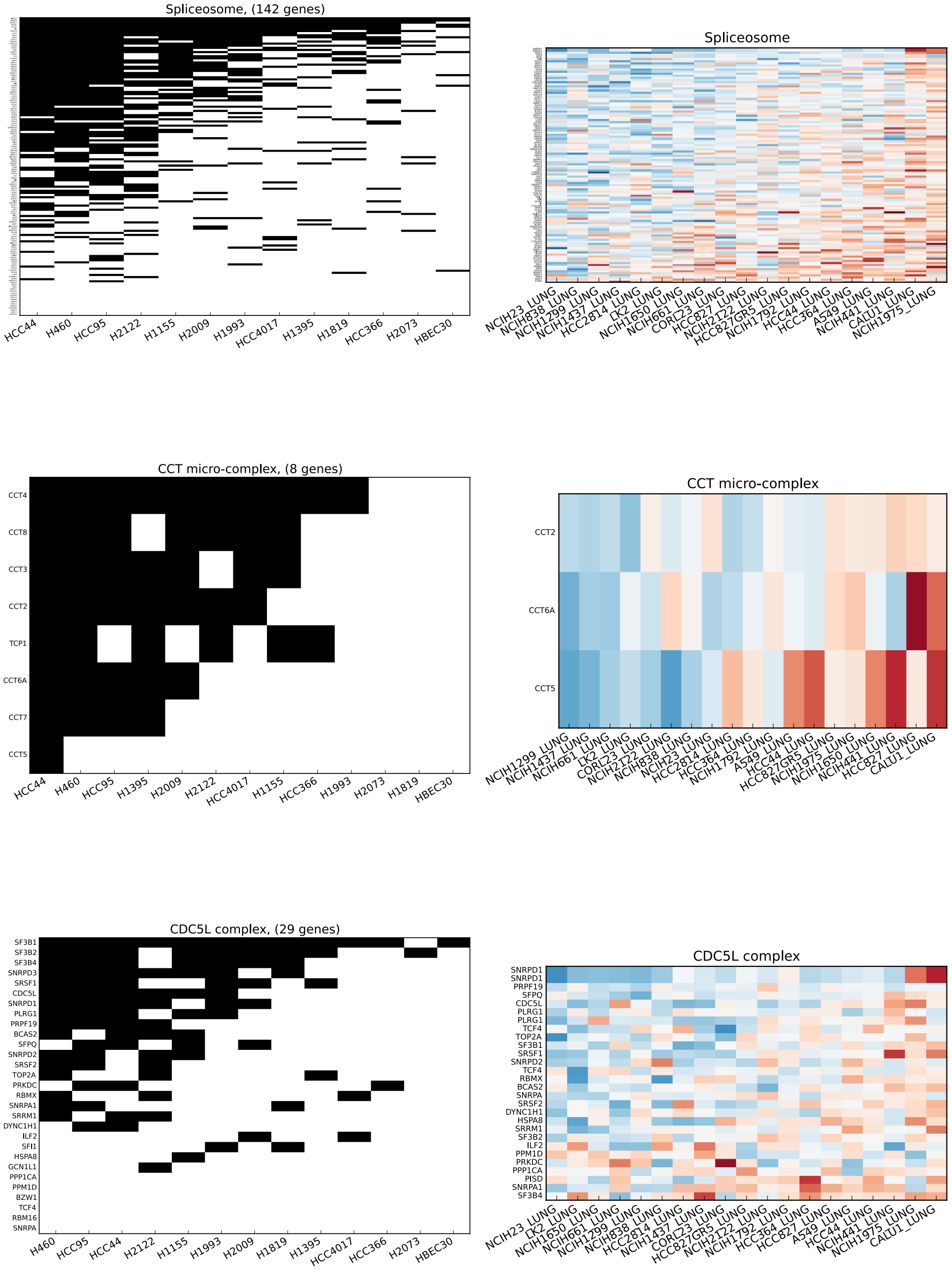


Figure A.1, continued on next page

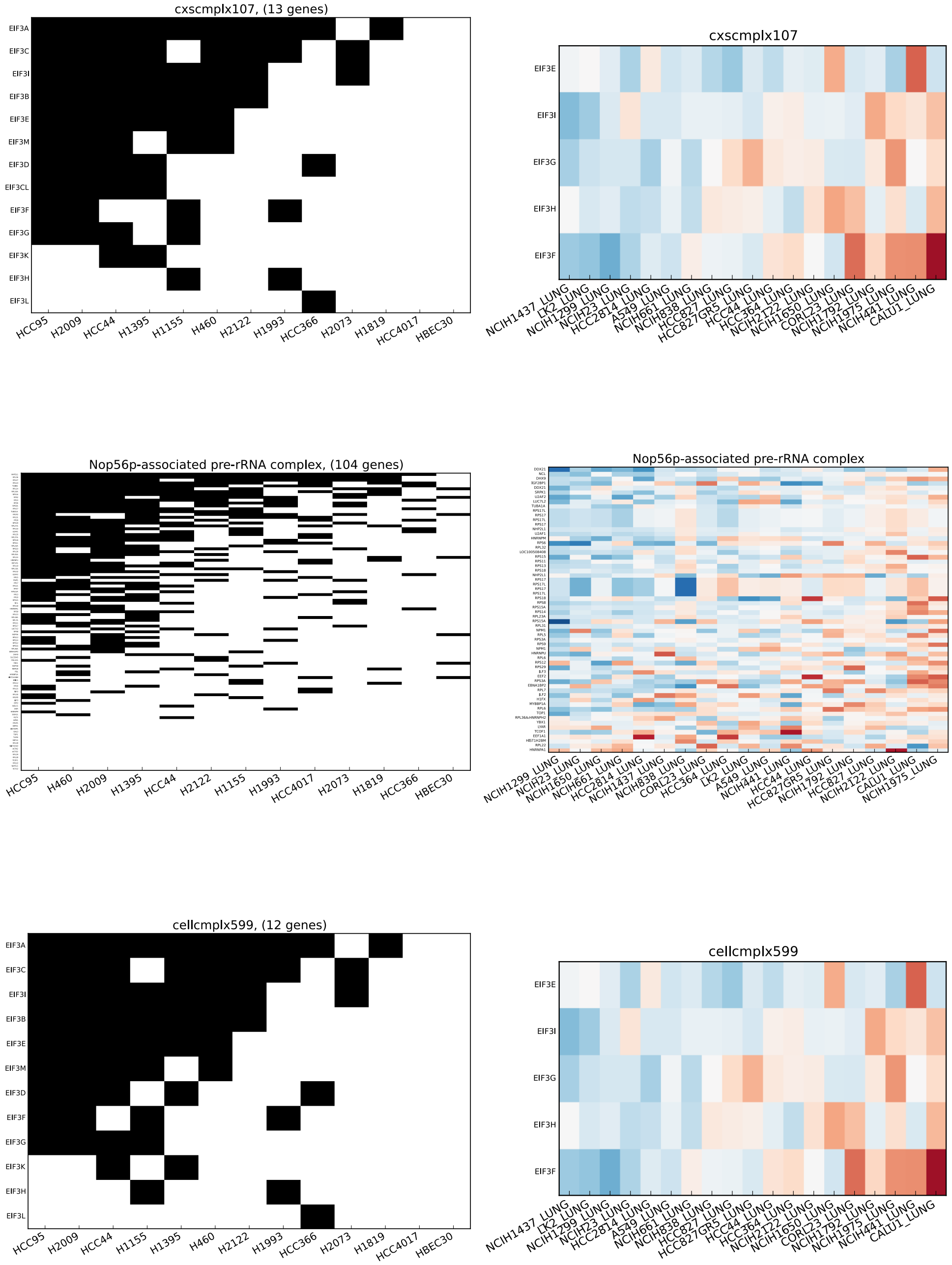


Figure A.1, continued on next page

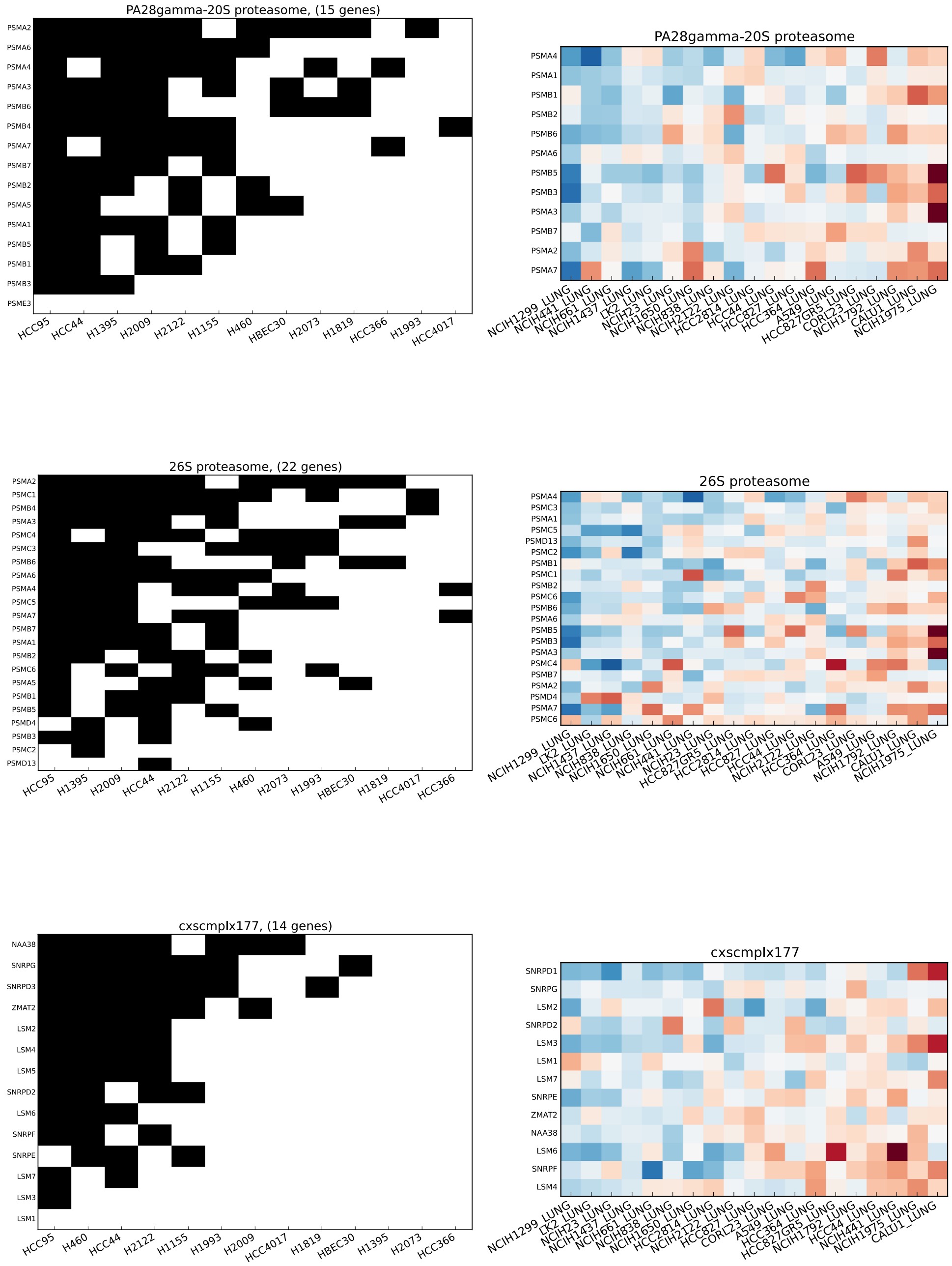


Figure A.1, continued on next page

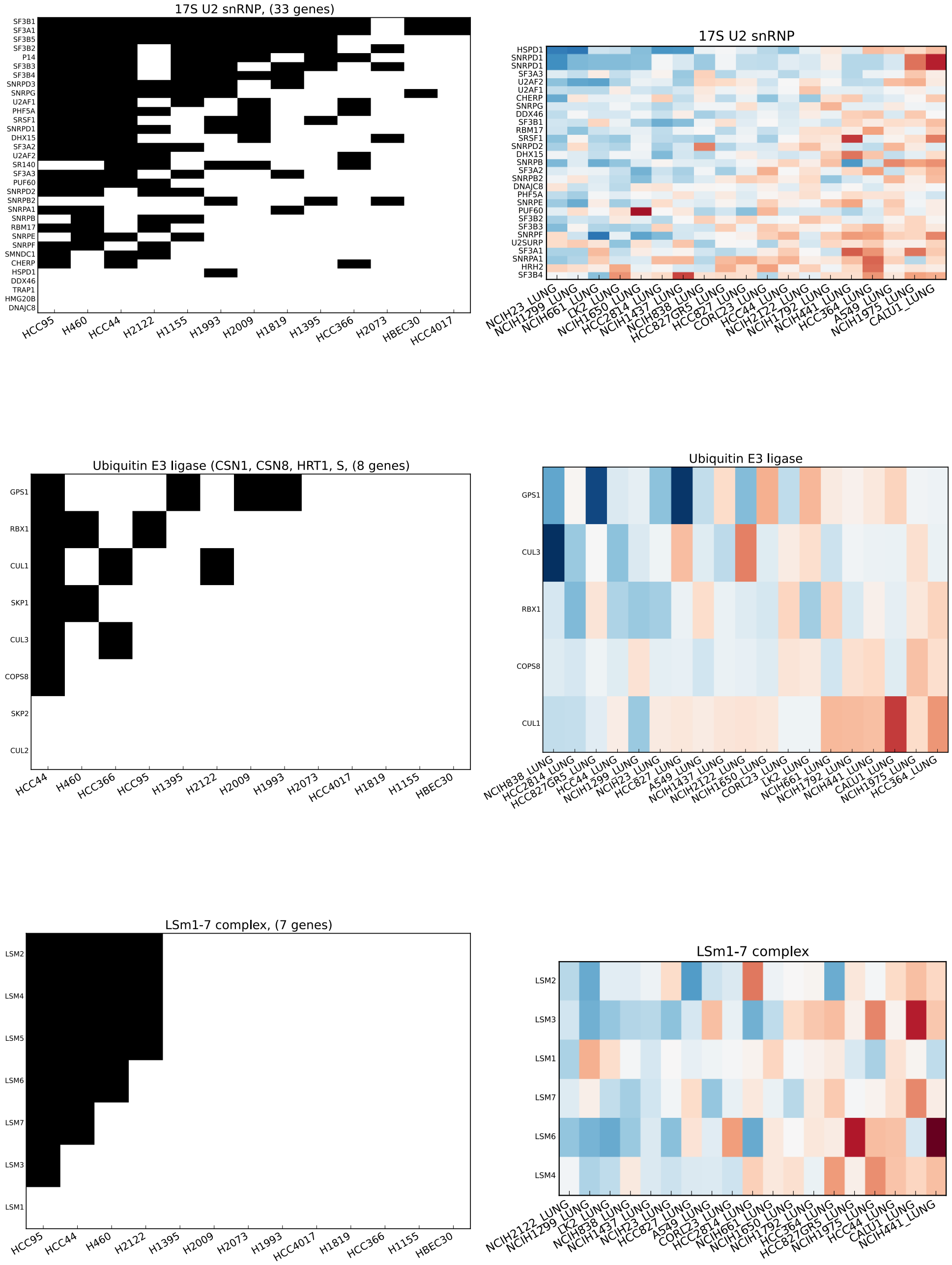


Figure A.1, continued on next page

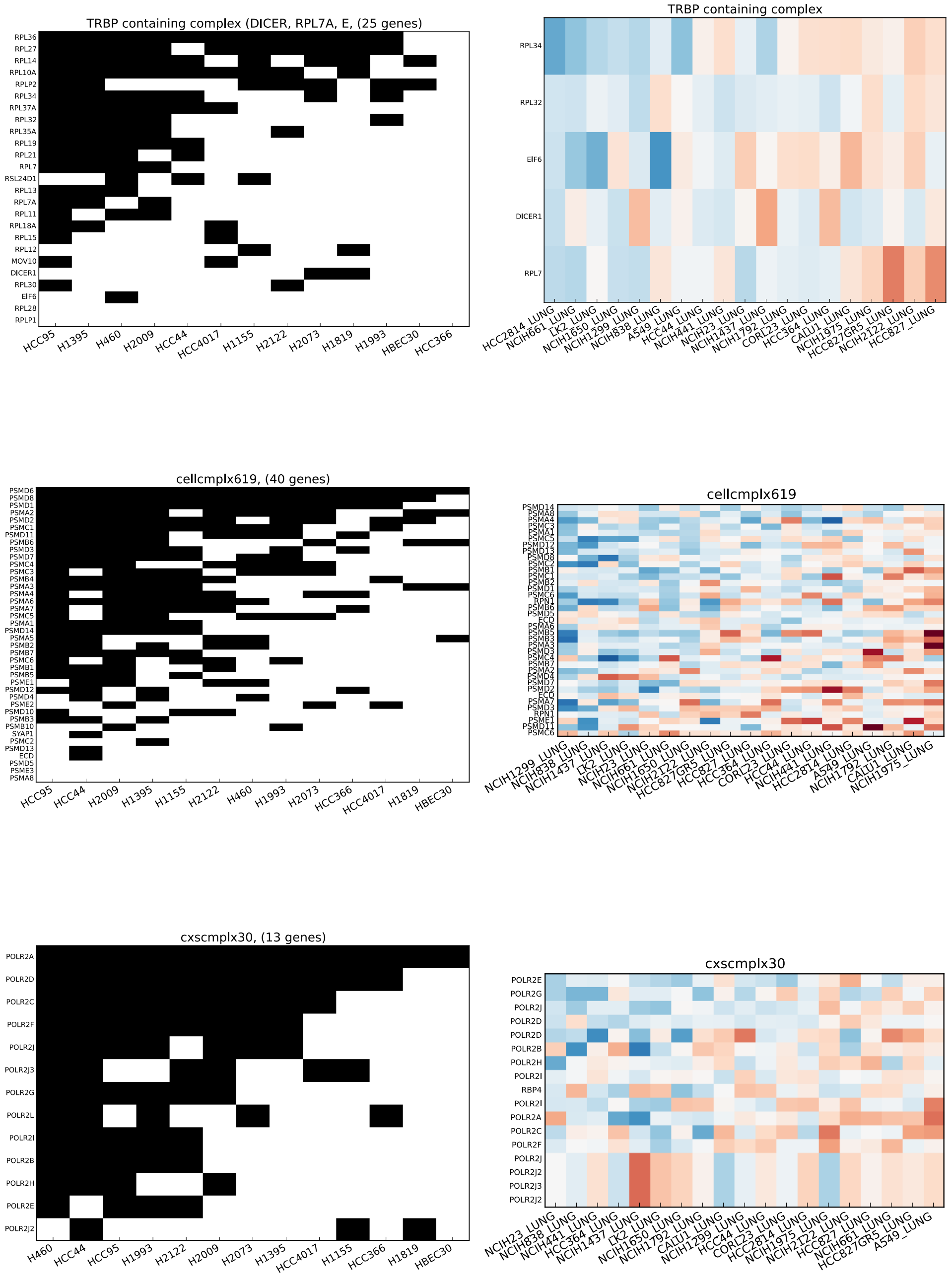


Figure A.1, continued on next page

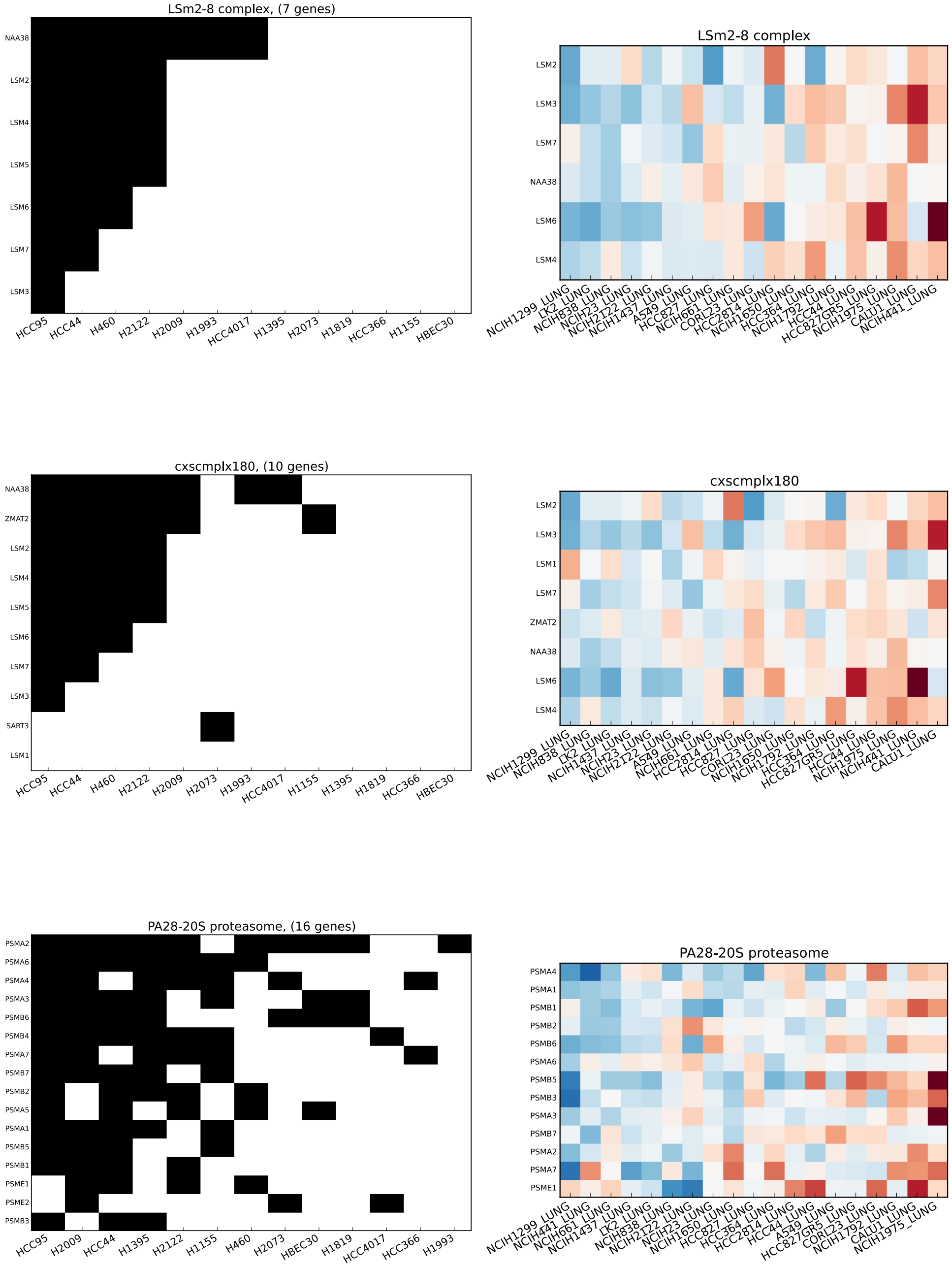


Figure A.1, continued on next page

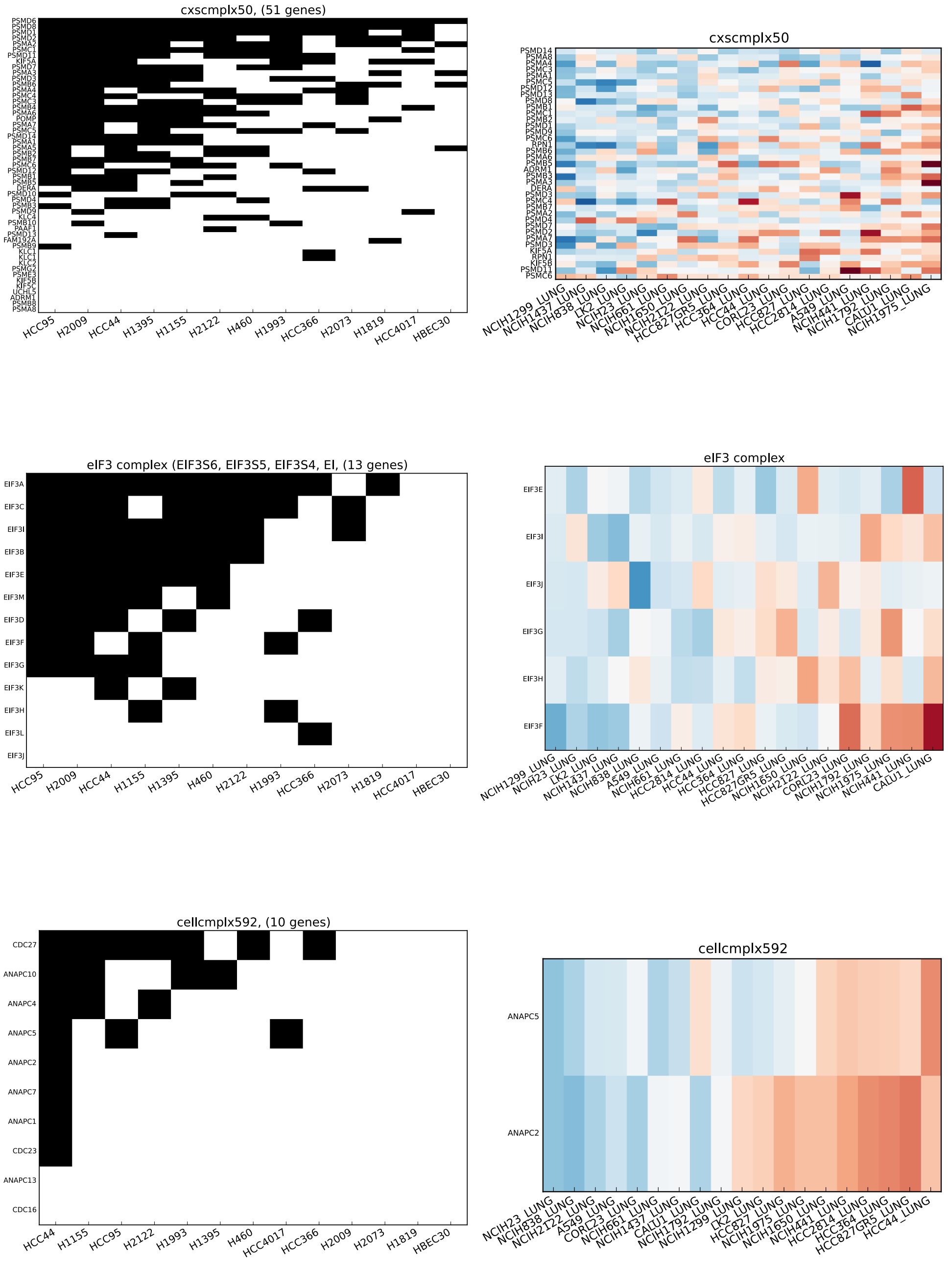


Figure A.1, continued on next page

Figure A.1: Genetic vulnerability patterns of protein complexes. Shown are all 35 protein complexes found from the 2-means clustering approach to be statistically significant at 10% FDR. Also shown for each complex is its corresponding shRNA knockdown sensitivity profile from Project Achilles. Note that some genes in certain protein complexes are absent in the Project Achilles heatmaps, which had knockdown data available for only 5711 genes. There are also some genes with multiple knockdown values for each cell line in Project Achilles. Complex labels indicate source (cellcmplx, Havugimana et al., 2012; cxscmplx, `http://metazoa.med.utoronto.ca`).

Appendix B

Genetic Interactions Predicted from YeastNet

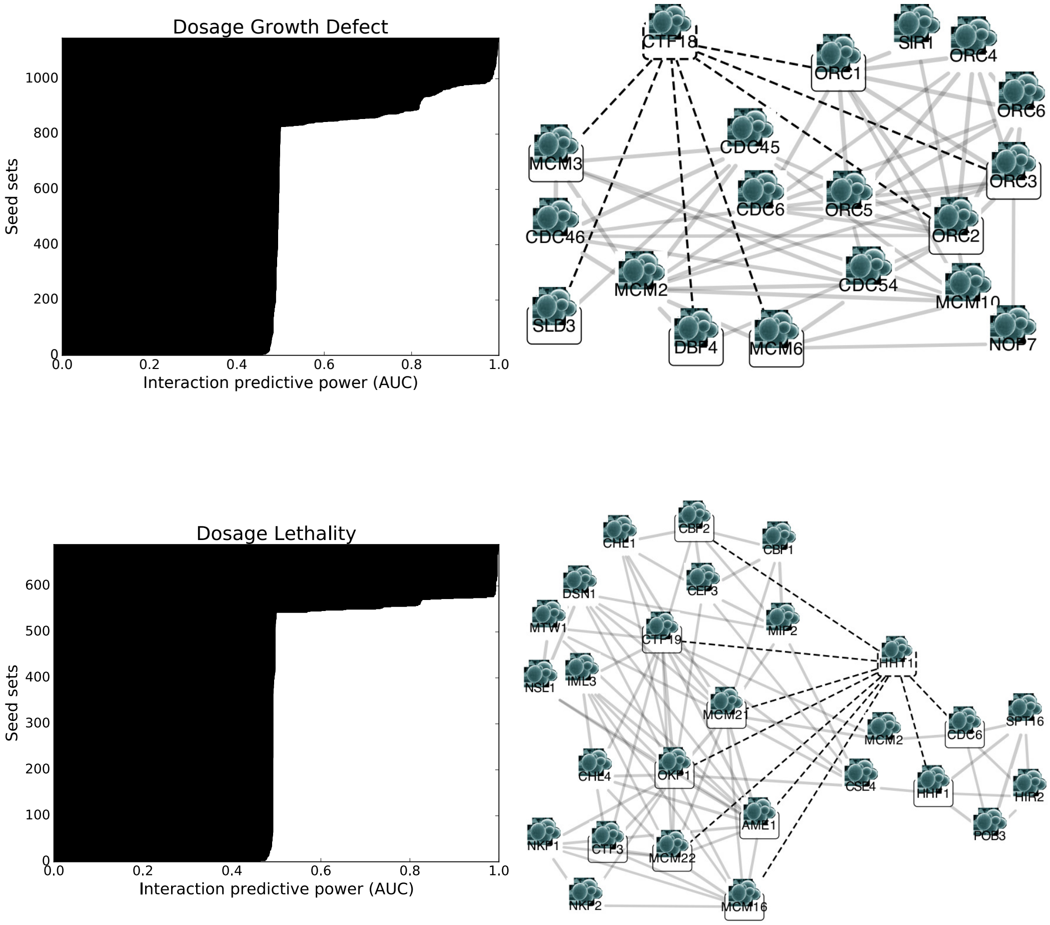


Figure B.1, continued on next page

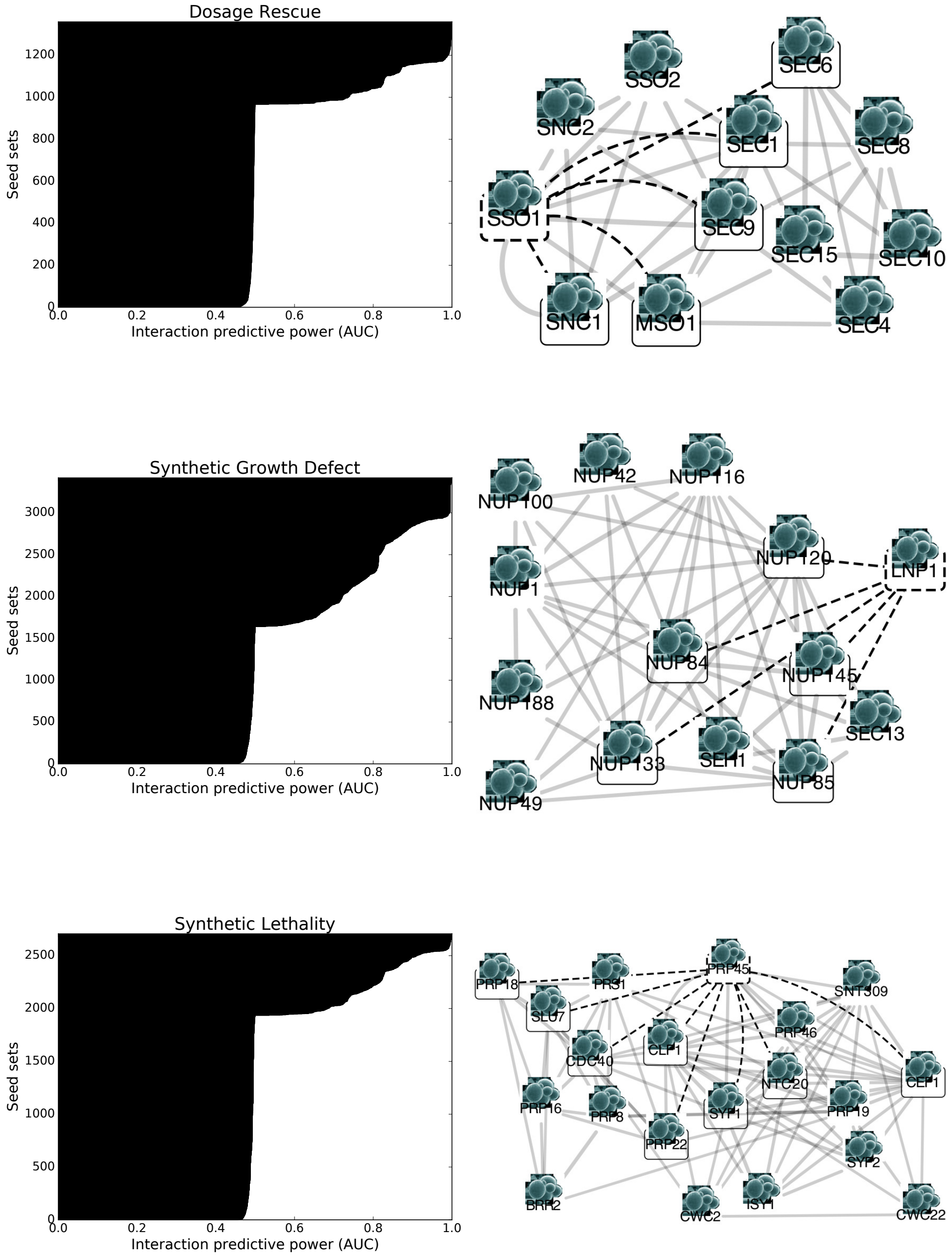


Figure B.1, continued on next page

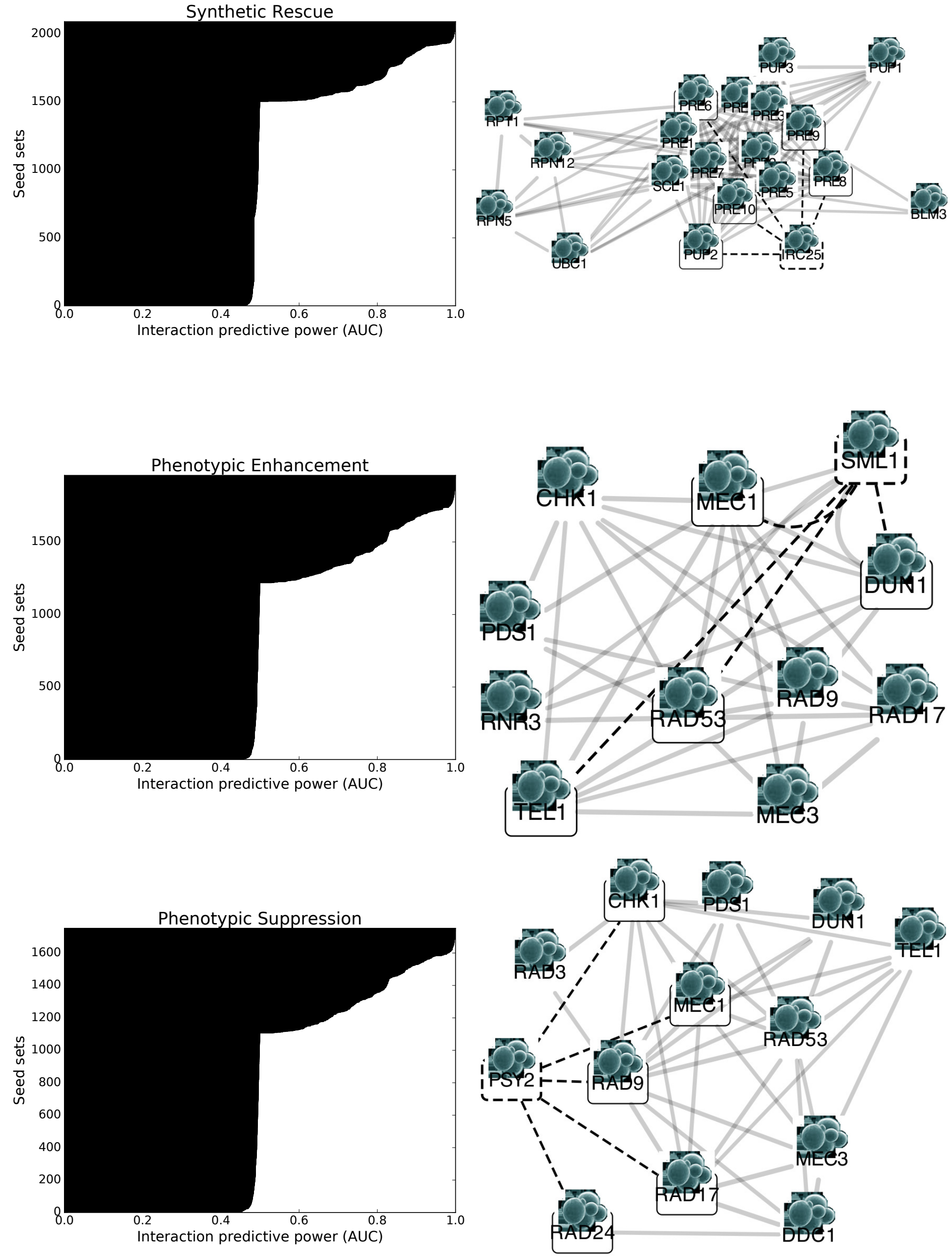


Figure B.1, continued on next page

Figure B.1: Yeast functional networks are predictive for diverse genetic interactions. The horizontal bar plots on the left hand side show the number of predictive functional net clusters for each genetic interaction. Each bar is a seed set, which is a group of genes all known to share genetic interactions with a particular gene. If the gene group is clustered in the yeast functional network, then it is predictive of additional novel genetic interactions as measured by the AUC, the area under a receiver operating characteristic curve. On the right are highly predictive functional net clusters for each genetic interaction type, with genetic interactions and functional connections indicated by dotted and solid lines, respectively. For clarity, only edges above a weight cutoff and genes with more than one interaction to the seed set are shown. In the case of phenotypic suppression for instance, the seed set consists of the gene group CHK1, MEC1, RAD9, RAD17 and RAD24, which are all known to phenotypically suppress PSY2.

Bibliography

[1] Christopher P Adams and Van V Brantner. Estimating the cost of new drug development: is it really $802 million? *Health Affairs*, 25(2):420–428, 2006.

[2] Brian J Albert, Peter A McPherson, Kristine O'Brien, Nancy L Czaicki, Vincent DeStefino, Sami Osman, Miaosheng Li, Billy W Day, Paula J Grabowski, Melissa J Moore, et al. Meayamycin inhibits pre–messenger rna splicing and exhibits picomolar activity against multidrug-resistant cells. *Molecular Cancer Therapeutics*, 8(8):2308–2318, 2009.

[3] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[4] Alan Ashworth, Christopher J Lord, and Jorge S Reis-Filho. Genetic interactions in cancer progression and treatment. *Cell*, 145(1):30–38, 2011.

[5] Anastasia Baryshnikova, Michael Costanzo, Chad L Myers, Brenda Andrews, and Charles Boone. Genetic interaction networks: toward an understanding of heritability. *Annual Review of Genomics and Human Genetics*, 14:111–133, 2013.

[6] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[7] Benjamin Besse, David Planchard, Anne-Sophie Veillard, Laurent Taillade, David Khayat, Muriel Ducourtieux, Jean-Pierre Pignon, Jean Lumbroso, Carole Lafontaine, Claire Mathiot, et al. Phase 2 study of frontline bortezomib in patients with advanced non-small cell lung cancer. *Lung Cancer*, 76(1):78–83, 2012.

[8] Yang Bian, Risa Kitagawa, Parmil K Bansal, Yo Fujii, Alexander Stepanov, and Katsumi Kitagawa. Synthetic genetic array screen identifies pp2a as a therapeutic target in mad2-overexpressing tumors. *Proceedings of the National Academy of Sciences*, 111(4):1628–1633, 2014.

[9] CI Bliss. The toxicity of poisons applied jointly. *Annals of Applied Biology*, 26(3):585–615, 1939.

[10] Sophie Bonnal, Luisa Vigevani, and Juan Valcárcel. The spliceosome as a target of novel antitumour drugs. *Nature Reviews Drug Discovery*, 11(11):847–859, 2012.

[11] Matias Casás-Selves, Jihye Kim, Zhiyong Zhang, Barbara A Helfrich, Dexiang Gao, Christopher C Porter, Hannah A Scarborough, Paul A Bunn, Daniel C Chan, Aik Choon Tan, et al. Tankyrase and the canonical wnt pathway protect lung cancer cells from egfr inhibition. *Cancer Research*, 72(16):4154–4164, 2012.

[12] Hiu Wing Cheung, Glenn S Cowley, Barbara A Weir, Jesse S Boehm, Scott Rusin, Justine A Scott, Alexandra East, Levi D Ali, Patrick H Lizotte, Terence C Wong, et al. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences*, 108(30):12372–12377, 2011.

[13] Scott Cohen. *Finding color and shape patterns in images.* PhD thesis, Stanford University, 1999.

[14] Murat Cokol, Hon Nian Chua, Murat Tasan, Beste Mutlu, Zohar B Weinstein, Yo Suzuki, Mehmet E Nergiz, Michael Costanzo, Anastasia Baryshnikova, Guri Giaever, et al. Systematic exploration of synergistic drug pairs. *Molecular Systems Biology*, 7(1):544, 2011.

[15] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, Judice LY Koh, Kiana Toufighi, Sara Mostafavi, et al. The genetic landscape of a cell. *Science*, 327(5964):425–431, 2010.

[16] Glenn S Cowley, Barbara A Weir, Francisca Vazquez, Pablo Tamayo, Justine A Scott, Scott Rusin, Alexandra East-Seletsky, Levi D Ali, William FJ Gerath, Sarah E Pantel, et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific Data*, 1, 2014.

[17] Adrienne D Cox, Stephen W Fesik, Alec C Kimmelman, Ji Luo, and Channing J Der. Drugging the undruggable ras: mission possible? *Nature Reviews Drug Discovery*, 13(11):828–851, 2014.

[18] Kyle R Cron, Kaya Zhu, Deepa S Kushwaha, Grace Hsieh, Dmitry Merzon, Jonathan Rameseder, Clark C Chen, Alan D D'Andrea, and David Kozono. Proteasome inhibitors block dna repair and radiosensitize non-small cell lung cancer. *PloS One*, 8(9):e73710, 2013.

[19] Angela M Davies, Primo N Lara, Philip C Mack, and David R Gandara. Incorporating bortezomib into the treatment of lung cancer. *Clinical Cancer Research*, 13(15):4647s–4651s, 2007.

[20] Vincent T DeVita and Edward Chu. A history of cancer chemotherapy. *Cancer Research*, 68(21):8643–8653, 2008.

[21] Brian J Druker, Shu Tamura, Elisabeth Buchdunger, Sayuri Ohno, Gerald M Segal, Shane Fanning, Jürg Zimmermann, and Nicholas B Lydon. Effects of a selective inhibitor of the abl tyrosine kinase on the growth of bcr-abl positive cells. *Nature Medicine*, 2(5):561–566, 1996.

[22] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[23] Ferran Fece de la Cruz, Bianca V Gapp, and Sebastian MB Nijman.

Synthetic lethal vulnerabilities of cancer. *Annual Review of Pharmacology and Toxicology*, 55:513–531, 2015.

[24] Peter C Fong, David S Boss, Timothy A Yap, Andrew Tutt, Peijun Wu, Marja Mergui-Roelvink, Peter Mortimer, Helen Swaisland, Alan Lau, Mark J O'Connor, et al. Inhibition of poly (adp-ribose) polymerase in tumors from brca mutation carriers. *New England Journal of Medicine*, 361(2):123–134, 2009.

[25] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[26] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569–576, 2015.

[27] Ana Rita Grosso, Sandra Martins, and Maria Carmo-Fonseca. The emerging role of splicing factors in cancer. *EMBO Reports*, 9(11):1087–1093, 2008.

[28] Michal Grzmil and Brian A Hemmings. Translation regulation as a therapeutic target in cancer. *Cancer Research*, 72(16):3891–3900, 2012.

[29] Scott M Hammer, Kathleen E Squires, Michael D Hughes, Janet M Grimes, Lisa M Demeter, Judith S Currier, Joseph J Eron Jr, Judith E

Feinberg, Henry H Balfour Jr, Lawrence R Deyton, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and cd4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine*, 337(11):725–733, 1997.

[30] G Traver Hart, Insuk Lee, and Edward M Marcotte. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, 8(1):1, 2007.

[31] Pierre C Havugimana, G Traver Hart, Tamás Nepusz, Haixuan Yang, Andrei L Turinsky, Zhihua Li, Peggy I Wang, Daniel R Boutz, Vincent Fong, Sadhna Phanse, et al. A census of human soluble protein complexes. *Cell*, 150(5):1068–1081, 2012.

[32] Maureen E Hillenmeyer, Eula Fung, Jan Wildenhain, Sarah E Pierce, Shawn Hoon, William Lee, Michael Proctor, Robert P St Onge, Mike Tyers, Daphne Koller, et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, 320(5874):362–365, 2008.

[33] Cheuk Hei Ho, Leslie Magtanong, Sarah L Barker, David Gresham, Shinichi Nishimura, Paramasivam Natarajan, Judice LY Koh, Justin Porter, Christopher A Gray, Raymond J Andersen, et al. A molecular barcoded yeast orf library enables mode-of-action analysis of bioactive compounds. *Nature Biotechnology*, 27(4):369–377, 2009.

[34] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[35] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[36] John C Hunter, Deepak Gurbani, Scott B Ficarro, Martin A Carrasco, Sang Min Lim, Hwan Geun Choi, Ting Xie, Jarrod A Marto, Zhe Chen, Nathanael S Gray, et al. In situ selectivity profiling and crystal structure of sml-8-73-1, an active site inhibitor of oncogenic k-ras g12c. *Proceedings of the National Academy of Sciences*, 111(24):8895–8900, 2014.

[37] John D Hunter et al. Matplotlib: A 2d graphics environment. *Computing in Science and Engineering*, 9(3):90–95, 2007.

[38] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 245–251. IEEE, 2013.

[39] David R Jones, Christopher A Moskaluk, Heidi H Gillenwater, Gina R Petroni, Sandra G Burks, Jennifer Philips, Patrice K Rehm, Juan Olazagasti, Benjamin D Kozower, and Yongde Bao. Phase i trial of induction

histone deacetylase and proteasome inhibition followed by surgery in non–small-cell lung cancer. *Journal of Thoracic Oncology*, 7(11):1683–1690, 2012.

[40] Kenneth I Kaitin. Deconstructing the drug development process: the new face of innovation. *Clinical Pharmacology and Therapeutics*, 87(3):356, 2010.

[41] Hyun Seok Kim, Saurabh Mendiratta, Jiyeon Kim, Chad Victor Pecot, Jill E Larsen, Iryna Zubovych, Bo Yeun Seo, Jimi Kim, Banu Eskiocak, Hannah Chung, et al. Systematic identification of molecular subtype-selective vulnerabilities in non-small-cell lung cancer. *Cell*, 155(3):552–566, 2013.

[42] Alexei F Kisselev, Wouter A van der Linden, and Herman S Overkleeft. Proteasome inhibitors: an expanding army attacking a unique target. *Chemistry & Biology*, 19(1):99–115, 2012.

[43] Michael Kuhn, Damian Szklarczyk, Sune Pletscher-Frankild, Thomas H Blicher, Christian von Mering, Lars J Jensen, and Peer Bork. Stitch 4: integration of protein–chemical interactions with user data. *Nucleic Acids Research*, page gkt1207, 2013.

[44] Scott A Laurie and Glenwood D Goss. Role of epidermal growth factor receptor inhibitors in epidermal growth factor receptor wild-type non–small-cell lung cancer. *Journal of Clinical Oncology*, 31(8):1061–1069, 2013.

[45] Insuk Lee, U Martin Blom, Peggy I Wang, Jung Eun Shim, and Edward M Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 21(7):1109–1121, 2011.

[46] Insuk Lee, Ben Lehner, Tanya Vavouri, Junha Shin, Andrew G Fraser, and Edward M Marcotte. Predicting genetic modifier loci using functional gene networks. *Genome Research*, 20(8):1143–1153, 2010.

[47] Insuk Lee, Zhihua Li, and Edward M Marcotte. An improved, bias-reduced probabilistic functional gene network of baker's yeast, saccharomyces cerevisiae. *PLoS One*, 2(10):e988, 2007.

[48] Hecheng Li, XiaoLi Zhu, Yawei Zhang, Jiaqing Xiang, and Haiquan Chen. Arsenic trioxide exerts synergistic effects with cisplatin on non-small cell lung cancer cells via apoptosis induction. *Journal of Experimental & Clinical Cancer Research*, 28(1):1, 2009.

[49] S Loewe. The problem of synergism and antagonism of combined drugs. *Arzneimittel-Forschung*, 3(6):285, 1953.

[50] Hendrik Luesch, Tom YH Wu, Pingda Ren, Nathanael S Gray, Peter G Schultz, and Frantisek Supek. A genome-wide overexpression screen in yeast for small-molecule target identification. *Chemistry & Biology*, 12(1):55–63, 2005.

[51] Pek Yee Lum, Christopher D Armour, Sergey B Stepaniants, Guy Cavet, Maria K Wolf, J Scott Butler, Jerald C Hinshaw, Philippe Garnier, Glenn D Prestwich, Amy Leonardson, et al. Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell*, 116(1):121–137, 2004.

[52] Ramamurthy Mani, Robert P St Onge, John L Hartman, Guri Giaever, and Frederick P Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences*, 105(9):3461–3466, 2008.

[53] Deepak Nijhawan, Travis I Zack, Yin Ren, Matthew R Strickland, Rebecca Lamothe, Steven E Schumacher, Aviad Tsherniak, Henrike C Besche, Joseph Rosenbluh, Shyemaa Shehata, et al. Cancer vulnerabilities unveiled by genomic loss. *Cell*, 150(4):842–854, 2012.

[54] Xuewen Pan, Stefanie Reissman, Nick R Douglas, Zhiwei Huang, Daniel S Yuan, Xiaoling Wang, J Michael McCaffery, Judith Frydman, and Jef D Boeke. Trivalent arsenic inhibits the functions of chaperonin complex. *Genetics*, 186(2):725–734, 2010.

[55] Gaurav Pandey, Bin Zhang, Aaron N Chang, Chad L Myers, Jun Zhu, Vipin Kumar, and Eric E Schadt. An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Comput Biol*, 6(9):e1000928, 2010.

[56] Woo Hyun Park and Suhn Hee Kim. Arsenic trioxide induces human

pulmonary fibroblast cell death via the regulation of bcl-2 family and caspase-8. *Molecular Biology Reports*, 39(4):4311–4318, 2012.

[57] Ainslie B Parsons, Renée L Brost, Huiming Ding, Zhijian Li, Chaoying Zhang, Bilal Sheikh, Grant W Brown, Patricia M Kane, Timothy R Hughes, and Charles Boone. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nature Biotechnology*, 22(1):62–69, 2004.

[58] Ainslie B Parsons, Andres Lopez, Inmar E Givoni, David E Williams, Christopher A Gray, Justin Porter, Gordon Chua, Richelle Sopko, Renee L Brost, Cheuk-Hei Ho, et al. Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. *Cell*, 126(3):611–625, 2006.

[59] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[60] R Pincheira, Q Chen, and JT Zhang. Identification of a 170-kda protein over-expressed in lung cancers. *British Journal of Cancer*, 84(11):1520, 2001.

[61] Bilal Piperdi, William V Walsh, Kendra Bradley, Zheng Zhou, Venu Bathini, Meredith Hanrahan-Boshes, Lloyd Hutchinson, and Roman

Perez-Soler. Phase-i/ii study of bortezomib in combination with carboplatin and bevacizumab as first-line therapy in patients with advanced non–small-cell lung cancer. *Journal of Thoracic Oncology*, 7(6):1032–1040, 2012.

[62] Yan Qi, Yasir Suhail, Yu-yi Lin, Jef D Boeke, and Joel S Bader. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Research*, 18(12):1991–2004, 2008.

[63] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[64] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner Mewes. Corum: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Research*, 38(suppl 1):D497–D501, 2010.

[65] Colm J Ryan, Nevan J Krogan, Pádraig Cunningham, and Gerard Cagney. All or nothing: protein complexes flip essentiality between distantly related eukaryotes. *Genome Biology and Evolution*, 5(6):1049–1059, 2013.

[66] Daniel A Schult and P Swart. Exploring network structure, dynamics,

and function using networkx. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, volume 2008, pages 11–16, 2008.

[67] Andrey A Shabalin, Victor J Weigman, Charles M Perou, and Andrew B Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, pages 985–1012, 2009.

[68] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[69] Alice T Shaw and Jeffrey A Engelman. Alk in lung cancer: past, present, and future. *Journal of Clinical Oncology*, 31(8):1105–1111, 2013.

[70] Zhi-Xiang Shen, Guo-Qiang Chen, Jian-Hua Ni, Xiu-Shong Li, Shu-Min Xiong, Qian-Yao Qiu, Jun Zhu, Wei Tang, Guan-Lin Sun, Kan-Qi Yang, et al. Use of arsenic trioxide (as2o3) in the treatment of acute promyelocytic leukemia (apl): Ii. clinical efficacy and pharmacokinetics in relapsed patients. *Blood*, 89(9):3354–3360, 1997.

[71] Junha Shin, Sunmo Yang, Eiru Kim, Chan Yeong Kim, Hongseok Shim, Ara Cho, Hyojin Kim, Sohyun Hwang, Jung Eun Shim, and Insuk Lee. Flynet: a versatile network prioritization server for the drosophila community. *Nucleic Acids Research*, page gkv453, 2015.

[72] Nitin Kumar Singh, Bo Yeun Seo, Mathukumalli Vidyasagar, Michael A White, and Hyun Seok Kim. simacro: A fast and easy data processing tool for cell-based genomewide sirna screens. *Genomics & Informatics*, 11(1):55–57, 2013.

[73] Steven L Soignet, Stanley R Frankel, Dan Douer, Martin S Tallman, Hagop Kantarjian, Elizabeth Calleja, Richard M Stone, Matt Kalaycio, David A Scheinberg, Peter Steinherz, et al. United states multicenter study of arsenic trioxide in relapsed acute promyelocytic leukemia. *Journal of Clinical Oncology*, 19(18):3852–3860, 2001.

[74] Steven L Soignet, Peter Maslak, Zhu-Gang Wang, Suresh Jhanwar, Elizabeth Calleja, Laura J Dardashti, Diane Corso, Anthony DeBlasio, Janice Gabrilove, David A Scheinberg, et al. Complete remission after treatment of acute promyelocytic leukemia with arsenic trioxide. *New England Journal of Medicine*, 339(19):1341–1348, 1998.

[75] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1):D535–D539, 2006.

[76] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[77] David C Swinney and Jason Anthony. How were new medicines discovered? *Nature Reviews Drug Discovery*, 10(7):507–519, 2011.

[78] Ronald J Tallarida. An overview of drug combination analysis with isobolograms. *Journal of Pharmacology and Experimental Therapeutics*, 319(1):1–7, 2006.

[79] Craig M Tamble, Robert P St Onge, Guri Giaever, Corey Nislow, Alexander G Williams, Joshua M Stuart, and R Scott Lokey. The synthetic genetic interaction network reveals small molecules that target specific pathways in sacchromyces cerevisiae. *Molecular BioSystems*, 7(6):2019–2030, 2011.

[80] RJ Van Alphen, EAC Wiemer, Herman Burger, and FALM Eskens. The spliceosome as target for anticancer treatment. *British Journal of Cancer*, 100(2):228–232, 2009.

[81] Haizhou Wang and Mingzhou Song. Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. *The R Journal*, 3(2):29–33, 2011.

[82] Jan Wildenhain, Michaela Spitzer, Sonam Dolma, Nick Jarvik, Rachel White, Marcia Roy, Emma Griffiths, David S Bellows, Gerard D Wright, and Mike Tyers. Prediction of synergism from chemical-genetic interactions by machine learning. *Cell Systems*, 1(6):383–395, 2015.

[83] Alan SL Wong, Gigi CG Choi, Cheryl H Cui, Gabriela Pregernig, Pamela Milani, Miriam Adam, Samuel D Perli, Samuel W Kazer, Aleth Gaillard,

Mario Hermann, et al. Multiplexed barcoded crispr-cas9 screening enabled by combigem. *Proceedings of the National Academy of Sciences*, 113(9):2544–2549, 2016.

[84] Sharyl L Wong, Lan V Zhang, Amy HY Tong, Zhijian Li, Debra S Goldberg, Oliver D King, Guillaume Lesage, Marc Vidal, Brenda Andrews, Howard Bussey, et al. Combining biological networks to predict genetic interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 101(44):15682–15687, 2004.

[85] Roberta J Worthington and Christian Melander. Combination approaches to combat multidrug-resistant bacteria. *Trends in Biotechnology*, 31(3):177–184, 2013.

[86] Chunlei Wu, Ian MacLeod, and Andrew I Su. Biogps and mygene. info: organizing online, gene-centric information. *Nucleic Acids Research*, page gks1114, 2012.

[87] Pamela J Yeh, Matthew J Hegreness, Aviva Presser Aiden, and Roy Kishony. Drug interactions and the evolution of antibiotic resistance. *Nature Reviews Microbiology*, 7(6):460–466, 2009.

[88] Lili Zhang, Xiaoyu Pan, and John WB Hershey. Individual overexpression of five subunits of human translation initiation factor eif3 promotes malignant transformation of immortal fibroblast cells. *Journal of Biological Chemistry*, 282(8):5790–5800, 2007.

[89] Weiwei Zhong and Paul W Sternberg. Genome-wide prediction of c. elegans genetic interactions. *Science*, 311(5766):1481–1484, 2006.