Title
The analysis and integration of high-throughput biological data for pathway discovery

Author
Kelley, Ryan Matthew

Publication Date
2009

UNIVERSITY OF CALIFORNIA, SAN DIEGO

THE ANALYSIS AND INTEGRATION OF HIGH-THROUGHPUT BIOLOGICAL DATA FOR PATHWAY DISCOVERY

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Bioinformatics

by

Ryan Matthew Kelley

Committee in charge:

Professor Trey Ideker, Chair
Professor Vineet Bafna, Co-chair
Professor Bing Ren
Professor Shankar Subramaniam
Profressor Jean Wang

2009

The dissertation of Ryan Matthew Kelley is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-chair

Chair

University of California, San Diego

2009

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to acknowledge the support of my advisor, Dr. Trey Ideker, without whom none of this work would have been possible. Dr. Ideker provided invaluable guidance and was always able to provide a fresh and insightful view on research problems. The many collaborations which were so valuable to my studies would not have been possible without Dr. Ideker. Finally, he provided a sense of enthusiasm and dedication to the study of biology and bioinformatics.

I would also like to thank all of the members of my committee for the time and support they have provided in the preparation of my dissertation. I would especially like to thank Dr. Jean Wang for her contribution. She provided valuable insight and encouraged me to consider additional angles and possibilities regarding my research.

I also owe a great debt to my many co-authors; first, for their valuable scientific work, and second, for extending their permission to use our joint work in this dissertation. Astrid Haugen was crucial in providing voluminous amounts of data and biological insight to ground the development and use of network analysis algorithms in the study arsenic toxicity. Although not a co-author, Owen Ozier initially suggested and implemented the neighborhood-scoring scheme used in that work. Sourav Bandyopadhyay was the driving force in the conceptualization and completion of the extension to my work on the analysis of genetic interaction networks. Without his dedication, the work would never have been started, much less completed.

Chapter 2, in full, is a reprint of the following work,

Kelley R, Ideker T. *Systematic interpretation of genetic interactions using protein networks*. **Nature Biotechnology** 2005; 23(5):561-6.

The dissertation author was the sole first author on this paper, responsible for designing, implementing, and running computational algorithms.

Chapter 3, in full, is a reprint of the following work,

Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T. *Functional maps of protein complexes from quantitative genetic interaction data*. **PLoS Computational Biology** 2008; 4(4).

The dissertation author was the second author on this work, responsible for designing and implementing computational algorithms.

Chapter 4, in full, is a reprint of the following work,

Kelley, R., Feizi H., Ideker T. Correcting for gene-specific dye bias in DNA microarrays using the method of maximum likelihood. **Bioinformatics** (2007).

The dissertation author was the sole first author on this paper, responsible for designing, performing, and analyzing experiments and algorithms.

Chapter 5, in full, is a reprint of the following work,

Haugen AC, Kelley R, Collins JB, Tucker CJ, Deng C, Afshari CA, Brown JM, Ideker T, Van Houten B. *Integrating phenotypic and expression profiles to map arsenic-response networks*. **Genome Biology** 2004;5(12):R95.

The dissertation author was the second author on this work, responsible for designing and implementing network analysis algorithms.

Chapter 6, in full, is a copy of the following manuscript currently under preparation,

Kelley R, Ideker T. Genome-wide fitness and expression profiling implicate Mga2 in adaptation to hydrogen peroxide. **In preparation**.

The dissertation author is the sole first author on this work, responsible for designing and executing experiments and computational algorithms.

VITA

2002	Bachelor of Science	University of Arizona
2009	Doctor of Philosophy	University of California, San Diego

PUBLICATIONS

Kelley R, Feizi H, Ideker T. *Correcting for gene-specific dye bias in DNA microarrays using the method of maximum likelihood*. **Bioinformatics** 2007.

Cline MS, *et al*. *Integration of biological networks and gene expression data using Cytoscape*. **Nature Protocols** 2007; 2(10): 2366-82.

Bandyopadhyay S, Kelley R, Ideker T. *Discovering regulated networks during HIV-1 latency and reactivation*. **Pacific Symposium on Biocomputing** 2006; 354-66

Kelley R, Ideker T. *Systematic interpretation of genetic interactions using protein networks*. **Nature Biotechnology** 2005; 23(5): 561-6.

Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T. *Conserved patterns of protein interaction in multiple species*. **PNAS** 2005;102(6):1974-9.

Haugen AC, Kelley R, Collins JB, Tucker CJ, Deng C, Afshari CA, Brown JM, Ideker T, Van Houten B. *Integrating phenotypic and expression profiles to map arsenic-response networks*. **Genome Biology** 2004; 5(12): R95.

ABSTRACT OF THE DISSERTATION

The analysis and integration of high-throughput biological data for pathway discovery

by

Ryan Matthew Kelley

Doctor of Philosophy in Bioinformatics

University of California, San Diego 2009

Professor Trey Ideker, Chair

Professor Vineet Bafna, Co-chair

The past decade has seen the creation and maturation of a number of new technologies designed to study life on a genome-wide scale. However, the sheer volume of data generated by these methods surpasses the analytical and critical abilities of a single researcher. For this reason, it is necessary to create new computational methods to

assist in the analysis of these new sources of data.

Both yeast-two hybrid and co-immunoprecipitation followed by mass spectrometry allow the determination of binding interactions between proteins. Functional (genetic) interactions are determined via SGA (Synthetic Genetic Array) and E-MAP (Epistasis Mini-array Profile). In Chapters 2 and 3, we develop algorithms to integrate these two types of interactions together for the purpose of biological pathway discovery. Moreover, our approaches create maps of genetic interactions that provide a picture of the global organization of pathways and complexes within the cell.

Expression arrays are a genome-wide quantitative assay for mRNA levels within the cell. Using fluorescent dyes, two different biological samples can be directly compared on a single array slide. In Chapter 4, we identify a gene-specific dye bias in this type of expression array data. We improve upon a maximum likelihood method in order to remove the effect of this bias. Using novel control experiments, we show that this enhanced analysis yields results that more reproducible.

Complementary to expression profiling, deletion fitness profiling quantifies the relative fitness defect of every deletion strain in *Saccharomyces cerevisiae* under a particular stress. In Chapters 5 and 6, we discuss how to use the type of pathway information uncovered in Chapters 2 and 3 to improve the analysis of both expression and deletion fitness profiling datasets. We apply these methods to the study of two different cellular stresses in *Saccharomyces cerevisiae*, arsenic exposure and adaptation to oxidative stress.

Chapter 1. Introduction

The study of biology arose from the wealth of diversity in the natural world. Why is man different than other animals? Why is one man sick while another is healthy? Our understanding of these questions took a great leap forward with the elucidation of the central dogma of biology. Every living organism has a genetic code, a central repository of information encoded in its DNA with four nucleotide bases. Using this genetic code as a template, a strand of messenger RNA is transcribed. Compared to DNA, messenger RNA is much less stable. Fortunately, it need only survive the trip to the ribosome. The information contained within the messenger RNA instructs the ribosome how to construct a corresponding protein. The nature of these proteins as well as how they interact with each other defines the phenotype of an organism. While additional variations upon the central dogma have been discovered, it is still incredibly important. A central question in modern biology is therefore to determine how the process of the central dogma results in a particular phenotype.

In the study of the central dogma, biologists have devised a number of methods to interrogate the process at every stage. Sanger sequencing can be used to identify the sequence of a particular gene[1]. A Northern blot reveals the quantity (expression) of an mRNA transcript within a cell[2]. Interactions between two different proteins can be detected with the yeast-two hybrid assay[3]. Even today, the application of these techniques greatly increases our understanding of the function of the living cell. However, a limitation to these technologies became readily apparent. With over 6,000 genes present in even a relatively simple organism such as *Saccharomyces cerevisiae*[4], performing these screens on a genome-wide scale is infeasible. Thus, these screens must be conducted in a relatively targeted fashion, making it difficult to identify novel patterns of expression or interaction.

In order to combat this limitation, new technologies were introduced to investigate the components of the central dogma on a genome-wide scale. Often, nano-technology and robotics were used to augment an existing technique so that it could be performed rapidly in parallel. For example, the expression microarray is similar to a Northern blot. In a Northern blot, a labeled nucleotide is used to probe for a sequence of interest in an RNA sample that has been separated on a gel and transferred to a membrane[2]. In the miniaturized expression array analog, probes for multiple different sequences are spotted onto a glass slide. The diameter of each spot is typically less than 100 uM in diameter, allowing for tens of thousands of spots on a single slide. The RNA sample is then labeled with a fluorescent dye. The binding of labeled transcripts to the probe sequences can be visualized with a confocal scanning microscope[5]. Similarly, robotic technology can dramatically increase the throughput of the yeast two-hybrid screen. A two-hybrid screen requires generating a cross of two yeast strains. In one strain (the prey), a protein is linked to a transcription activation domain, while in the other strain (the bait), a (possibly) different protein is chained to a DNA binding domain. If the two proteins interact, these two domains are brought together, allowing transcription of the marker gene downstream of the bound promoter[3]. Even in a relatively small genome such as *Saccharomyces cerevisiae*, testing for all possible protein interactions requires the generation of a huge number of crosses (over 36,000,000). The generation of these crossed strains can be greatly accelerated with a replica pinning robot[6].

These technologies represent a huge promise of new biological discoveries. However, they also lead to an additional set of complications. First, when a technique is miniaturized or performed by robot, it may no longer be as reliable as when it is performed by the hands of a skilled biologist. Early yeast two-hybrid studies are estimated to have an accuracy of between 1%-10% at a coverage of just 1%[7]. However, even if a massive increase in throughput does not add additional noise into an assay, there is still the problem of "multiple hypothesis testing."

The problem of multiple hypothesis testing is related to the statistical analysis of experimental data. The analyses of most experiments are designed to differentiate between two possible hypotheses, the null hypothesis (the observed data is no different than one would expect to observe at random) and the alternate hypothesis (the observed data supports a predicted result). The p-value, which represents the probability of obtaining a given result under the null hypothesis, it typically used to make this distinction. By convention, the p-value threshold is typically set to 0.05, meaning there is only a 5% chance of generating the observed data under the null hypothesis. However, in a genome-wide screen, this practice can lead to complications. Imagine 100 experiments in which the null hypothesis is actually true in every case. Setting a p-value threshold of 0.05 would result in an incorrect rejection of the null hypothesis 5% of the time, creating a large number of false positives. In a single targeted experiment, there is usually a reason to investigate a particular interaction or gene, leading to a relatively high proportion of cases where the null hypothesis is actually false. By extending our search to the entire genome, we examine many more cases where the null hypothesis is true, greatly increasing the false positive rate.

The simplest way to address this problem is with a greater number of replicates. With a greater number of replicates, it is generally possible to set a more stringent p-value threshold, reducing the problem of false positives. Unfortunately, performing more replicates is time consuming and expensive, especially in the case of genome-wide studies.

Conversely, improved statistical analysis of the data may also allow for a more stringent p-value threshold. The determination of a p-value implicitly depends on the error model of the underlying process. In a genome-wide screen, we are presented with a wealth of data that can be used to accurately determine an appropriate error model for our data. In Chapter 2, we describe this process for a microarray expression study. Through analysis of control experiments, we identify the presence of a systematic bias in the data. We find that the relative efficiency for labeling with the Cy3 and Cy5 dyes varies on a gene-by-gene basis, an effect termed gene-specific dye bias. By incorporating the effect of this gene-specific dye bias into our error model, we are able to identify more differentially expressed genes. Furthermore, the sets of differentially expressed genes are more reproducible over repeated experiments using different labeling methods. Despite the potential for increased reliability, we must face the reality that in a genome-wide screen, a single measurement is often unreliable.

Interpretation of genome-wide data can be greatly improved by an understanding of how genes and proteins are organized in the cell. In order to accomplish a large task, such as response to disease or stress, proteins rarely act alone. Instead, groups of proteins often work together in pathways to accomplish the same goal. While an experimental measurement corresponding to a single gene may be unreliable; an aggregate measure over an entire pathway is more robust to both experimental noise and missing data. Furthermore, since there are fewer pathways than there are genes, the problem of multiple hypothesis testing is reduced. Unfortunately, the exact constituents of the pathways present in the cell are often partially or fully unknown. Thus, methods for the identification of the constituents of cellular pathways can greatly aid in the interpretation of genome-wide data.

In Chapter 2, we describe a method for identifying protein complexes/pathways from a combined dataset of physical and genetic interactions. The set of physical interactions includes regulatory (DNA-binding), metabolic (shared substrate), and protein-protein binding interactions. The genetic interaction dataset contains both synthetic sick and synthetic lethal interactions derived using SGA (synthetic genetic array) technology[8]. In a synthetic lethal or sick interaction, a double deletion mutant is found to be inviable or sick, while both corresponding single deletion mutants are healthy. Such an interaction represents a functional link between the two genes. We propose two models of genetic interactions, within-pathway and between-pathway interaction. In within-pathway interaction, both physical and genetic interactions occur among the proteins of a single pathway. In between-pathway interactions, physical interactions still occur primarily between members of the same complex, but genetic interactions occur between members of a two different pathways. We search for individual pathways or pairs of pathways corresponding to the within- and between-pathway genetic explanation. We find that both models can be used to detect pathways from the combined dataset; however, between-pathway interactions are the predominant mode of genetic interaction.

In Chapter 3, we expand upon the previous work of Chapter 2 to address certain limitations present in the approach. Following our original work, new technologies were created for the identification of genetic interactions. In the previous SGA approach, genetic interactions between two genes are represented as a binary value. However, in the new technology of E-MAP, genetic interaction is quantitative[9]. Each tested gene pair is assigned a real value, corresponding to the aggravating (negative) or alleviating (positive) extent of the genetic interaction. In Chapter 3, we describe how our previous algorithm can be updated to take advantage of this quantitative information. Furthermore, the previous algorithm was limited in that it would search for individual pathways or pairs of pathways, corresponding to the within- and between-pathway searches, respectively. Multiple pathways may be identified that are very similar to each other. In comparison, the global search algorithm we describe in Chapter 3 simultaneously identifies a mutually exclusive set of pathways. In addition, our previous work conducted separate within- and between-pathway searches. In Chapter 3, we combine both searches to identify previously unknown complexes based on the strength of both types of explanations.

Chapters 2 and 3 illustrate how we can identify potential pathways from a diverse set of interaction data. However, it is also necessary to identify the context in which the cell is utilizing those pathways. The final chapters of this dissertation describe how we can identify active pathways from a combined dataset of expression and deletion-fitness profiling data. Building on the approach of Ideket *et al.*[10], Chapter 5 describes the identification of active pathways in the arsenic-stress response. In previous work, Ideker *et al.* identified pathways involved in galactose utilization by integrating microarray expression studies with an interaction network. Any connected component in a restricted network of interactions between genes with a suspected role in the galactose response was considered a potential pathway. The overall expression of each potential pathway was measured with a modified average of normalized expression values. However, in the expanded interaction network we employ in Chapter 5, the approach of Ideker *et al.* fails to identify activated subnetworks. We believe this is because the definition of a pathway is too liberal, leading to the spurious identification of many uninformative results. To combat this problem, we utilize a neighborhood scoring strategy, in which the set of possible pathways is restricted to the complete set of neighbors of any connected group of genes. While this restricts the possible space of pathways, it is still able to capture many of the true pathways present in the interaction data. For example, it allows us to identify a transcription factor and all of its targets or a densely connected set of proteins that form a protein complex. Specifically, we identify the activation of the stress response transcription factors Yap1, Msn2/4, and Hsf1 among others. The identification of the RPN4 transcription factor and the proteasomal protein complex implicates the process of protein degradation in the arsenic response.

However, using this approach with a combination of deletion fitness data and a network of metabolic interactions fails to identify any significant results. Instead, we modify our approach to only consider linear pathways through the metabolic network, representing chains of metabolic interactions. Using this approach we identify pathways of metablic reactions responsible for serine, threonine and glutamate metabolism. Further results indicate a role for the shikimate pathway, which is essential for the production of *p*-aminobenzoic acid (PABA) among other metabolic end products.

In Chapter 6, we apply similar approaches in the analysis of adaptation to hydrogen peroxide. Adaptation describes a process by which a mild dose of oxidant is able to confer protection against a later acute dose[11]. Oxidative stress is an important factor in a variety of human diseases, including aging, neural degeneration, and cardiovascular disease[12-14]. The study of adaptation may reveal how the influence of oxidative stress on these disease processes can be mitigated. Since adaptation to an oxidant requires active transcription[11], we primarily apply network analysis methods to the transcriptional network. By examining sets of transcription factor targets, we identify transcription factors with a large number of differentially expressed targets. In integrating the expression values with deletion fitness profiling data, we find that those transcription factors with differentially expressed targets also tend to correspond to sensitive deletion strains. Furthermore, the activities of these transcription factors are confirmed with expression profiling of transcription factor deletion strains.

Chapter 2. Systematic interpretation of genetic interactions using protein networks

Abstract

Genetic interaction, in which two mutations have a combined effect not exhibited by either mutation alone, is a powerful and widespread tool for establishing functional linkages between genes. In the yeast *Saccharomyces cerevisiae*, ongoing screens have generated >4,800 genetic interactions. We demonstrate that by combining these data with information on protein-protein, protein-DNA, or metabolic networks, it is possible to uncover physical mechanisms behind many of the observed genetic effects. Using a probabilistic model, we find that 1,922 genetic interactions are significantly associated with either between- or within-pathway explanations encoded in the physical networks, covering ~40% of known genetic interactions. These models predict new functions for 343 proteins and suggest that between-pathway explanations are better than within-pathway explanations at interpreting genetic interactions identified in systematic screens. This study provides a roadmap for how genetic and physical interactions can be integrated to reveal pathway organization and function.

Introduction

A major challenge of biotechnology and genetics is to interpret observed genetic interactions in a physical cellular context[15-17]. Genetic interactions consist of several major varieties: synthetic-lethal interactions, in which mutations in two nonessential genes are lethal when combined; suppressor interactions, in which one mutation is lethal but combination with a second restores cell viability; and an array of other effects such as enhancement and epistasis. Genetic interactions have been used extensively to shed light on pathway organization in model organisms[15-18]. In humans, genetic interactions are critical in linkage analysis of complex diseases[19] and in discovery of new pharmaceuticals[20]. Although genetic interactions are classically identified by mutant screens[21], recent studies have applied systematic "reverse" methods such as Synthetic Genetic Arrays (SGA)[22] or Synthetic Lethal Analysis by Microarrays (SLAM)[23] to catalog ~4,000 synthetic-lethal and synthetic-sick interactions in Saccharomyces cerevisiae.

Due to the high-throughput nature of SGA, discovery of new genetic interactions is largely automated. However, interpreting the functional significance of each result remains a relatively slow process. The problem is compounded by the large number of genetic interactions measured when screening one gene versus all others (~34 on average[8]) as well as possible false positives if the interactions are not confirmed by tetrad or random spore analysis. Thus, without further methods to aid in characterizing synthetic lethals, large-scale interpretation is a daunting prospect.

A promising solution may be to integrate synthetic lethals with other types of high-throughput interactions. For instance, direct physical interactions among proteins are being mapped by systematic two-hybrid[24-28] or immunoprecipitation studies[29, 30], while physical interactions between transcription factors and promoter sites are determined using chromatin-immunoprecipitation in conjunction with DNA microarrays[31, 32]. These interactions comprise a physical network which correlates with the network of genetic interactions and provides potential clues as to the mechanisms behind particular synthetic-lethal effects. Previous studies have demonstrated this correlated structure in yeast, by showing that two proteins in the same region of the genetic network are likely to also physically interact[8, 22]; that genes with similar patterns of genetic interactions often occur with the same protein complex[8]; and that a protein with many interactions in the physical network typically has many interactions in the genetic network also[33].

These studies suggest that it may be possible to interpret observed synthetic-lethal relationships explicitly using physical interactions. In this regard, previous authors[16, 34] have noted that synthetic-lethal genetic interactions are typically associated with one of three types of physical interpretations:

1) Between-pathway interpretations. The genetic interaction bridges genes operating in two pathways with redundant or complementary functions. Deletion of either gene is expected to abrogate the function of one but not both pathways.
2) Within-pathway interpretations. The genetic interaction occurs between protein subunits within a single pathway. A single gene is dispensable for the function of the overall pathway, but the additive effects of several gene deletions are lethal.
3) Indirect effects. The synthetic lethal phenotype is not mediated by a localized mechanism in the physical network. Indirect effects can occur because a deletion

phenotype represents not just the absence of one particular gene, but also the response of the cell to its absence, involving many diverse pathways[34].

Here, we demonstrate a computational framework for assembling genetic and physical interactions into models corresponding to the between- versus within-pathway interpretations. Regions of the physical network which correspond to each type of model are identified using a probabilistic scoring scheme. These models predict new protein functions and suggest that genetic interactions are more likely to bridge redundant or complementary processes than to combine additively within the same process.

Results

Construction of genetic and physical networks.

We assembled a genetic interaction network from two primary data sources (Figure 2.1). The first was generated by SGA, a large-scale screen[8] crossing 132 yeast gene deletion strains versus each of the ~4,700 available deletion strains[35] and resulting in 2,012 observed synthetic-lethal interactions and 2,113 synthetic-sick interactions. The second data source consisted of an additional 687 synthetic-lethal interactions culled from the literature and catalogued at the Munich Information Center for Protein Sequences (MIPS)[36]. The combined genetic network synthesizing these data consisted of 1,434 proteins (genes) linked by 4,812 synthetic-lethal interactions.

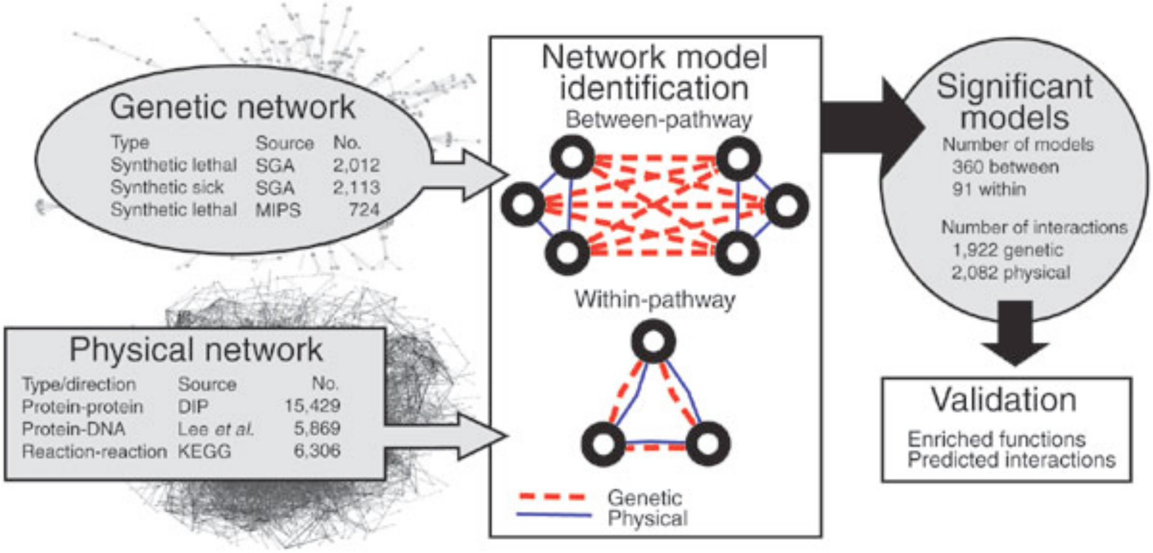


Figure 2.1. Method overview.

A combined physical and genetic network is searched to identify between- or within-pathway models of genetic interactions. The between-pathway model implies two groups of proteins (pathways) with many physical connections within each pathway (solid blue links) and genetic interactions spanning between pathways (dotted red links). The within-pathway model implies many physical and genetic interactions within the same group of proteins. In the search, 360 and 91 network models were identified that correspond to between- or within-pathway searches, respectively.

We also assembled a physical network of 5,993 yeast proteins connected by physical interactions of three types: 15,429 protein-protein interactions (the two proteins a and b display physical binding); 5,869 protein-DNA regulatory interactions (a binds upstream of the gene encoding b), and 6,306 shared-reaction metabolic relationships (a and b are enzymes that operate on at least one metabolite in common). The protein-protein interactions were downloaded from the DIP database[37] as of July 2004 and predominantly included data from large-scale experiments[25, 28-30]. The protein-DNA interactions were obtained from a large-scale chromatin-IP study of 106 transcription factors[32] (interactions with P = 0.001). Enzymatic reactions linked by common metabolites were obtained from KEGG[38], excluding metabolite cofactors such as ATP or

H2O (listed in Supplemental Table 2.1). The combined physical network covered 94.4% of all proteins in the genetic network. Both networks are provided at http://www.cellcircuits.org/Kelley2005/ in Cytoscape[39] (SIF) format.

Between-pathway interpretations for genetic interactions.

Preliminary statistical analyses confirm a limited relationship between genetic and physical interactions (see Supplemental Figure 4.1 and Tong et al.[8, 22]), but demonstrate a need for structured models to efficiently separate signal from noise. Towards this goal, we implemented a probabilistic modeling procedure to capture the between-pathway interpretation of genetic interactions. As described in Methods, this procedure involved a search for pairs of physical pathways that were densely connected by genetic interactions, in which a “pathway” was loosely defined as any densely-connected set of proteins in the physical network (this definition generically covers many network structures, including protein complexes). Pairs of pathways (constituting a single network model; see Figure 2.1) were assigned a score proportional to the density of physical interactions falling within each pathway and the density of genetic interactions bridging between pathways. This search generated 360 significant models covering 401 pathways and incorporating a total of 1,573 genetic interactions (196 MIPS, 687 SGA synthetic lethal, 690 SGA synthetic sick) and 1,931 physical interactions (1,248 protein binding, 77 regulatory, 606 shared reaction). Significance of these models was assessed by comparison to random genetic and physical networks. Detailed information for all models is provided at http://www.cellcircuits.org/Kelley2005/.

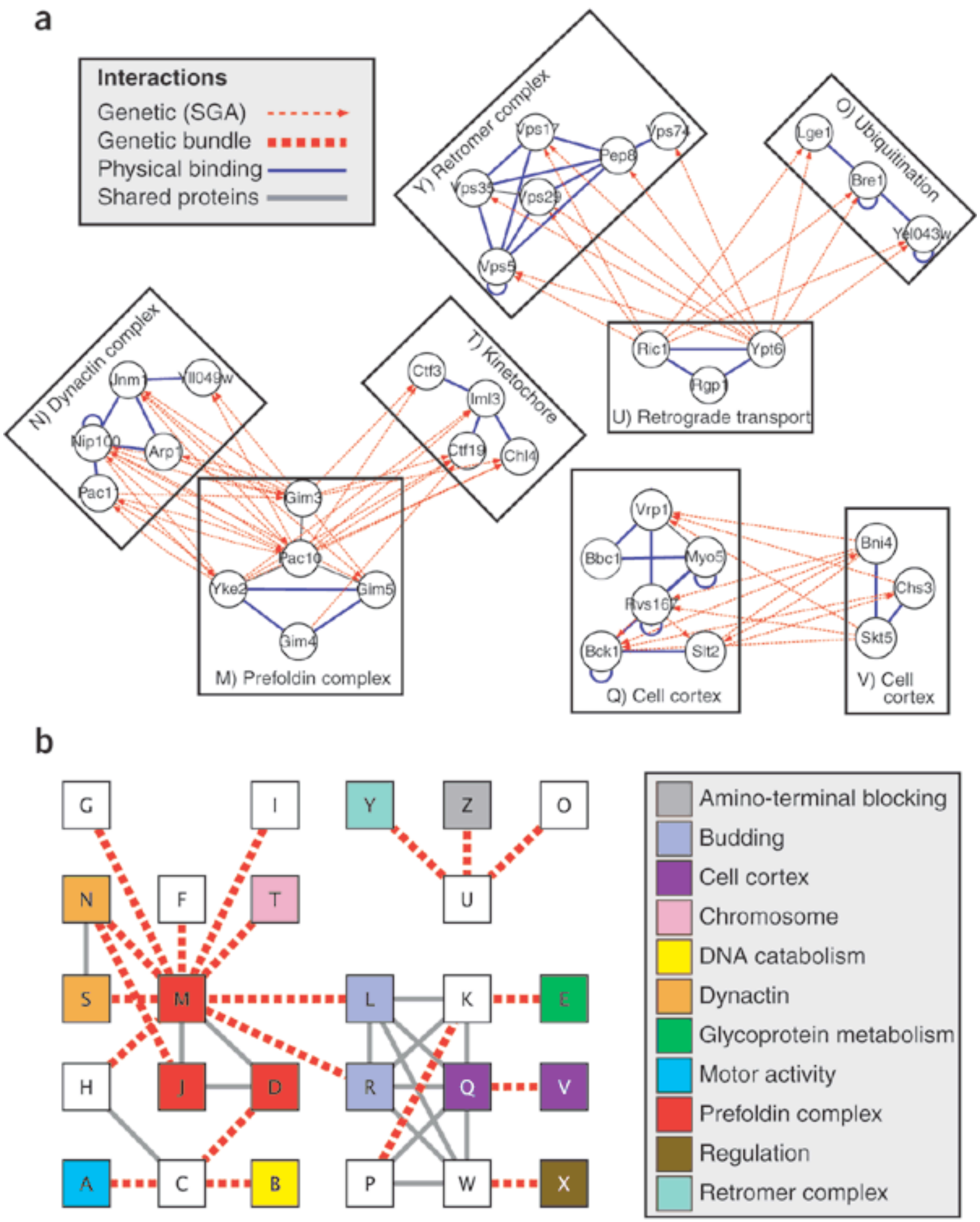


Figure 2.2. Between-pathway explanations for genetic interactions.

(a) Several high-scoring models are shown (M,N,T; Q,V; O,U,Y). Blue solid and red dotted links indicate physical and genetic interactions, respectively. (b) Bird's-eye view of all between-pathway models obtained from a search on a reduced network composed of SGA and DIP interactions. Each node [A]−[Z] represents a physical pathway; groups of genetic interactions between pathways are condensed into a single link called a 'bundle.' Node colors indicate significant Gene Ontology annotations. Solid gray lines connect pathways that share one or more proteins; such pathways may represent different components of a larger mechanism.

Pooling diverse genetic and physical interaction data sets widens the search but also has the potential to decrease the coverage of network models, because not all data sets may be equally predictive and high-scoring network models are more likely to arise at random in large networks. To investigate the effect of data pooling, we repeated the search on a reduced network including large-scale synthetic-lethal (SGA) and protein-binding (DIP) interactions only. This reduced search identified 20 models containing a total of 137 synthetic-lethal and 120 protein-binding interactions (Figure 2.2). In comparison to the complete search, fewer protein-binding and SGA synthetic-lethal interactions were incorporated into models, demonstrating the synergy obtained by data pooling (although models generated by the restricted search performed somewhat better in validation). Supplemental Table 2.2 analyzes the impact of removing each physical and genetic data set from the modeling procedure.

Within-pathway interpretations.

We next searched the physical and genetic networks for within-pathway explanations. As described in Methods, this procedure assigned a high score to single sets of proteins that were densely connected by both physical and genetic interactions (see Fig. 1). This search yielded 91 significant models. In all, these contained 272 MIPS, 225 SGA synthetic lethal, and 169 SGA synthetic-sick interactions associated with 318 protein binding, 37 regulatory, and 36 shared-reaction interactions. Four representative within-pathway models are shown in Figure 2.3.

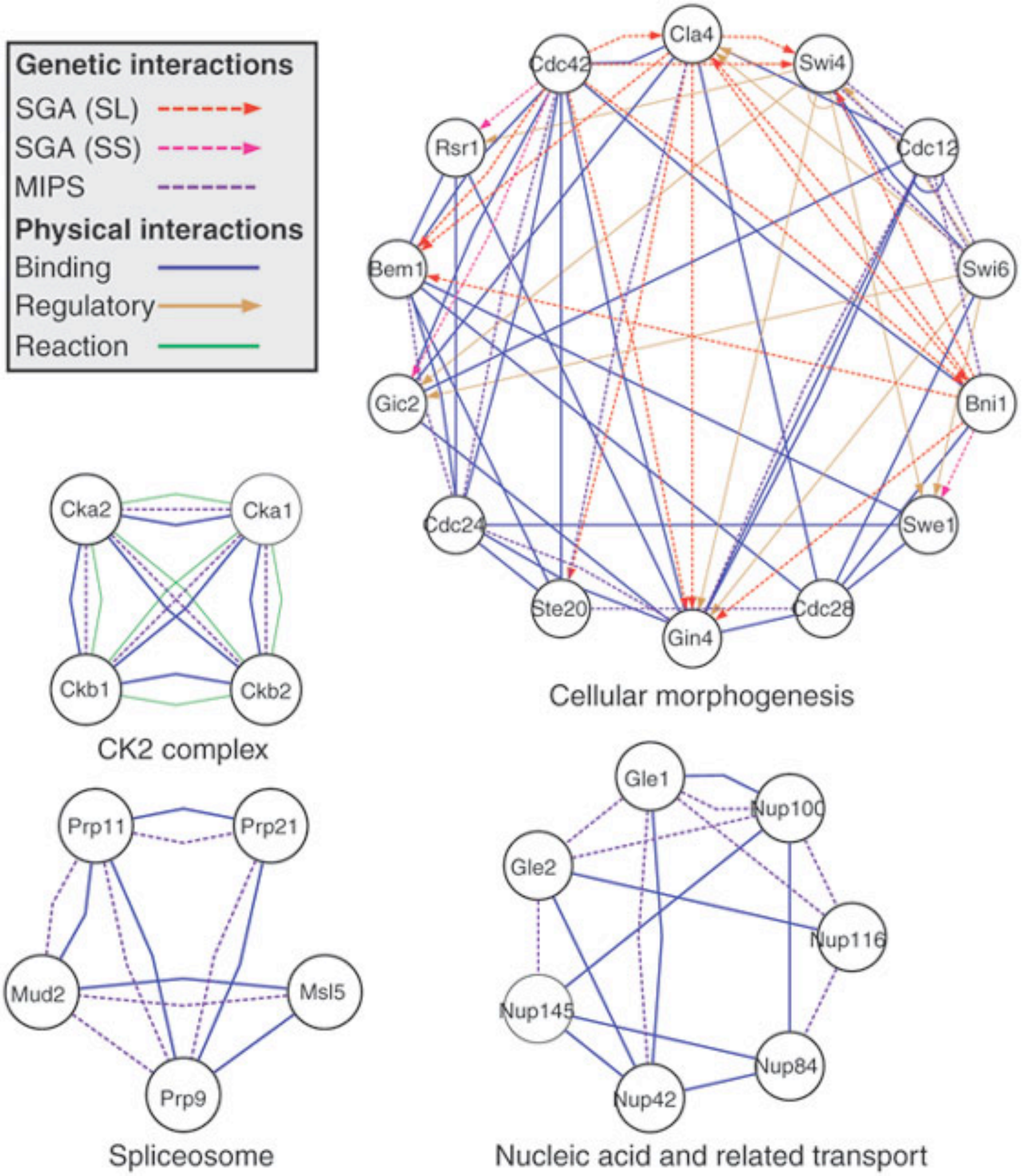


Figure 2.3. Within-pathway explanations for genetic interactions.

A total of 91 pathways were identified, of which four examples are displayed. Color is used to indicate the data set from which each interaction was drawn.

Functional enrichment of models.

As initial validation of the between- and within-pathway models, we found that both types were significantly enriched for particular functional annotations recorded in the Gene Ontology database[40]. Two-hundred and fifty-one out of 401 pathways in between-pathway models were enriched for proteins with a common Molecular Function, Biological Process, or Cellular Component annotation using the hypergeometric test (P = 0.05; Bonferroni-corrected for multiple testing)[41]. Similarly, 52 of the 91 within-pathway models were enriched for Gene Ontology annotations. Moreover, these functional enrichments were higher than expected based on the physical interaction network as a whole (see Supplemental Table 2.3).

Prediction of new protein functions.

Having established that proteins in many of the between- and within-pathway models were enriched for specific annotations, we used this concept to predict new protein functions. Specifically, for physical pathways in which a majority of proteins were already assigned a common significant. For between-pathway models, this approach predicted 745 Molecular Function, Biological Process, or Cellular Component annotations among 282 proteins. In comparison, the within-pathway models predicted 285 annotations involving 127 proteins, bringing the total to 973 annotations for 343 proteins accounting for repeated predictions. A list of novel functional predictions is provided at http://www.cellcircuits.org/Kelley2005. Less than a quarter of these predictions were attainable using a similar approach based on the physical network only (Supplemental Table 2.4).

Accuracy of these predictions was estimated using cross-validation[42]. Using a standard five-way procedure, the set of yeast proteins was partitioned such that annotations were hidden for one-fifth of the proteins and annotations for the remaining four-fifths of proteins were used to predict the hidden information. Each prediction for a protein in the "hidden set" was scored as a success or failure depending on whether it recovered a hidden annotation. Using this approach, the success rate was estimated to be 63% for between-pathway models, 69% for within-pathway models.

Prediction of new genetic interactions.

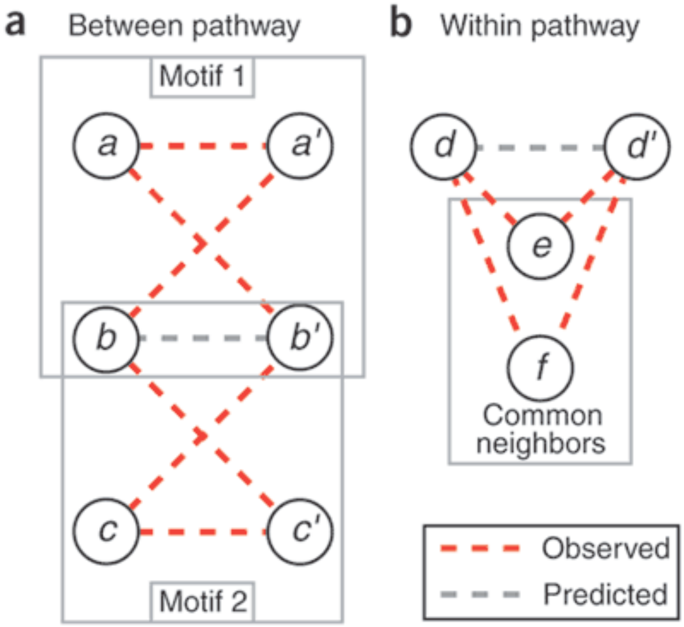


Figure 2.4. Genetic interaction prediction schemes.

Two different schemes are proposed for predicting genetic interactions, depending on the underlying network model. Observed genetic interactions are shown in red, while the corresponding predicted genetic interactions are shown in gray. (a) Under the between-pathway model, two incomplete bipartite motifs are shown which predict a genetic interaction between genes b and b'. (b) Under the within-pathway model, common genetic neighbors are used to predict a genetic interaction between genes d and d'. Note that these diagrams contain additional incomplete motifs which have been omitted for clarity: the motifs in a can be rearranged to predict genetic interactions (a to c') and (c to a'); the motifs in b can be rearranged to predict (e to f).

Finally, we investigated whether the network models could predict the existence of new genetic interactions. According to the between-pathway model, proteins in a first pathway genetically interact with many of the same partners in a second pathway. This leads to the occurrence of "complete bipartite motifs" in the genetic interaction network, defined as four-protein subnetworks in which the first two proteins are connected to the second two proteins by all four possible genetic interactions (Figure 2.4a; see Milo et al.[43] for an introduction to network motifs). When an incomplete motif (IM) is observed, for which only three of the four genetic interactions are present, the motif implies that the remaining interaction is true. Physical network information is incorporated by requiring that valid IMs fall within (i.e., are subgraphs of) a between-pathway model.

We applied the technique of five-way cross-validation to estimate the accuracy of genetic interaction prediction versus the minimum number of required IMs (Figure 2.5). In each of five cross-validation trials, approximately one-fifth of the genetic interaction data were withheld, including both positive and negative interactions measured for each genetic "bait" in SGA. These positive and negative interactions were subsequently used to test prediction accuracy. For instance, at a prediction threshold of eight or more IMs, the between-pathway models predicted 43 new genetic interactions with 87% estimated accuracy (Figure 2.5). To assess the contribution of the physical models in the prediction process, we also predicted "naïve" genetic interactions by relaxing the requirement that IMs fall in a between-pathway model. The estimated accuracy fell to 5% for these naïve predictions, evaluated at the same threshold of eight IMs.

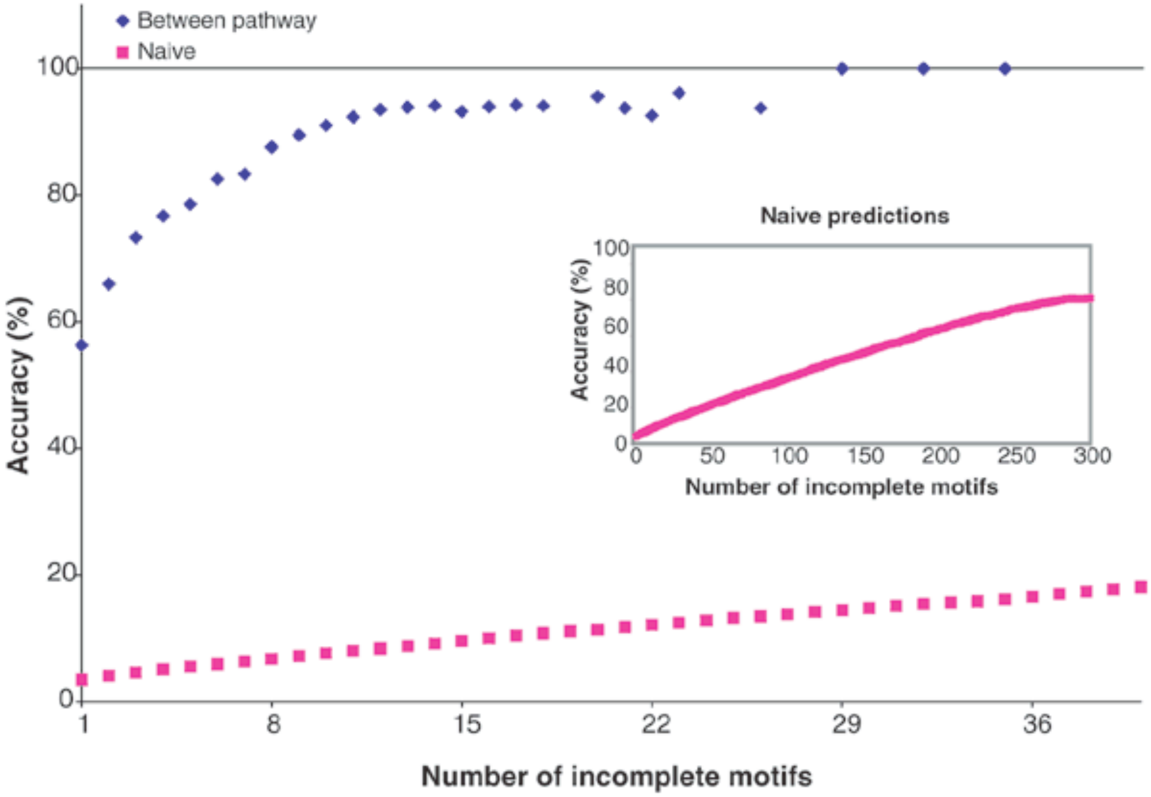


Figure 2.5. Success rate of genetic interaction prediction versus the stringency of prediction

Success rate is measured through cross-validation as (predicted positives)/(predicted positives and negatives). Stringency is defined by the minimum number of incomplete bipartite motifs required for prediction. Blue diamonds mark the success rate for predictions in which incomplete motifs must occur in a between-pathway model. The success rate is dramatically higher than for naive predictions (magenta) which predict interactions in the same manner, but are not constrained by the physical network. Even for much more stringent prediction criteria, the success rate of naive predictions fails to exceed that of the between-pathway predictions (inset)..

For the within-pathway models, genetic interactions were implied between proteins that had genetic interactions with one or more common neighbors (Figure 2.4b). The physical network was incorporated by restricting the proteins and neighbors to fall into a single within-pathway model. The number of common neighbors was used as a measure of confidence in the implied genetic interaction, and cross validation was used to estimate the prediction accuracy as a function of this number. The maximal prediction accuracy was 38%, achieved at a prediction threshold of three or more common neighbors (Supplemental Figure 2.3). The corresponding success rate for naïve predictions, made without constraining the proteins to occur in within-pathway models, was 15%. Thus, both types of models enhance the accuracy of prediction of genetic interactions, but between-pathway models appear to be better predictors than within-pathway models.

Discussion

Given a systematic approach for associating genetic interactions with physical interpretations, it is of interest to ask which type of interpretation is most common. Focusing on large-scale SGA measurements, roughly three and a half times as many genetic interactions are associated with between- as opposed to within-pathway models (1,377 vs. 394 SGA interactions). These figures can be viewed as an a priori expectation that a newly-determined SGA interaction will fall between vs. within pathways, suggesting that SGA interactions typically span between multiple physical network regions instead of occurring within a single complex or pathway. One reason for the preference towards between-pathway models may be that SGA interactions are mainly targeted to non-essential genes (due to their use of complete gene deletions as opposed to, e.g., point mutations made by classical techniques).

Using physical models, it is possible to characterize approximately 40% of the genetic interactions as occurring between or within pathways. Whether the remaining interactions belong to between-pathway models, within-pathway models, or are best characterized as “indirect” (Table 1) cannot be reliably determined at this stage. For example, consider the case of two related pathways, each with only one protein required for pathway function. In this case, only the required proteins would be connected by a (single) genetic interaction across the pathways, making it difficult for the between-pathway model to achieve statistical significance.

Further examination of the between-pathway models reveals that many of the genetically-linked pathways have clear interdependent functional relationships. For example, pathway M shown in Figure 2a contains members of the prefoldin complex, which have synthetic-lethal interactions with members of pathways N and T forming parts of the dynactin complex and kinetochore, respectively. The prefoldin complex promotes folding of α and β tubulin into functional microtubules[44]. These are important for the function of dynactin, an adaptor complex involved in translocating the spindle and other molecular cargos along microtubules[45], as well as the kinetochore, which anchors chromosomes to spindle microtubules during metaphase[46]. Apparently, deletion of proteins in the prefoldin complex reduces microtubule stability which leads to synthetic-lethal interactions with pathways that are directly dependent on microtubule function.

These pathways also predict a new function for the uncharacterized protein Yll049w (pathway N). This protein binds Jnm1, a dynactin protein which is required for spindle partitioning in anaphase[45]. In addition, it has synthetic-lethal interactions with members of the prefoldin complex in a manner similar to dynactin genes. Together, these relationships suggest that Yll049w is associated with dynactin during spindle partitioning. However, because Jnm1 has 12 physical interactions overall, and Yll049w has a total of 14 interactions in the genetic network, this prediction would have been difficult to make without an integrated approach.

Pathways O, U, and Y provide another example of synergistic pathways linked by genetic interactions (Fig. 2a). Pathways U and Y mediate retrograde transport of proteins to the Golgi apparatus[47, 48]. Pathway O (Bre1, Lge1) is involved in histone ubiquitination and cell size control, where cell size is influenced by the histone ubiquitination activity via an unknown process[49]. The abundant genetic interactions between pathways O and U indicate a possible role for retrograde transport in histone ubiquitination, or reciprocally, for histone ubiquitination in retrograde transport. Moreover, the uncharacterized protein Yel043w is physically associated with Bre1 and Lge1 and also has the same pattern of genetic interactions, suggesting that the three proteins may function together.

In summary, we have presented a methodology for integrating large-scale genetic and physical networks to capture the physical context behind observed genetic interactions. Approximately 40% of yeast synthetic-lethal genetic interactions can be incorporated into high-level physical pathway models and are approximately three and a half times as likely to span pairs of pathways than to occur within pathways. Further studies will be needed to address other types of genetic effects to extend this approach from yeast to the growing number of other organisms for which protein networks are now available. As systematic approaches generate ever larger databases of interactions across a variety of species, integrative modeling approaches such as the one proposed here will be indispensable for selecting and organizing the information into predictive models.

Methods

Scoring within-pathway explanations.

The within-pathway model implies dense interactions within a single group of proteins in both the physical and genetic networks. We adopt a previously described log-odds score[50] to assess the likelihood that a group of proteins is more densely connected than would be expected at random:

$$
\begin{aligned}
S_{within}(V,E) &= \log\frac{P(V,E\mid Model_{dense})}{P(V,E\mid Model_{random})} \\
&= \frac{\prod_{(a,b)\in V\times V}\beta I_E(a,b)+(1-\beta)(1-I_E(a,b))}{\prod_{(a,b)\in V\times V} r_{a,b}I_E(a,b)+(1-r_{a,b})(1-I_E(a,b))}
\end{aligned}
$$

where V is a set of proteins and E a set of interactions among those proteins (genetic or physical). $I_E(a,b)$ is an indicator function which equals 1 if and only if the interaction (a,b) occurs in E and otherwise 0. For $Model_{dense}$, interactions are expected to occur with high probability (β) for every pair of proteins in V. In this work, β is set to 0.9 (Supplemental Figure 2.2 shows how the results depend on choice of β). For $Model_{random}$, the probability of observing each interaction ($r_{a,b}$) is determined by estimating the fraction of all networks with identical degree distribution which also contain that interaction. Comparable random networks are generated by "crossing" pairs of edges in a process similar to that described by Milo et al.[43] In this randomization, only edges of the same type are allowed to be crossed. In addition, for undirected types, either interacting node is allowed to serve as the "source" in crossing the edges. Such randomization generates a family of random networks which resemble the original network and corrects for the presence of highly-connected proteins in the network, which score highly under both models. The interaction density is evaluated independently for the physical and genetic networks, yielding an overall score for the within-pathway model.

$$S = S_{within}(V, E_{physical}) + S_{within}(V, E_{genetic})$$

Scoring between-pathway explanations.

The between-pathway model implies dense genetic interactions connecting two separate, non-overlapping groups of proteins, where each group is densely connected by physical interactions. The density of physical interactions is scored independently within two sets of proteins V_1 and V_2 using the above function S. A related log-odds score is used to evaluate the probability that the genetic interactions $E_{genetic}$ bridging between these sets are denser than random:

$$S_{within}(V_1, V_2, E) = \frac{\prod_{(a,b)\in V_1 \times V_2} \beta I_{E_{genetic}}(a,b) + (1-\beta)(1 - I_{E\,genetic}(a,b))}{\prod_{(a,b)\in V_1 \times V_2} r_{a,b} I_{E_{genetic}}(a,b) + (1 - r_{a,b})(1 - I_{E_{genetic}}(a,b))}$$

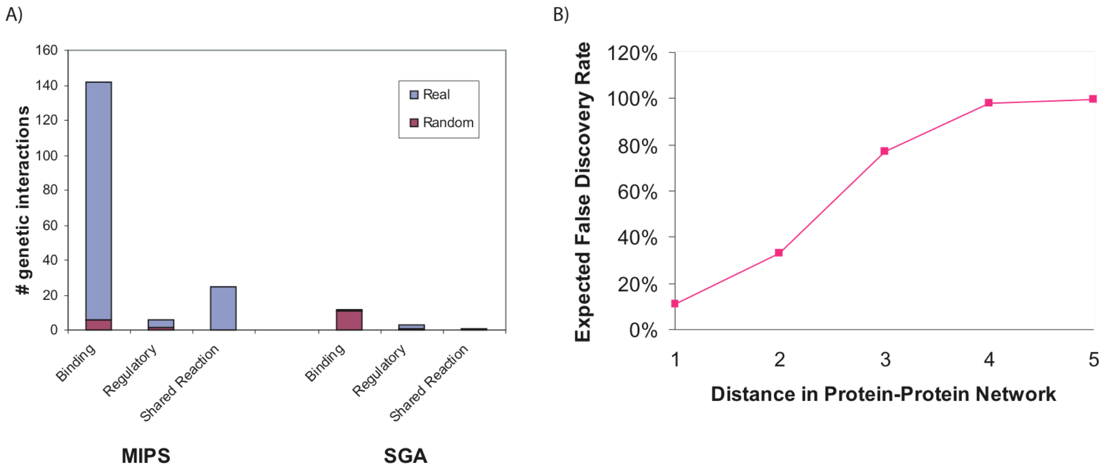
The final scoring function for the between-pathway model is then:

$$S(V_1, V_2, E_{all}) = \sum_{i=1,2} S_{within}(V_i, E_{physical}) + S_{between}(V_1, V_2, E_{genetic})$$

Search and Significance.

Sets of proteins that are well explained by either the within-pathway or between-pathway models are identified using a greedy network search procedure. The search is as previously described by Sharan et al.[50] except that it is seeded from each pair of genetically-interacting proteins. Pathways that share more than 50% of genetic interactions with a higher-scoring result are discarded. To determine the significance threshold, identical searches are performed over 100 random trials in which both the genetic and physical networks are randomized as described above. Models that score higher than the maximal-scoring models in 95% of random trials are reported as significant.

Supplemental Figures



Supplemental Figure 2.1. Direct overlaps between genetic and physical interactions, while statistically significant, are limited in systematic data and probably biased.

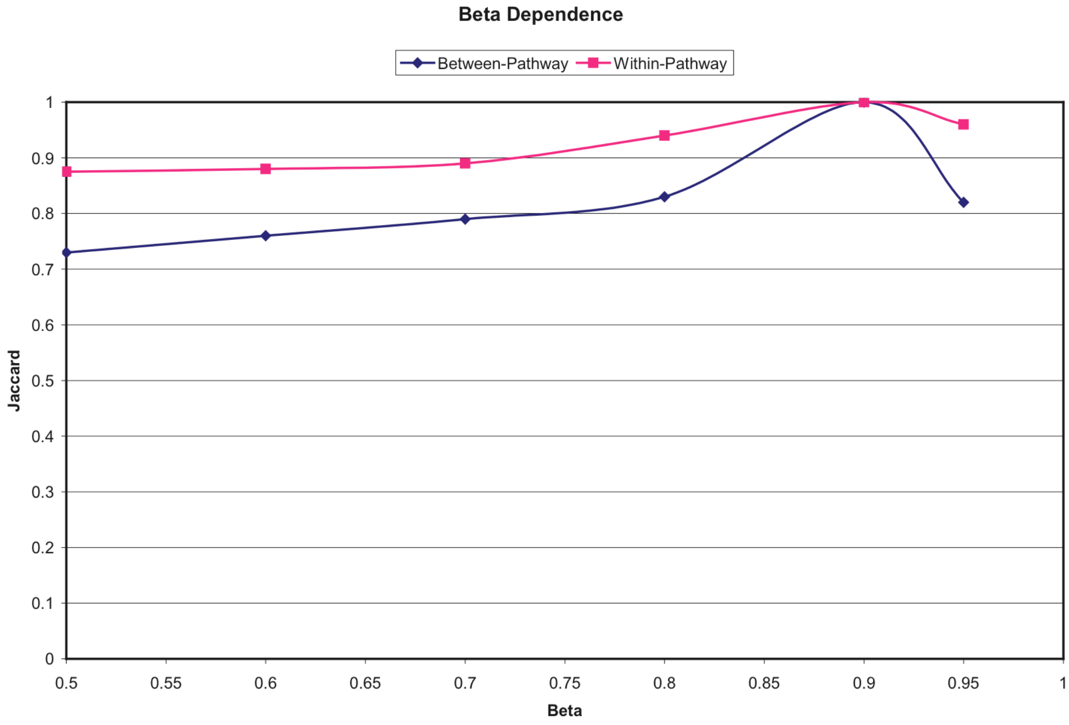
As a preliminary assessment of whether synthetic-lethal genetic interactions could be explained by physical interactions, we investigated whether proteins connected by genetic interactions were also at close proximity in the physical network. As shown in Figure S1 [a], genetic interactions were co-incident with a total of 189 physical interactions (154 protein binding, 9 regulatory, 26 metabolic). These counts were significant in comparison to randomized genetic networks (yielding 19.2 +/- 6.9 overlapping physical interactions; mean +/- stdev). In the figure, results are tabulated separately for two types of genetic interactions (MIPS, SGA) and three types of physical network (protein-protein binding, protein-DNA regulatory, and shared-reaction metabolic).

However, further investigation suggested that much of this overlap might be due to bias in determination of the physical or genetic network. First, 93% (176/189) of the coincident genetic interactions were derived from small-scale studies curated by MIPS. This percentages was highly enriched (p¡4x10ˆ-65) compared to the relatively small percentage (26%) of genetic data measured in small-scale studies overall. Similarly, 87% (164/189) of the coincident physical interactions were identified in small-scale studies (as recorded by the DIP database).

Thus, the coincident physical interactions are biased towards small-scale studies, probably because physical interactions are sometimes tested explicitly as a follow-up to observing a genetic interaction. Direct correspondence between systematic genetic and physical interactions is much weaker (e.g., between SGA and protein-binding interactions in DIP)

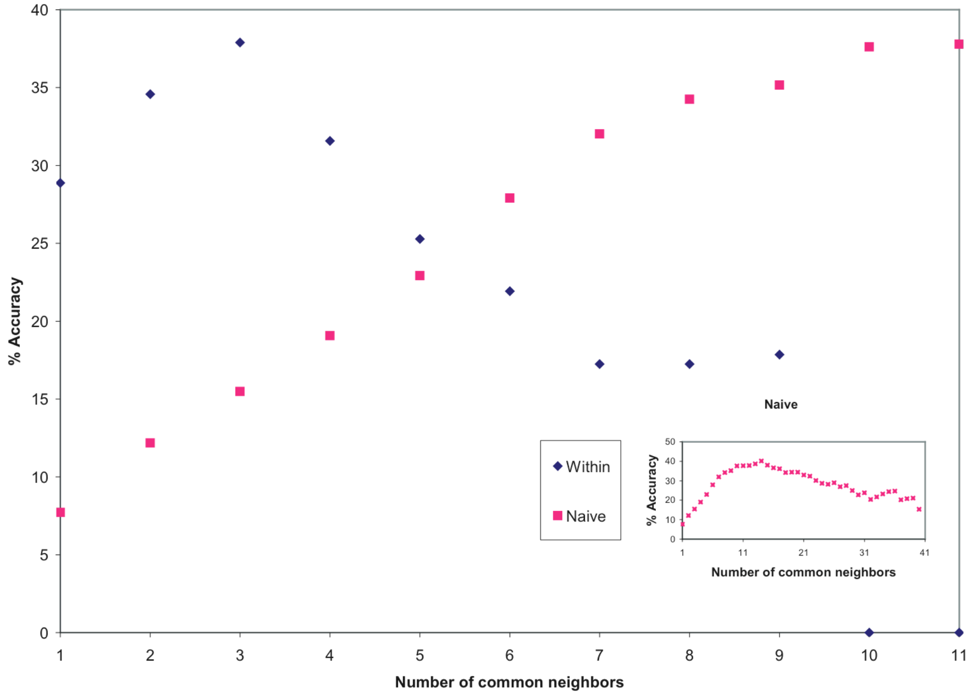
Such conclusions hold even after extending the analysis from direct interactions to longer paths. As shown in Figure S1 [b], for each pair of synthetic-lethal proteins we recorded the length of the shortest path connecting these proteins in the protein-protein network. The False Discovery Rate of genetic/physical overlap is shown for genetic interactions connected by direct (length 1) or longer paths of protein-protein interactions (lengths 2-6). False Discovery Rate (FDR) is the expected percentage of these relationships that are spurious based on randomized networks (described further in the Methods). Genetic interactions in MIPS match a greater number of short paths than would be expected at random, while the number explained by physical paths of length > 3 (SGA: all lengths) is essentially no different than for random networks.

Thus, the number of synthetic-lethal pairs connected by paths of up to three protein-protein interactions was larger than expected. However, this trend was too weak to be used in identifying the physical cause of any partical synthetic-lethal interaction: Even limiting consideration to paths of length two, nearly a third of the paths were likely due to random chance.



Supplemental Figure 2.2. Influence of beta on result set.

This figure compares the set of proteins included in significant network models for different values of beta versus a beta of 0.9. The similarity of the two result sets is summarized by the jaccard (intersection/union). From these results, we see that the between-pathway models are more sensitive to changes in the value of beta. However, even for a very small value of beta, the results are still largely overlapping. We chose a beta of 0.9 (which tends to create a smaller result set) to enhance the stringency of our results.



Supplemental Figure 2.3. Estimate prediction accuracy for naive and pathway-based within-pathway genetic predictions.

This graph displays the estimated accuracy of pathway-based and naive within-pathway genetic predictions. Within-pathway predictions were made by predicting genetic interactions between genes with common genetic interaction neighbors. The physical network was incorporated in the pathway-based predictions by restricting the proteins and neighbors to fall into a single within-pathway model. The number of common neighbors is therefore used as a measure of confidence in the implied genetic interactions. The maximum prediction accuracy obtained for within-pathway based predictions is ≈ 38% using a prediction threshold of three common neighbors. The inset displays the estimated prediction accuracy over an expanded range for just the naive predictions.

Supplemental Tables

Supplemental Table 2.1. Compounds excluded from the physical interaction network.

Kegg Compound ID	Description
C00001	H2O
C00002	ATP
C00008	ADP
C00009	orthophosphate
C00013	pyrophosphate
C00003	NAD
C00004	NADH
C00005	NADPH
C00006	NADP
C00010	CoA
C00011	CO2
C00020	AMP
C00025	Glutamate
C00014	Ammonia
C00024	Acetyl CoA

Supplemental Table 2.2. Results from reduced searches.

This table displays the results from a number of searches run in reduced networks. The column labeled ”Interaction Set” identifies the specific reduced network, where Interaction Type(-) indicates that type of interaction was removed to create the reduced network. ”Average GI Predictions (Cross Validation)” refers to the expected number of genetic predictons made using this reduced network in cross-validation.”GI Prediction Accuracy” refers to the accuracy of these predictions. In cases where the cross-validation was unable to generate predictions of the required stringency, this result in not available (n/a). In addition, GO prediction accuracy was also assessed. This result is displayed in the final column. With respect to the set of physical interactions, protein-protein interactions are crucial for model performance. For reduced searches in the genetic network, both synthetic sick and synthetic lethal interactions enhance model performance. Note that the ”PP+SL” models perform particularly well in both genetic and GO prediction.

Between-Pathway			
Interaction Set	Average GI Predictions (Cross Validation)	GI Prediction Accuracy	GO prediction accuracy
Complete	117.6	88%	63%
Regulatory(-)	110.2	90%	48%
Reaction(-)	122.6	89%	53%
Binding(-)	0	n/a	100%
SS(-)	7	100%	33%
SL(-)	1.2	83%	25%
Binding+SL only	21.2	99%	78%
Within-Pathway			
Interaction Set	Average GI Predictions (Cross Validation)	GI Prediction Accuracy	GO prediction accuracy
Complete	108.2	38%	69%
Regulatory(-)	66.6	32%	72%
Reaction(-)	113.6	35%	68%
Binding(-)	2.6	0%	92%
SS(-)	57.6	49%	65%
SL(-)	1.4	0%	50%

Supplemental Table 2.3. Functional Enrichment.

For each protein in a between-pathway model, we computed the average GO similarity of this protein versus two different sets of interacting neighbors: (1) all of its neighbors in the protein-protein interaction network; and (2) the specific subset of these neighbors arising in the same model. Out of 712 proteins in between-pathway models, 563 (79%; p<0.01) had a higher average similarity with the interacting proteins in their model than with all neighbors in the network as a whole. In within-pathway models, 233 of 298 proteins (78%; p<0.01) showed similar functional enrichment. These trends apply whether the network neighborhood is defined to include only proteins with direct interactions (as reported above) or also those reachable at ”distance 2” through an intermediate protein (yielding 91% and 97% for between- and within-pathways models). Here, functional similarity between two proteins a and b was defined as inversely proportional to the number of proteins covered by the most specific GO term covering a and b.

	Physical Distance	
Search	1	2
Between-Pathway	79%	91%
Within-Pathway	78%	97%

Supplemental Table 2.4. Basis of annotation predictions.

Using the between-pathway models, 745 annotations were predicted with an estimated accuracy of 63%. Conversely, 285 annotations were predicted using the within-pathway models with an estimated accuracy of 69%. Are these functional predictions based mainly on physical interactions, or do they require the genetic network also? To address this question, we looked for dense subnetworks of interactions in the physical network separately (analogous to the within-pathway search but scoring only one interaction type - see Methods). Of the above 745 between-pathway annotations, 194 were also predicted from significant physical pathways alone. In the case of the within-pathway search, there are only 29 such predictions. The remaining predictions rely, at least in part, on genetic evidence for support. Considering these remaining predictions only, cross-validation accuracy was 50% and 59% for between- vs. within-pathway models. In this table, the ”Model” column delineates which type of model was used to generate the prediction. The ”Search” column tells whether the annotations were predicted only from the combined physical/genetic search. The ”Count” and ”Accuracy” columns give these values for the various sets of annotations.

Model	Search	Count	Accuracy
Between-Pathway	Combined	745	63%
	Combined Only	551	50%
Within-Pathway	Combined	285	69%
	Combined Only	256	59%

Acknowledgements

We thank Jonathan Wang, Owen Ozier, and Gopal Ramachandran for preliminary investigations and Vineet Bafna, Ben Raphael, and Vikas Bansal for insightful commentary. Craig Mak, Silpa Suthram, and Taylor Sittler provided helpful reviews of the text. Funding was provided by the National Institute of General Medical Sciences (GM070743-01) and the National Science Foundation (NSF 0425926).

Chapter 2, in full, is a reprint of the following work,

Kelley R, Ideker T. *Systematic interpretation of genetic interactions using protein networks*. **Nature Biotechnology** 2005; 23(5):561-6.

The dissertation author was the sole first author on this paper, responsible for designing, implementing, and running computational algorithms.

Chapter 3. Functional maps of protein complexes from quantitative genetic interaction data

Abstract

Recently, a number of advanced screening technologies have allowed for the comprehensive quantification of aggravating and alleviating genetic interactions among gene pairs. In parallel, TAP-MS studies (Tandem Affinity Purification followed by Mass Spectroscopy) have been successful at identifying physical protein interactions which can indicate proteins participating in the same molecular complex. Here, we propose a method for the joint learning of protein complexes and their functional relationships by integration of quantitative genetic interactions and TAP-MS data. Using three independent benchmark datasets, we demonstrate that this method is >50% more accurate at identifying functionally related protein pairs than previous approaches. Application to genes involved in yeast chromosome organization identifies a functional map of 91 multimeric complexes, a number of which are novel or have been substantially expanded by addition of new subunits. Interestingly, we find that complexes that are enriched for aggravating genetic interactions (i.e., synthetic lethality) are more likely to contain essential genes, linking each of these interactions to an underlying mechanism. These results demonstrate the importance of both large-scale genetic and physical interaction data in mapping pathway architecture and function.

Introduction

Genetic interactions are logical relationships between genes that occur when mutating two or more genes in combination produces an unexpected phenotype[15, 51, 52]. Recently, rapid screening of genetic interactions has become feasible using Synthetic Genetic Arrays (SGA) or diploid Synthetic Lethality Analysis by Microarray (dSLAM)[22, 23]. SGA pairs a gene deletion of interest against a deletion to every other gene in the genome (in turn). The growth / no growth phenotype measured over all pairings defines a *genetic interaction profile* for that gene, with no growth indicating a synthetic-lethal genetic interaction. Alternatively, all combinations of double deletions can be analyzed among a functionally-related group of genes[9, 53, 54]. A recent variant of SGA termed E-MAP[54] has made it possible to measure continuous rates of growth with varying degrees of epistasis (based on imaging of colony sizes). "Aggravating" interactions are indicated if the growth rate of the double gene deletion is slower than expected, while for "alleviating" interactions the opposite is true[55, 56].

One popular method to analyze genetic interaction data has been to hierarchically cluster genes using the distance between their genetic interaction profiles. Clusters of genes with similar profiles are manually searched to identify the known pathways and complexes they contain as well as any genetic interactions between these complexes. This approach has been applied to several large-scale genetic interaction screens in yeast including genes involved in the secretory pathway[9] and chromosome organization[53]. Segré et al.[57] extended basic hierarchical clustering with the concept of "monochromaticity", in which genes were merged into the same cluster based on minimizing the number of interactions with other clusters that do not share the same classification (aggravating or alleviating).

Another set of methods has sought to interpret genetic relationships using physical protein-protein interactions[58]. Among these, Kelley and Ideker[59] used physical interactions to identify both "within-module" and "between-module" explanations for genetic interactions. In both cases, modules were detected as clusters of proteins that physically interact with each other more often than expected by chance. The "within-module" model predicts that these clusters directly overlap with clusters of genetic interactions. The "between-module" model predicts that genetic interactions run between two physical clusters that are functionally related. This approach was improved by Ulitsky *et al.*[60] using a relaxed definition of physical modules. In related work, Zhang et al.[61] screened known complexes annotated by the Munich Information Center for Protein Sequences (MIPS) to identify pairs of complexes with dense genetic interactions between them.

One concern with the above approaches, and the works by Kelley and Ulitsky in particular, is that they make assumptions about the density of interactions within and between modules which have not been justified biologically. Ideally, such parameters should be learned directly from the data. Second, between-module relationships are identified by separate, independent searches of the network seeded from each genetic interaction. This "local" search strategy can lead to a set of modules that are highly overlapping or even completely redundant with one another. Finally, genetic interactions are assumed to be binary growth / no growth events while E-MAP technology has now made it possible to measure continuous values of genetic interaction with varying degrees of epistasis. Here, we present a new approach for integrating quantitative genetic and physical interaction data which addresses several of these shortcomings. Interactions are analyzed to infer a set of modules and a set of inter-module links, in which a module represents a protein complex with a coherent cellular function, and inter-module links capture functional relationships between modules which can vary quantitatively in strength and sign. Our approach is supervised, in that the appropriate pattern of physical and genetic interactions is not predetermined but learned from examples of known complexes. Rather than identify each module in independent searches, all modules are identified simultaneously within a single unified map of modules and inter-module functional relationships. We show that this method outperforms a number of alternative approaches and that, when applied to analyze a recent EMAP study of yeast chromosome function, it identifies numerous new protein complexes and protein functional relationships.

Results

Characterization of Genetic and Physical Networks.

We first sought to quantitatively confirm whether, and to what degree, physical and genetic interactions could indicate common membership in a protein complex. To provide genetic data for analysis, we obtained the previously-published results from a large E-MAP of yeast chromosomal biology[53]. This study consisted of genetic interactions measured among 743 genes (including 74 essential genes), yielding quantitative values for 182,669 gene pairs (66% of all possible pair-wise measurements). Each gene pair was assigned an S-score, where positive scores indicate protein pairs for which the double mutant grows better than expected (i.e., an alleviating interaction) and negative scores indicate pairs for which the double mutant grows worse than expected (i.e., a synthetic sick/lethal or aggravating interaction) where the expectation is that the double-deletion of unrelated proteins will have a growth rate equivalent to the multiplicative product of the two individual growth rates[62]. In all, 14,237 gene pairs (8%) showed strong genetic interactions with |S| > 2.5. Physical interactions were taken from a recent computational integration of two large datasets measured by co-immunoprecipitation followed by mass spectrometry[63]. This study assigned to each pairwise interaction a Purification Enrichment (PE) score, with larger values representing a greater likelihood of true binding.

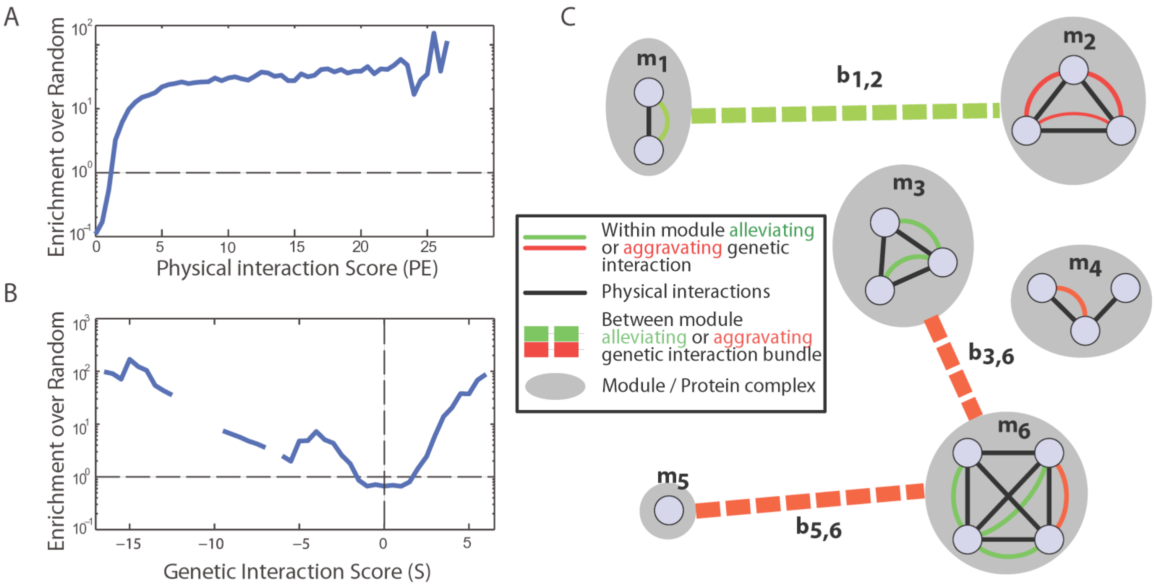


Figure 3.1. Combining physical and genetic interactions to define protein complexes.

Correspondence of the physical interaction score (A) and the genetic interaction score (B) with the known small-scale, manually annotated protein complexes in MIPS. To compute the enrichment over random (y-axis), one first computes the fraction f of interactions at each score x that fall inside a MIPS small-scale complex (bin size of 1.5). The enrichment is the ratio f/r, where r is the fraction of random protein pairs within MIPS complexes. (C) Proteins are grouped into physically interacting sets called modules (gray ovals; m_1–m_6). Pairs of modules may be linked to indicate a functional relationship (dotted lines; b_1–b_6). The assignment of proteins to modules along with the list of inter-module links comprises the state of the system.

Figure 3.1a confirms that protein pairs with higher PE-scores are more likely to operate in a known small-scale protein complex recorded in the MIPS database[64] (versus protein pairs chosen at random). This result is expected considering that PE-scores were trained based on these complexes[63]. Figure 3.1b shows that protein pairs with both positive and negative S-scores are more likely to operate within a known complex. Positive (alleviating) interactions are well-known to occur between subunits of a complex[53]. Negative (aggravating) interactions are to a lesser degree so, although the mechanism has not been as clear as for the alleviating case[65]. By comparing the magnitudes of enrichment between Figure 3.1a and b, it is apparent that extreme S-scores are at least as indicative of co-complex membership as strong PE-scores, if not more so (~100-fold enrichment versus ~50-fold enrichment, respectively). Together, these exploratory findings suggest that the physical and genetic information can indeed provide a basis for the identification of protein pairs involved in the same complex.

Functional maps of protein complexes involved in yeast chromosomal biology.

To capture these trends, we formulated an approach to classify a protein pair as operating either within the same module or between functionally-related modules given its genetic and physical interaction scores. This approach seeks to categorize interactions supported by both strong genetic and physical evidence as operating within a module (i.e., complex). Interactions with a strong genetic but weak physical signal are better characterized as operating between two functionally-related modules. Given within-module and between-module likelihoods for individual interactions, an agglomerative clustering procedure seeks to merge these interactions into increasingly larger modules and to identify pairs of modules interconnected by bundles of many strong genetic interactions (Figure 3.1c). Full details are provided in the Methods.

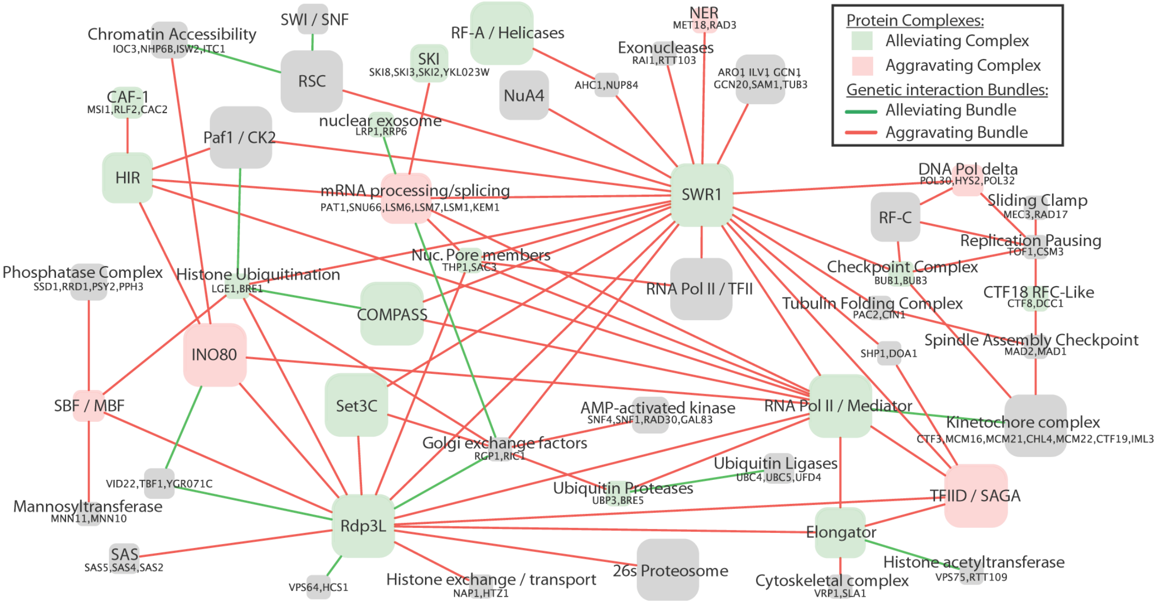


Figure 3.2. Global map of protein complexes involved in yeast chromosome biology.

Each node represents a predicted multimeric protein complex, while each link represents a significantly alleviating or aggravating bundle of genetic interactions between complexes, indicative of an inter-complex functional relationship. Node colors indicate enrichment for alleviating or aggravating genetic interactions among members of the same complex. Node sizes are proportional to the number of proteins in the complex. When known, nodes are labeled with the common name of the complex. For complexes that are newly identified by our study and thus unnamed, the constituent proteins are listed. For clarity, the co-chaperone prefoldin complex (PFD1, PAC10, YKE2, GIM3, GIM4, GIM5, BUD27) and the 25 links associated with it have been removed.

Applying this method, we identified 91 distinct modules with an average size of 4.1 proteins per module. Figure 3.2 gives an overview of a subset of the identified modules and inter-module links. Complete results are catalogued at http://www.cellcircuits.org/Bandyopadhyay2008/html/. Overall, these results suggest ten novel complexes not recorded in either the small-scale or high-throughput MIPS compendium, covering 23 proteins in total. The results also identify 84 new subunits of known complexes (Supplemental Materials). Through permutation testing, 19 versus 9 of the identified modules could be categorized as enriched for alleviating or aggravating genetic interactions, respectively. A total of 313 significant genetic relationships were identified between modules, 94 versus 219 of which were enriched for alleviating or aggravating interactions.

Comparison to related approaches.

The method of choice for interpreting quantitative genetic interactions has been hierarchical clustering (HCL) of genes based on pair-wise distances between their genetic interaction profiles[9, 53]. We compared the clusters obtained using HCL to the modules obtained with our present approach (Bandyopadhyay *et al.*) using three gold-standard metrics: gene co-expression (Figure 3.3a), co-functional annotation (Figure 3.3b), or membership in the same previously-identified complex (Figure 3.3c). To ensure a fair comparison between the two approaches, HCL and Bandyopadhyay *et al.* were evaluated across a range of coverages (number of gold-standard gene pairs recovered by the predicted clusters/modules; see Methods). For all three benchmarks, our performance was substantially higher than that of the HCL-based approach at most levels of coverage (and at a level of coverage corresponding to the 91 modules reported above; dotted vertical line in Figure 3.3).

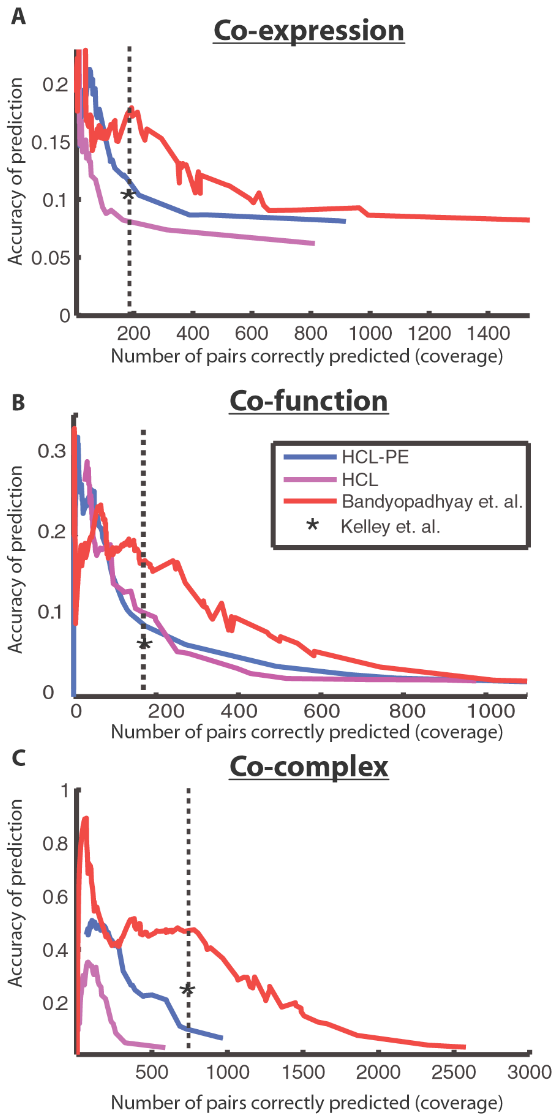


Figure 3.3. Performance of complex identification.

The proposed approach is compared to several competing methods of discovering protein complexes within genetic interaction networks: HCL implements hierarchical clustering with a distance measure computed from the genetic interaction profiles only (S-scores), while HCL-PE extends HCL by merging clusters only if there is a physical interaction between them (PE-score>1). For the modules defined by each method, accuracy versus coverage is plotted over a range of values for tuning the module size (see Methods). Accuracy is estimated as the fraction of protein pairs in a predicted module that are in a gold-standard set; coverage is estimated as the number of gold-standard pairs that fall in the same predicted module. Gold-standard sets are defined by protein pairs that are either (A) co-expressed, (B) functionally-related, or (C) assigned to the same complex in high-throughput data sets (as annotated in MIPS). The performance at the chosen parameter setting ($\alpha = 1.6$) is indicated by the dotted vertical line. The performance of the method of Kelley et al. is reported for the same level of coverage as the present approach (asterisk). Since it operates on binary interaction data, we converted quantitative genetic and physical interaction scores to binary values based on a threshold of |S|>2.5 and PE>1.

We considered that one reason why HCL performed less favorably might be that it was not given access to the same information (i.e., the physical network). This is especially true for the metric based on previously-identified complexes, in which complexes were annotated based on the same high-throughput protein interactions used here. To investigate this possibility, we extended HCL to incorporate physical interactions in a straightforward fashion, by merging only those clusters which share a physical interaction between them (HCL-PE). Although this approach outperformed hierarchical clustering without physical interactions, it was outperformed by the present approach by at least 50% across the three metrics. Finally, our method also shows improvement over the previous approach of Kelley and Ideker[59] for integrating genetic and physical interactions (Figure 3.3).

Aggravating complexes tend to be essential.

Nineteen versus nine of the learned modules had significant enrichment for alleviating versus aggravating genetic interactions, respectively. Identification of "alleviating" modules is expected, since subunits of a complex operate together and the phenotypic effect of removing any pair of proteins in a complex should be no worse than removing any single protein individually. The presence of aggravating interactions within modules was more intriguing. One way in which aggravating interactions could occur among the subunits of a complex is if its function is essential, i.e., the loss of the complex's function causes a lethal phenotype. In these cases, some protein subunits should be encoded by essential genes, while other subunits might be redundant and thus essential in pairwise combinations[65].

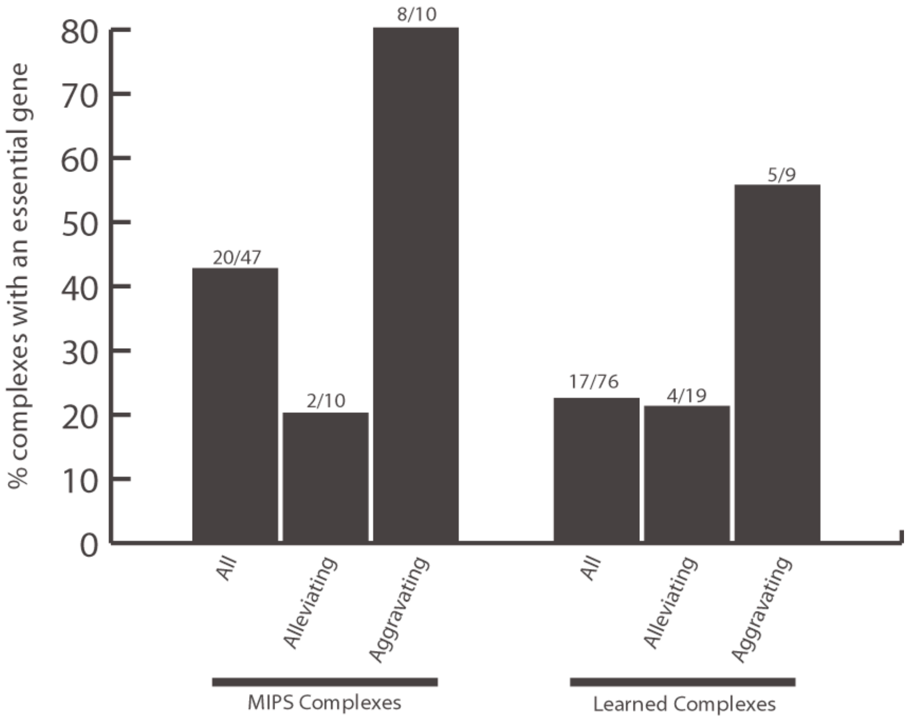


Figure 3.4. Aggravating complexes are more likely to contain essential genes.

The percentage of complexes that contain at least one essential gene is shown, for various groups of complexes defined within small-scale complexes in MIPS (left three bars) or complexes identified in this study (right three bars). In MIPS, approximately 80% of “aggravating” complexes (see text) contain an essential gene, versus 20% for “alleviating” complexes. The trend is similar for the complexes reported in this study, with 55% versus 22% of aggravating versus alleviating complexes containing an essential gene. The list of all essential genes was taken from (http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html).

To test the hypothesis that essential genes are more likely in aggravating modules, we analyzed both MIPS small-scale complexes and our learned modules for the presence of essential genes (Figure 3.4). We found that 80% of aggravating MIPS complexes contained an essential gene, compared to only 20% of alleviating MIPS complexes (a four-fold increase). Similarly, of the aggravating modules determined by our approach, 55% contained an essential gene compared to only 21% of alleviating modules (a 2.6-fold increase). These results are not correlated with module size, as the median size of aggravating learned modules is less than the median size of alleviating learned modules.

They suggest that, regardless of the technique for identifying complexes, those containing essential genes tend to be composed of primarily aggravating genetic interactions. Mechanistically, this might occur through a variety of means, including proteins with separate but functionally-redundant roles in maintaining complex integrity, or subunit substitution by paralogous proteins.

Discussion

Figure 3.5 presents detailed diagrams of example functional relationships elucidated by our module mapping method. Figure 3.5a shows the alleviating relationship between the RTT109-VPS75 histone acetyltransferase complex[53, 66, 67] and Elongator, a complex that is associated with RNA Polymerase II and is involved in transcriptional elongation[68]. Since several subunits both of Elongator and RTT109/VPS75 have been shown to be involved in histone acetylation levels[67, 69], these two complexes may operate together to effectively clear histones from actively transcribed regions. To identify further mechanisms of their cooperation, future studies may search for specific residues of histone H3 whose acetylation levels are modulated by both complexes. This example highlights the utility of an integrated approach, since although RTT109 and VPS75 are known to form a complex their genetic interaction profiles are not congruent (correlation of profiles of -0.1) and had been missed by hierarchical clustering. Figure 3.5b highlights non-essential components (LRP1 and RRP6) of the exosome, which contributes to the quality-control system that retains and degrades aberrant mRNAs in the nucleus[70]. These components have alleviating interactions with a complex composed of Lsm proteins involved in mRNA decay.

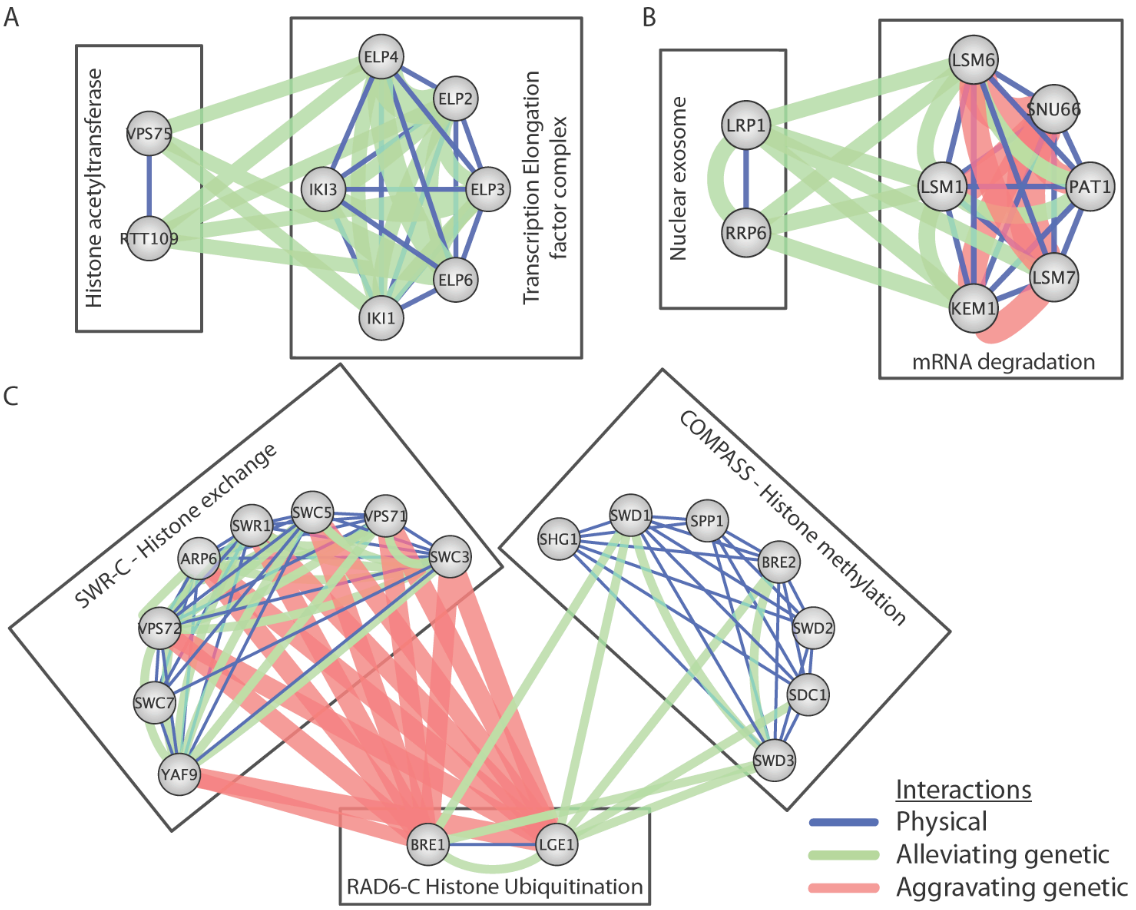


Figure 3.5. Pathway models identify novel functional associations among cellular machinery.

Each panel represents complexes and between-complex links taken from Figure 2. Physical interactions with PE>1 are shown and strong genetic interactions (|S|>2.5) are shown with increased thicknesses corresponding to stronger genetic interactions. (A) Histone acetyltransferase complex RTT109 – VPS75 showing strong alleviating interactions with the Elongator transcription elongation factor complex. (B) Between-complex model highlighting alleviating interactions between the LRP1 – RRP6 nuclear exosome complex and an mRNA degradation complex. (C) Complexes associated with the RAD6-C histone ubiquitination complex (BRE1/LGE1).

Figure 3.5c centers on BRE1/LGE1, subunits of the Rad6 Histone Ubiquitination Complex (RAD6-C; the Rad6 protein itself was not covered by the original E-MAP screen)[49, 71]. RAD6-C is functionally connected with two other complexes, SWR-C and COMPASS. SWR-C functions to regulate gene expression through the incorporation of transcriptionally-active histone variant H2AZ[72-74], while COMPASS is involved in mediating transcriptional elongation and silencing at telomeres through methylation of histone H3[75]. Interactions between RAD6-C and SWR are aggravating, suggesting synergy or redundancy towards an essential cellular function. Interactions between RAD6-C and COMPASS are alleviating, suggesting they operate in a potentially serial fashion. Consistent with this analysis, it has been shown that histone H2B ubiquitination by RAD6-C is a prerequisite for histone H3 methylation by COMPASS[76, 77].

Several trends emerge from the performance analysis in Figure 3.3. First, genetic interaction data alone can yield substantial information about molecular pathways. Functionally similar proteins often have similar profiles of genetic interaction, a feature we have previously exploited to identify functional interactions between complexes as well as to identify new members of complexes based on a combination of weak physical and genetic data[59]. On the other hand, the ability to detect complexes can be greatly improved by adding information about protein physical interactions. Even the straightforward HCL-PE method was able to greatly improve the accuracy and coverage according to most metrics, while the greatest performance was achieved by the improved probabilistic framework we have presented in this study. This framework has led to the inclusion of YKL023W as a potential new member of the SKI complex and YGR071C in a complex with VID22/TBF1 (Figure 3.2), for a total of 84 novel protein subunit assignments to complexes (Supplemental Data). Both of these examples have both physical and genetic support and would have been missed by an approach based on either type of interaction alone.

Future work may seek to incorporate yet additional types of linkages such as protein-DNA interactions[31, 78], kinase-substrate phosphorylations[79], or other genetic perturbation data such as eQTLs[80]. There are also opportunities to refine the modeling framework further. Here, a gold-standard set of complexes was used to explicitly learn the relationship between physical interactions, genetic interactions, and module membership. This supervised approach could be extended to also learn which features best indicate the inter-module functional relationships, perhaps through curation of a gold-standard set of interacting complexes.

Methods

Problem definition.

We analyze the interaction data to infer *a set of protein modules* and *a set of inter-module links* (Figure 3.1c). A protein module is defined as a set of proteins that are connected through protein-protein interactions and are likely to represent a protein complex with a coherent cellular function. Inter-module links capture functional relationships between modules and may be of two types, aggravating or alleviating. The complete state of the system is described by a set *M* of modules, each module defining a set of proteins, and a set *N* of pairs of modules that are functionally linked.

Scoring module co-membership.

For each pair of proteins (*a*,*b*) we compute a log ratio *W* of the likelihood that *a* and *b* fall *within* the same module versus the likelihood that they are unrelated (i.e., occur in the background). The function uses two sources of information that are indicative of protein complex co-membership: the strength of protein-protein physical interaction (*PE*) and the strength of genetic interaction (*S*):

$$W(a,b) = LLR_{PE}(a,b) + LLR_{S}(a,b) \quad (1)$$

For a given data type (*PE* or *S*) the log likelihood ratio (LLR) is defined as:

$$LLR(a,b) = \log \frac{P_{within}(a,b)}{P_{background}(a,b)} \qquad (2)$$

The probability P_{within} is determined using logistic regression training on 217 complexes curated from small-scale studies in MIPS[64]. $P_{background}$ is the probability of randomly observing the observed value (*PE* or *S*) for the pair (*a*,*b*) in the background of all gene pairs. As shown in Figure 3.1 and 1b, it is clear that higher values of *PE* are predictive of MIPS complex membership. As both positive and negative values of *S* are predictive, $LLR_S(a,b)$ is trained on the absolute value of *S*. A third predictor based on the correlation of genetic interaction profiles was also evaluated but did not result in any gain in performance (Supplemental Figure 3.1).

Scoring inter-module links.

A similar function *B*() is formulated to assess the likelihood that two proteins fall *between* modules that are functionally linked. The function inputs the same two sources of information on protein-protein and genetic interactions (*PE* and *S*). Unfortunately, there is no curated set of functionally-related complexes that can be used as positive training examples for regression. Instead, *B*() is derived from the within-module LLRs, assuming that between-module interactions have a similar pattern of genetic interactions but lack physical interactions:

$$B(a,b) = -LLR_{PE}(a,b) + LLR_S(a,b) \quad (3)$$

This function captures both aggravating and alleviating genetic interactions between two functionally-related modules. It also ensures such modules are physically separate—if not, they would be better considered as a single module.

Global optimization of module memberships and links.

Given the above functions *W*() and *B*(), we compute the likelihood of the complete system (i.e., given a particular choice *M* of modules and *N* of inter-module links):

$$L = \left(\sum_{m \in M} \sum_{(a,b) \in m \times m} W(a,b) \right) + \left(\sum_{(m_1,m_2) \in N} \sum_{(a,b) \in m_1 \times m_2} B(a,b) \right) + \left(\sum_{m \in M} |m|^{\alpha} \right) \qquad (4)$$

The first term accumulates the within-module scores among gene pairs assigned to the same module. The second term accumulates the inter-module scores for gene pairs spanning any two modules. Gene pairs spanning unlinked modules do not contribute to *L*. The final term is a tunable reward which scales with module size. Larger values of *α* result in fewer, larger complexes. The final module map shown in Figure 3.2 was generated using α=1.6, based on its good coverage and performance across all three metrics in Figure 3.3.

Module search.

Assignment of gene to modules and of inter-module links is performed using a simple variant of UPGMA hierarchical clustering[81]: (a) Initially, each gene is assigned to a separate module; (b) Each pair of modules (m_1, m_2) is evaluated for merging into a single module $m = m_1 \cup m_2$; the pair-wise merging that results in the largest increase in *L* is chosen; (c) Repeat step b until no module merge operation increases *L*.

At each iteration of step b, *L* is optimized over all possible ways of assigning inter-module links (i.e., module pairs are linked whenever the second term in Eqn. 4 is positive). Because each inter-module link is scored independently, additions or deletions of links from the system need only be evaluated for modules that are under evaluation for merging.

Subsequent to the above procedure, each between-module link is evaluated to assess its significance and whether it represents predominantly aggravating or alleviating genetic interactions. A two-tailed p-value is computed by indexing the sum of *S*-scores for gene pairs falling across the two modules against a distribution of 10^6 sums of equal numbers of *S*-scores drawn from random gene pairs. To account for multiple testing, we use the distribution of between-module p-values to compute a local false discovery rate (FDR)[82]. All reported between-module links have an inferred FDR of <10% with the global map in Figure 3.2 constrained to links with an FDR of <1%. Module maps in Figure 3.2 and Figure 3.5 are visualized using the Cytoscape package[39, 83].

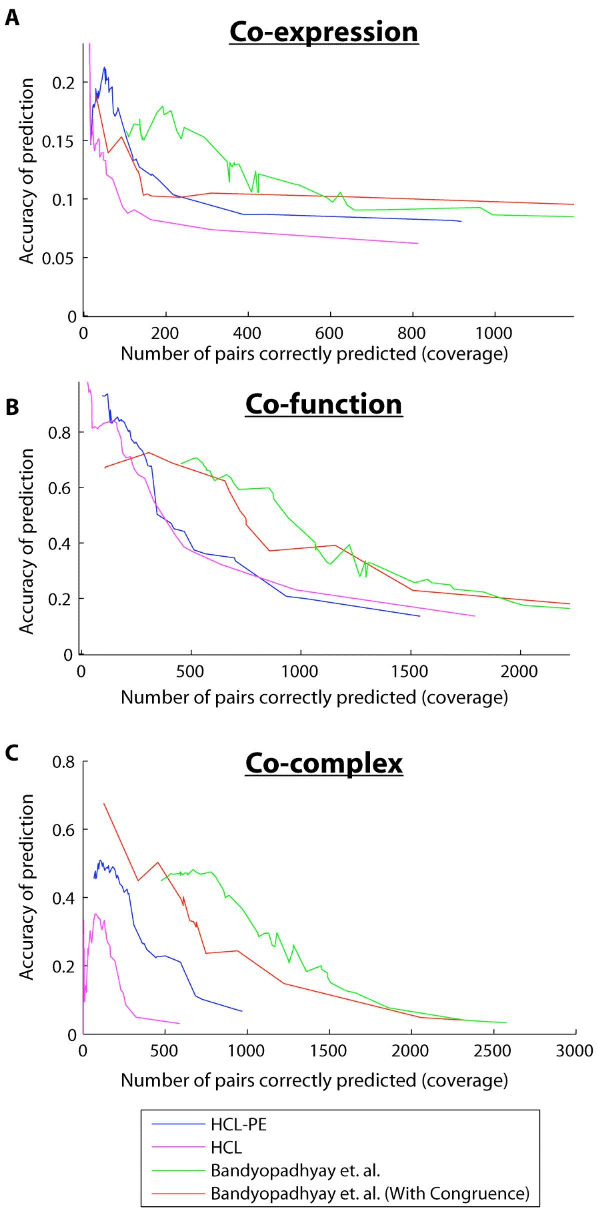
To label modules as "aggravating" or "alleviating" (Figure 3.2), the sum of *S*-scores for gene pairs assigned to the same module is compared to a distribution of sums of equal numbers of randomly drawn *S*-scores. Modules with a two-tailed p-value < 0.05 are labeled as either alleviating (right tail) or aggravating (left tail).

Validation using co-expression, co-function, or co-complex annotations.

Co-expressed gene pairs were defined using gene expression datasets culled from the Stanford Microarray Database covering ~790 conditions[84]. The validation set was taken as the top 5% (13,014) of pairs ranked by Pearson correlation coefficient. The co-function set was based on yeast Gene Ontology annotations from November 2005 which predates the publication of large scale TAP-MS studies that were used to generate the PE-score. This set was taken as the top 5% (13,052) most functionally similar gene pairs covered in the E-MAP. Functional similarity was determined by comparison to the background probability of picking two genes with the same shared functional annotation from the entire yeast genome (via a hypergeometric test). Similar analysis using current Gene Ontology annotation was also performed (Supplemental Figure 3.2). The co-complex validation set was defined as gene pairs from 846 MIPS complexes annotated using high-throughput approaches (with interactions also appearing in small-scale studies removed) for a total of 2,885 gold-standard pairs.

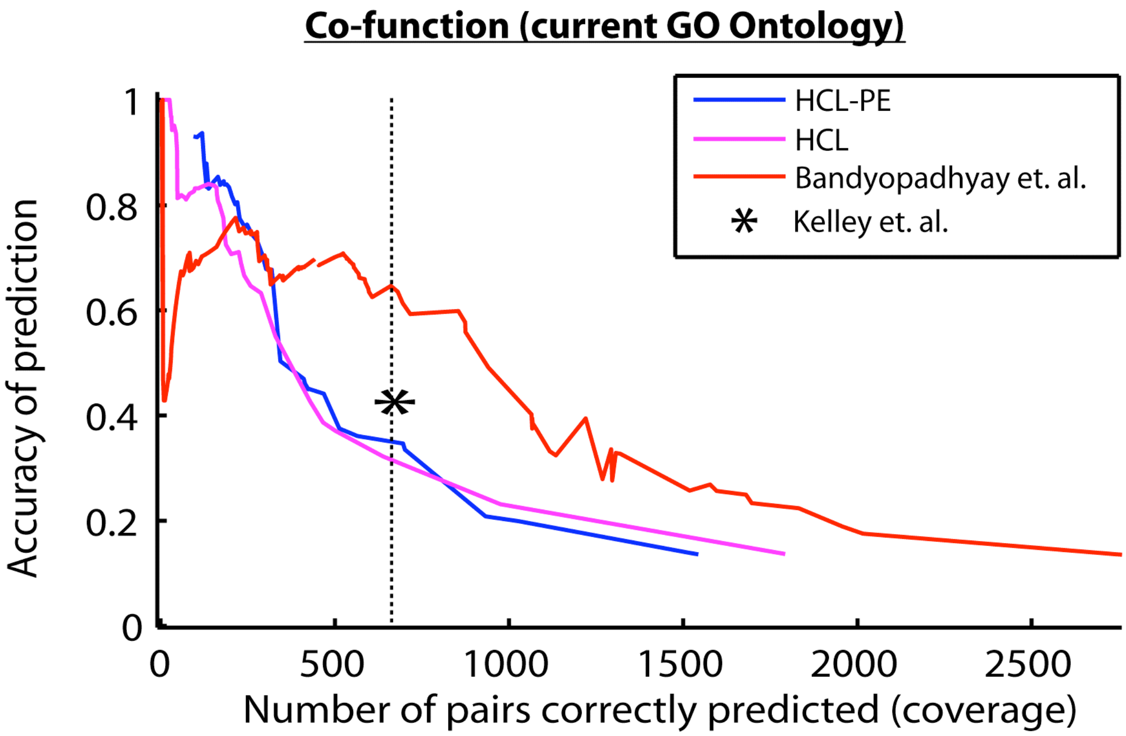
The size and number of final modules was varied by altering the α parameter (see above). To assess performance at low coverage we ran the method with no reward contribution (remove the third term in eq. 4 by setting $\alpha = -\infty$) and plotted the performance of the algorithm at each merge step, which ultimately connects with the performance of the method as α is increased. For HCL and HCL-PE methods, the size and number of modules were varied by changing the level at which the hierarchy was cut.

Supplemental Figures



Supplemental Figure 3.1. Addition of congruence as a predictor of pathway membership.

A variant of this algorithm which includes congruence (measured as the pearson correlation of genetic interaction profiles) was included as a third predictor (beyond pairwise physical and genetic interaction scores)**.** The results indicate that, especially in determining co-complex membership, the addition of congruence does not help to find functionally related modules. A possible rationale for this result is that by scoring between-complex interactions explicitly, the method is already rewarding for similarity of genetic interaction profiles so that the addition of the third congruence predictor results in overfitting and no additional gain in performance.



Supplemental Figure 3.2. A current version of the Gene Ontology shows similar performance.

The figure is the same as Figure 3B using the current version of the Gene Ontology (March 2007).

Acknowledgements

We gratefully acknowledge funding from the National Institute of Environmental Health Sciences (ES14811), the National Institute of General Medical Sciences (GM070743) and from a Sandler Family Fellowship. We also thank Sean Collins for helpful discussions.

Chapter 3, in full, is a reprint of the following work,

Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T. *Functional maps of protein complexes from quantitative genetic interaction data*. **PLoS Computational Biology** 2008; 4(4).

The dissertation author was the second author on this work, responsible for designing and implementing computational algorithms.

Chapter 4. Correcting for gene-specific dye bias in DNA microarrays using the method of maximum likelihood

Abstract

In two-color microarray experiments, well known differences exist in the labeling and hybridization efficiency of Cy3 and Cy5 dyes. Previous reports have revealed that these differences can vary on a gene-by-gene basis, an effect termed gene-specific dye bias. If uncorrected, this bias can influence the determination of differentially expressed genes. We show that the magnitude of the bias scales multiplicatively with signal intensity and is dependent on which nucleotide has been conjugated to the fluorescent dye. A method is proposed to account for gene-specific dye bias within a maximum likelihood error modeling framework. Using two different labeling schemes, we show that correcting for gene-specific dye bias results in the superior identification of differentially expressed genes within this framework. Improvement is also possible in related ANOVA approaches. A software implementation of this procedure is freely available at http://cellcircuits.org/VERA.

Introduction

Two color microarray experiments are an instrumental tool in modern biology[5]. In a typical experiment, RNA is extracted from two samples (populations of cells); labeled with Cy3 or Cy5 fluorescent dyes, respectively; hybridized to an array of DNA probes; and imaged with a confocal scanning device. Due to differences in dye chemistry, the measured intensity distributions for each dye are not directly comparable. Several normalizations are commonly applied to address this issue. First, each intensity distribution is median centered[85, 86]. Second, the LOESS procedure is used to normalize the intensity dependent bias of each dye[87]. In LOESS, the bias at each intensity is estimated from a window of data points with similar intensity values. This estimate is then used to correct the measured values at that intensity. In order to obtain meaningful results from two-color microarrays, it is important that both of these biases are corrected.

Recently, an additional source of systematic error in two-color microarray experiments has been identified[88-90]. Although still dye-dependent, unlike the aforementioned sources of error its magnitude varies according to each individual measured transcript. Accordingly, this bias has been termed Gene-Specific Dye Bias (henceforth abbreviated GSDB), and even data that have been median-centered and LOESS-corrected will display a consistent bias in either the Cy3 or Cy5 direction for a given probe. This effect has been observed on a variety of platforms and labeling systems, including PCR-spotted and short oligonucleotide arrays used in conjunction with either direct or indirect labeling methods[88]. In addition to this work with two-color arrays, sequence specific effects have been reported within single color array systems such as Affymetrix GeneChips[91, 92]. These effects can confound the discovery of differentially expressed genes (false negatives) or, depending on the experimental design, lead to their erroneous identification (false positives)[89].

In a proper experimental design, the dyes used to label a given sample are balanced. That is, every microarray experiment is duplicated by one that reverses the Cy3 vs. Cy5 labeling orientation of the samples (i.e., such that Cy5 labels the first sample and Cy3 labels the second). Dye balancing mitigates gene-specific dye bias because the direction of bias alternates from replicate to replicate such that the average effect is zero. However, although the mean bias is zero the variance across replicate measurements is now greatly increased by the presence of gene-specific dye bias. Increased variance, in turn, decreases the sensitivity in identifying differentially expressed genes.

Recognizing the limitations of dye balancing experiments, the problem of GSDB has been addressed using a variety of sophisticated experimental and bioinformatic techniques. Rosenzweig et al.[90] proposed to handle GSDB with a modified experimental design utilizing the addition of control microarrays. They found that employing their strategy with 10 replicate microarrays could yield comparable technical accuracy to a 16 replicate experiment performed with a traditional balanced design. Using an analysis of variance (ANOVA) model, Martin-Magniette et al.[93] developed a test statistic (the label bias index) to measure the extent of GSDB across a microarray and discussed possible ramifications on the design of indirect comparison experiments. In a related approach, Dobbin et al.[88] characterized GSDB as well as other sources of systematic error such as cell-line specific bias. Correcting for GSDB within an ANOVA framework, they found significant differential expression for approximately 18% more genes than if such correction was not applied. Without a gold standard set of differentially expressed genes, however, it is unclear whether this represents an increase in the number of true or false positives.

One limitation of ANOVA is that the general linear framework does not capture all of the complex errors that could possibly influence a microarray experiment. Therefore, in parallel to ANOVA, several groups have proposed more advanced microarray error models, e.g., that capture both additive and multiplicative errors influencing each measured dye intensity[94-96]. A maximum-likelihood approach is then used to optimize model parameters and to score differentially-expressed genes. On the one hand, these models have the potential to more closely reflect the true error structure. On the other, it is unclear whether the additional complexity is warranted, and none of these models have been updated to account for the presence of GSDB.

Here, we present our efforts to both characterize gene-specific dye bias and to extend a maximum-likelihood error modeling approach to correct for its influence. By conducting the identical gene expression experiment using two different labeling systems, we demonstrate that correcting for the presence of GSDB results in the improved detection of differentially-expressed genes.

Methods

Error model

The proposed error model expands upon previous work to determine differentially expressed genes through the incorporation of both multiplicative and additive error (the VERA error model)[95]. To extend this model to capture GSDB, it is conceptually possible to model this bias as either a multiplicative or additive error term. Equations (4.1)-(4.4) display a concise representation of the error model as originally proposed, with additional terms to capture GSDB as multiplicative error.

$$x_{ij} = \mu_{x_i}(1 + \varepsilon_{x_{ij}} + I(Cy5)\beta_i) + \delta_{x_{ij}} \tag{4.1}$$

$$y_{ij} = \mu_{y_i}(1 + \varepsilon_{y_{ij}} + I(Cy5)\beta_i) + \delta_{y_{ij}} \tag{4.2}$$

$$\varepsilon_x \sim N(0,\sigma_{\varepsilon_x}), \varepsilon_y \sim N(0,\sigma_{\varepsilon_y}), Corr(\varepsilon_x,\varepsilon_y) = \rho_\varepsilon \tag{4.3}$$

$$\delta_x \sim N(0,\sigma_{\delta_x}), \delta_y \sim N(0,\sigma_{\delta_y}) \tag{4.4}$$

Alternatively, to model bias as additive error, equations (4.1) and (4.4) are replaced with (4.5) and (4.6), respectively.

$$x_{ij} = \mu_{x_i}(1 + \varepsilon_{x_{ij}}) + I(Cy5)\beta_i + \delta_{x_{ij}} \tag{4.5}$$

$$y_{ij} = \mu_{y_i}(1 + \varepsilon_{y_{ij}}) + I(Cy5)\beta_i + \delta_{y_{ij}} \tag{4.6}$$

Here, (x_{ij}, y_{ij}) are the observed dye intensities for gene i in replicate j. The variable μ is the true underlying intensity for each dye, while ε and δ represent multiplicative and additive error terms, respectively. Each of these error terms is normally distributed with mean zero and distinct standard deviation s. The multiplicative errors ε_x and ε_y may be highly correlated (with coefficient ρ_ε). It is possible to also include a correlation term for the additive errors; however, in practice, this correlation is near zero. Extending beyond previous work, the model is given the additional gene-specific bias term β. This correction is only applied if the values are taken from Cy5 intensity data, as enforced by the indicator function I(Cy5). The symmetric model, in which the correction is applied to the Cy3 channel only, would perform identically with the exception that the learned bias terms would be negated.

To fit the model to gene expression data, for each gene a total of three parameters (μ_x, μ_y, β) must be learned, in addition to the five global error parameters ($\sigma_{\varepsilon x}$, $\sigma_{\varepsilon y}$, ρ_ε, $\sigma_{\delta x}$, $\sigma_{\delta y}$) shared over all genes. Maximum likelihood estimates of all parameters are derived via an iterative procedure implemented in the MATLAB programming language[95]. Briefly, after selection of initial values for all parameters, the global error parameters are optimized to maximize the likelihood function utilizing a conjugate gradient approach[97]. These new global error estimates are then held constant during a similar estimation of the gene-specific parameters (μ_x, μ_y, β). These two optimizations continue to alternate in an iterative fashion until estimates for all parameters have converged. Through simulation, it is apparent that the parameters estimated in this fashion are subject to bias due to small-sample size (i.e., small numbers of replicates). Appropriate corrections are applied to remove this bias, as described in Supplemental Figs. 4.1 and 4.2.

Following parameter estimation, a generalized likelihood ratio test is used to assess the extent of differential expression for each gene. According to this test statistic, the likelihood of the expression data for a gene under the optimal model parameters (numerator of the likelihood ratio) is compared to the likelihood of the same data under an alternative model with the constraint $\mu_x = \mu_y$ (the “null” hypothesis of no differential expression; denominator of the likelihood ratio).

Assessing Dye Bias

The VERA error model incorporating bias as an additive term was applied to the set of control data. For each gene, a single bias term β was learned. To determine the relationship between overall intensity and the magnitude of bias, the “lowess” function in

R (with default parameters) was used to calculate a smoothed estimate of the absolute value of bias as a function of the average value of μ_x and μ_y.

ANOVA analysis

Within an ANOVA framework, different methods can be used to estimate differential expression based on how the residual error for each gene is determined. The R/maanova package defines four such measures: F1, F2, F3, and Fs[98]. F1 is the usual F statistic, which determines the residual error independently for each gene, while the remaining measures represent different ways of pooling the residual error over multiple genes[99]. F3 models a single residual averaged over all genes, while F2 sets the residual for each gene as an average of its F1 and F3 estimates. The Fs statistic is similar to the F2, but uses the heterogeneity of the error estimates to inform the exact weighting of the average. As a fifth measure, the R/VarMixt package (Delmar, et al., 2005) was used to model residual error as a mixture of different sub-populations of genes, as employed by Martin-Magniette et al.[93] in their earlier assessment of GSDB (see Introduction). In each of these five cases, a fixed ANOVA model was employed using the factors Array, Dye, and Sample. In the case of the non-dye-bias-corrected analysis, Dye was not used as a factor.

Sample Growth and Treatment

In total, twelve microarray experiments were performed, four control (comparing untreated vs. untreated) and eight treatment (comparing untreated vs. mild hydrogen peroxide treatment). In each control microarray experiment, a single colony of BY4741

(ATCC, Manassas, Virginia, USA) was used to inoculate 10 mL of YPD media. Following overnight growth at 30o C, this culture was then resuspended in 100 mL media at an OD600 of 0.1 and placed in an orbital shaker at 30o C. Following growth to OD600 = 0.6, the culture was split into two 50 mL portions and allowed to continue growth to OD600 = 1.0. Cells were then harvested by centrifugation at 3,000 rpm for 5 minutes. Pellets were immediately frozen in liquid nitrogen and stored at -80° C. Handling of the mild hydrogen peroxide treatment samples was similar, except that one member of each aliquoted pair was treated with 0.1 mM hydrogen peroxide 1 hour prior to collection.

RNA extraction, labeling, and hybridization

RNA from each sample was isolated via phenol extraction followed by mRNA purification (Poly(A)Purist, Ambion, Catalog # 1916). Purified mRNA from the control experiments was labeled with dUTP incorporating either Cy3 or Cy5 dye (CyScribe First-Strand cDNA labeling kit, Amersham Biosciences). The eight hydrogen peroxide treatment pairs were broken into two equal-sized groups of four pairs each. In one group, dUTP-labeled dye was used to label the transcripts, while in the other group, dCTP-labeled dye was substituted. Within each group, Cy3 and Cy5 labelings were assigned to create a balanced design. Complementary labelings (Cy3 vs. Cy5) were hybridized to an Agilent oligonucleotide expression array (Catalog # G4140B).

Data acquisition and analysis

Arrays were scanned using a GenePix 4000A and quantified with the GenePix 6.0 software package. Prior to further analysis, the data from each array were subjected to background and quantile normalization[100].

Comparing replicates

Each error model (VERA and the five ANOVA variants) was used to rank genes according to their significance of differential expression, for both the dUTP-labeled and dCTP-labeled sets of replicate microarray experiments (hydrogen-peroxide treated versus untreated). For a given rank cutoff, a superior GSDB correction method should result in higher overlap between the sets of differentially expressed genes identified by the two labeling methods. To ensure that this overlap is due to the enhanced identification of true positives and not shared false positives, a "baseline overlap" value was also calculated between ordered lists derived from the dCTP-labeled treatment series and the control series. Since there are no truly differentially expressed genes in the control series, any overlap in this comparison represents shared false positives or random overlap events. The actual overlap was reported after subtracting this baseline value.

To assign significance values of differential expression to the control series, two of the four arrays must be arbitrarily assigned as the "forward" labeling. Since there are three equally valid such assignments, the baseline overlap was determined in all three configurations and the average was used.

Results

Characterizing gene-specific dye bias

We first performed a series of microarray controls to confirm and further characterize the extent of gene-specific dye bias. Two samples of mRNA extracted from yeast undergoing exponential growth in identical conditions, were directly labeled with either Cy3 or Cy5 dyes conjugated to dUTP. These labeled samples were co-hybridized to an Agilent v2 Yeast Oligo Microarray, and ln(Cy3/Cy5) ratios were determined for each gene following median and quantile normalization. Additional cultures, mRNA extractions, and hybridizations were analyzed to generate a total of four separate microarray replicates.

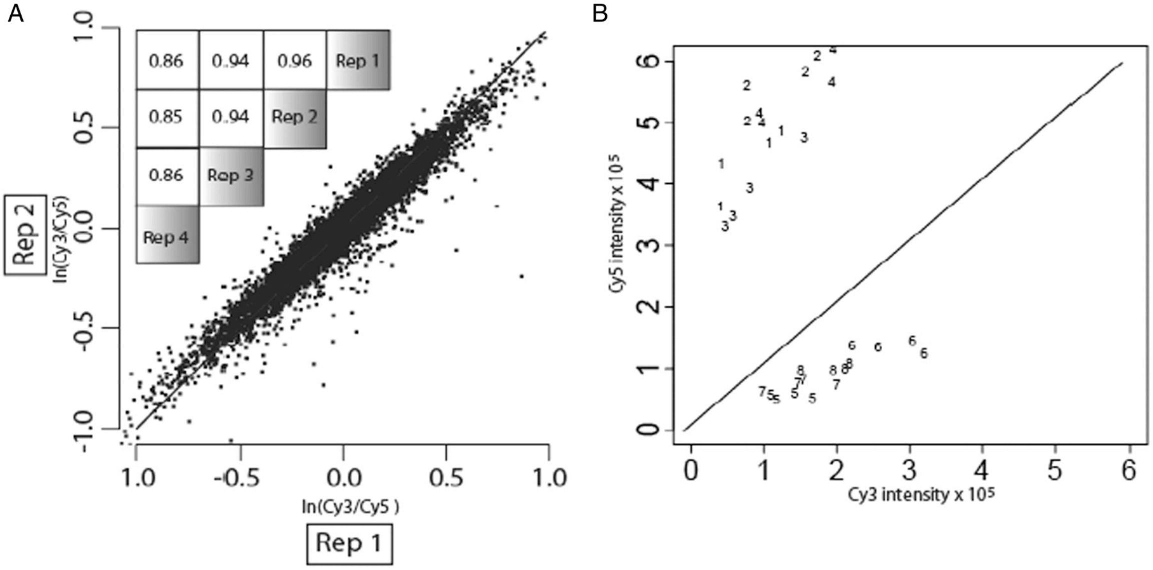


Figure 4.1. Gene-specific dye bias in oligonucleotide arrays.

Gene-specific dye bias is present and highly reproducible in an oligonucleotide expression microarray system. The scatter plot of panel A details a comparison of log ratio values from two separate control experiments. The inset in the upper left quantifies all six pair-wise correlations among the four replicate control experiments. As a different perspective on the same information, panel B presents the four replicateCy3 versus Cy5 intensity values for several genes (numbers 1–8) with apparent large gene-specific dye bias.

Since mRNA for each labeling was extracted from identical conditions, the true log ratio for all genes is zero. When examining multiple replicates, the observed log ratio deviates from zero due to various sources of error, such as uncontrollable biological variation between replicates and noise in the experimental analysis. If there is no gene-specific bias, the value of this deviation will vary around zero and will not be reproducible across replicates. However, as shown in Figure 4.1, this is strikingly not the case. When comparing two control experiments, the correlation over all log ratio values is at least 0.85, illustrating the presence of clear gene-specific bias. Since the only difference between the numerator and denominator of the log ratio is the dye used for labeling, this gene-specific effect must be dye bias. For the most affected genes, the bias effect alone can cause the ratio to deviate by more than two-fold. Such a deviation can easily influence determination of differential expression.

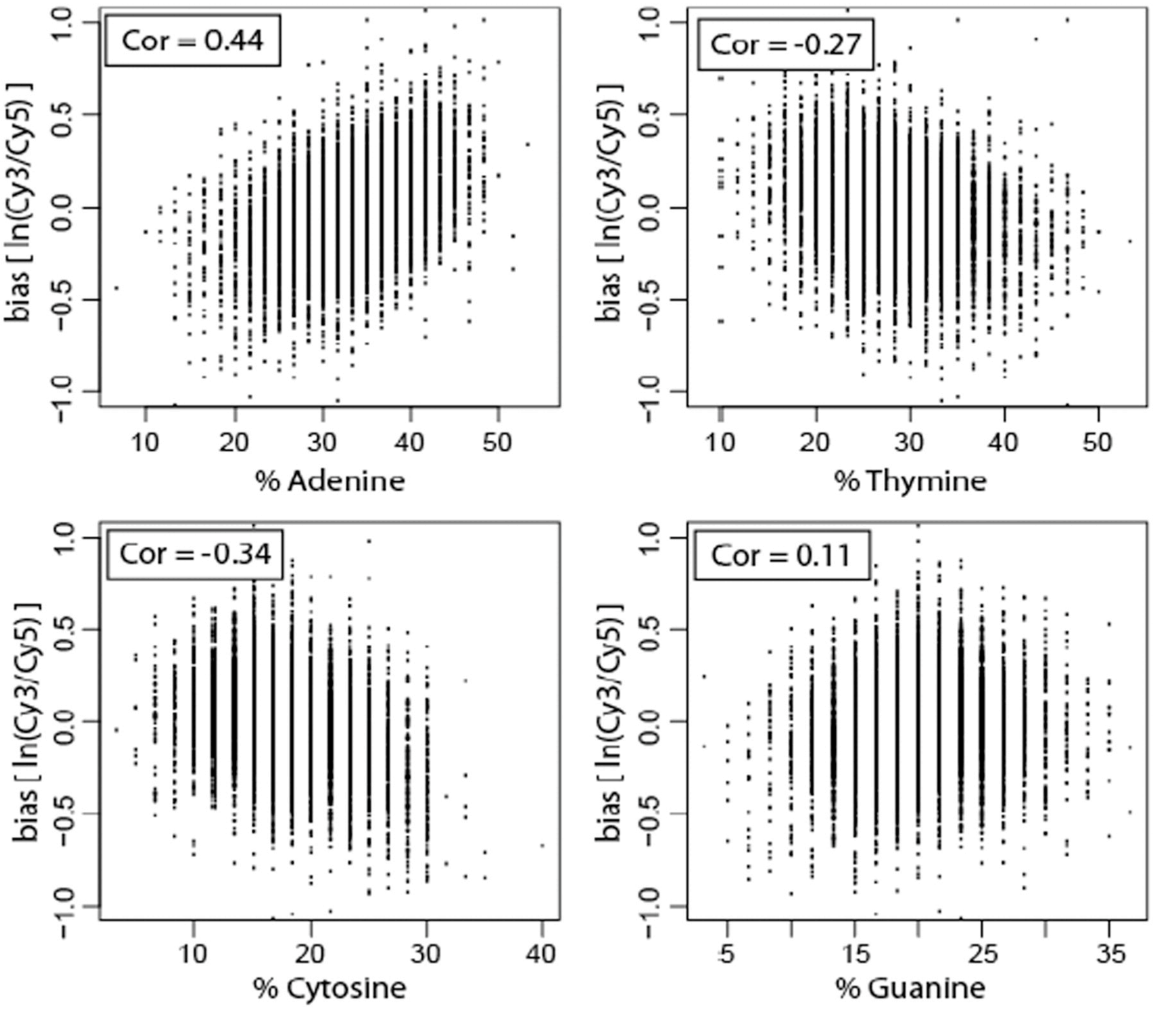


Figure 4.2. Bias strength is related to labeled nucleotide.

The upper left panel shows that strongest correlation between gene-specific dye bias in a dUTP-labeled control experiment and nucleotide content is with the frequency of adenine.

To further investigate the source of bias, we computed the correlation between the dye bias of each gene and the frequency of each nucleotide (A,C,G,T) in the sequence representing the gene on the microarray (Figure 4.2). Gene-specific dye bias was measured as the average natural log ratio (Cy3/Cy5) over the four replicate control hybridizations. The most significant correlation was found with adenine content (Figure 4.1A). Since the cDNA was labeled with Cy3 or Cy5 dyes conjugated to dUTP (the complement of adenine), the bias is thus proportional to the number of incorporated dye molecules. This result is then consistent with the less efficient incorporation of Cy5 dye by the polymerase.

Formulating an error model

It is possible to model bias as either a multiplicative or additive error term (see Methods). If the values of μ_x and μ_y vary substantially, the effect of an additive bias term will be different than a multiplicative one (i.e., only a multiplicative bias term will scale with the magnitude of μ). However, this distinction is irrelevant if the true intensity values for each dye (μ_x and μ_y) are equal. While this is generally not true, it is the case for the control experiments presented previously. Therefore, control data can be used to decide if it is more appropriate to model bias as a multiplicative or additive error term. Using an additive error model, we learned bias values for each gene in the control data. Figure 4.3 shows the relation between the absolute magnitude of this bias and the mean signal intensity. Across different genes, there is a clear multiplicative relationship between the magnitude of bias and the mean signal intensity. An equivalent result was determined when a multiplicative error model was applied instead. Since bias terms tend to increase multiplicatively with mean intensity, it is likely more appropriate to model bias as a multiplicative error term.

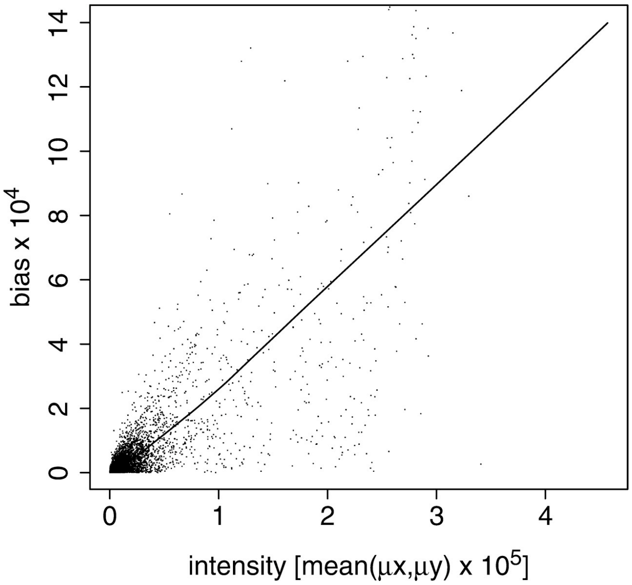


Figure 4.3. Gene-specific dye bias is multiplicative in nature.

The VERA error modeling procedure is applied to control data and used to determine the values of the parameters µx, µy and β for each gene. Here, the smoothed estimate of the absolute value of β is plotted as a function of the mean value of µx and µy. The data used to generate this smoothed line is also displayed as individual points.

Benchmarking model performance

We next set out to determine whether the VERA model was able to correct for the presence of gene-specific dye bias in experimental data. The original set of control expression profiles was analyzed with both the corrected (multiplicative bias) and uncorrected (no bias) models. Figure 4.4 displays the distribution of $\ln(\mu_x/\mu_y)$ values from each analysis. In the case of the corrected VERA method, the spread of log ratio values is much tighter around the origin. Quantitatively, the variance of the uncorrected log ratios is 5.2*10-3, compared to 3.4*10-3 for the corrected algorithm. Thus, following bias correction the observed ratios tend to be closer to the true expected value of zero.

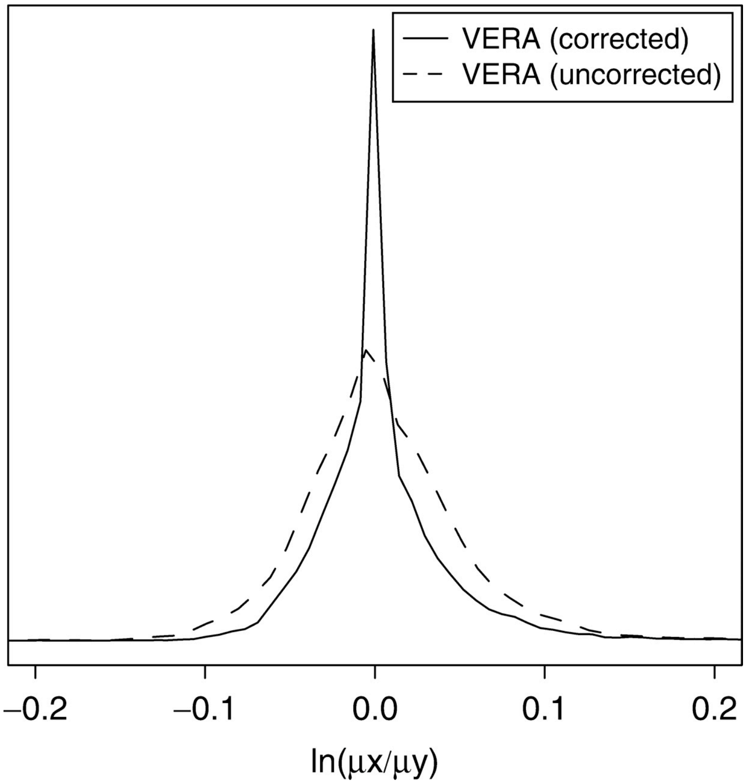


Figure 4.4. Application of dye-bias correction reduces variance in a control experiment.

The solid curve represents the probability distribution of log ratio values determined following application of the corrected VERA method to control data. Conversely, application of the uncorrected VERA approach to the same data results in a distribution of log ratio values with larger variance (dashed line).

To further validate our approach and to benchmark it against other methods that have been proposed for correcting dye bias, we performed two additional sets of experiments. In each experimental set, we profiled the response of yeast to mild oxidative stress (0.1 mM hydrogen peroxide vs. nominal conditions) over four replicate microarrays. The only difference between sets was that in one case, dUTP was used in the labeling process, while in the other dCTP was used. Since the frequency of the labeled nucleotide within a sequence is related to its gene-specific bias, the two labeling schemes create different gene-specific dye biases while preserving the same true changes in gene expression. A method which correctly accounts for and eliminates the effect of gene-specific dye bias should maximize the agreement between these two data sets.

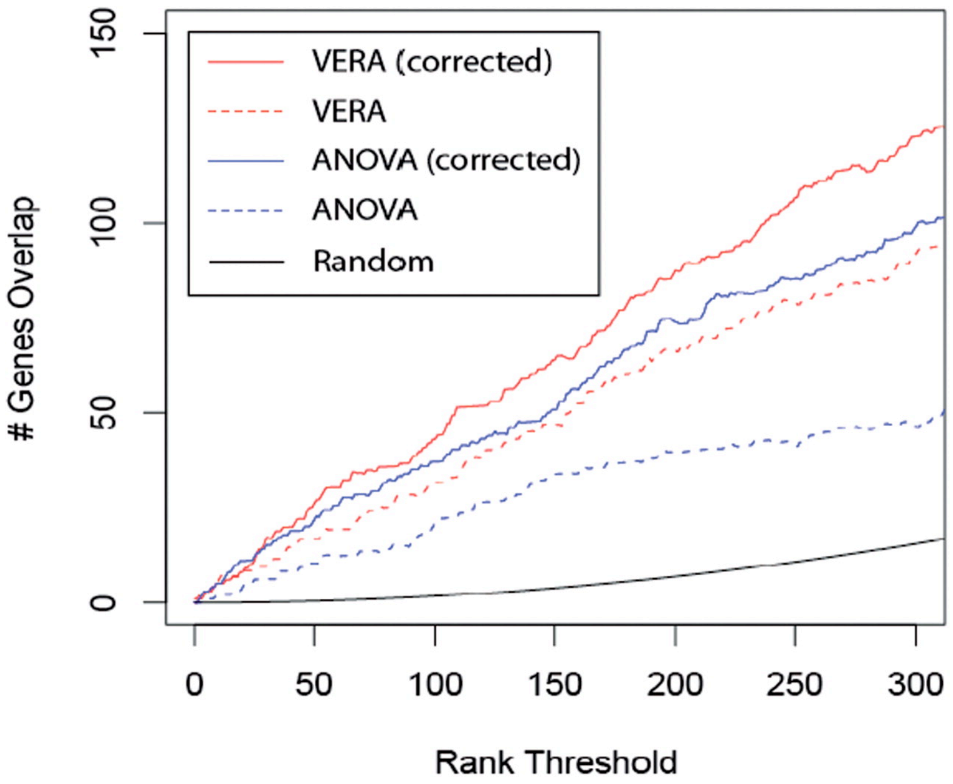


Figure 4.5. The dCTP- versus dUTP-labeled expression data is compared for different analysis methods.

Since the true number of differentially expressed genes is unknown, the calculation is performed over a range of values (x axis). The y axis shows the number of genes assumed to be significant in both labeling approaches after correcting for any bias in the method.

Figure 4.5 compares the ability of different methods to recover differentially-expressed genes in the dUTP-labeled set that were identified in the dCTP-labeled set also. Previous methods to correct for GSDB model the effect as an ANOVA factor. To implement this approach, we relied upon the MAANOVA and VarMixt packages[98, 101]. Since the true number of differentially expressed genes is unknown, this comparison was performed over a range of thresholds for calling differentially expressed genes[102]. At nearly all possible points in this range, the bias corrected VERA approach displayed the best performance. This was followed by the corrected ANOVA statistic and the uncorrected VERA approach. ANOVA results are reported for the Fs statistic; as it previously showed the best performance over a wide range of simulated data[99]. At a rank threshold of 300, the overlaps for all methods are significantly enriched over random (hypergeometric p-value = 5.4*10-9 for uncorrected Fs statistic). The improvement of performance of the corrected VERA algorithm over the uncorrected one is also significant at the same rank threshold (binomial p-value = 3.5*10-5). Comparison to alternative versions of the F-statistic (F1, F2, F3, and VarMixt) are available in Supplemental Figure 4.3.

When the choice of labeled nucleotide is changed from dUTP to dCTP, one would expect the correlations between dye bias and nucleotide content to be altered as well. Indeed, in the dCTP labeling experiments, we observed the strongest dye bias correlation was with guanine frequency (correlation = 0.39) rather than adenine frequency as observed earlier for dUTP. This reinforces the finding that the choice of labeled nucleotide has a strong impact on gene-specific dye bias.

Discussion

The performance of VERA improved significantly when corrected for GSDB. For the ANOVA F2, Fs, and VarMixt statistics, dye-bias correction also improved performance (Figure 4.5 and Supplemental Figure 4.3), while little to no improvement was observed for the F1 and F3 statistics. For the F1 statistic, it is likely that the lack of shared error estimates across genes in combination with the small sample size made accurate error estimation difficult, even with dye-bias correction. For the F3 statistic, the estimate of error is identical for all genes by definition. Therefore, since the dye-bias correction in the ANOVA framework affects only the relative determination of gene-specific residual error, the F3 rankings of differential expression must be identical with and without correction. VERA's greater agreement between dCTP- and dUTP-labeled experiments (compared to ANOVA) is likely due to its more complex error model, which accounts for both additive and multiplicative errors. The ANOVA models account for multiplicative error only (which becomes additive after log transformation of the intensity values). On the other hand, ANOVA provides a flexible framework which can be easily extended to handle additional factors influencing an experiment (e.g., cell-line, treatment, dye, array).

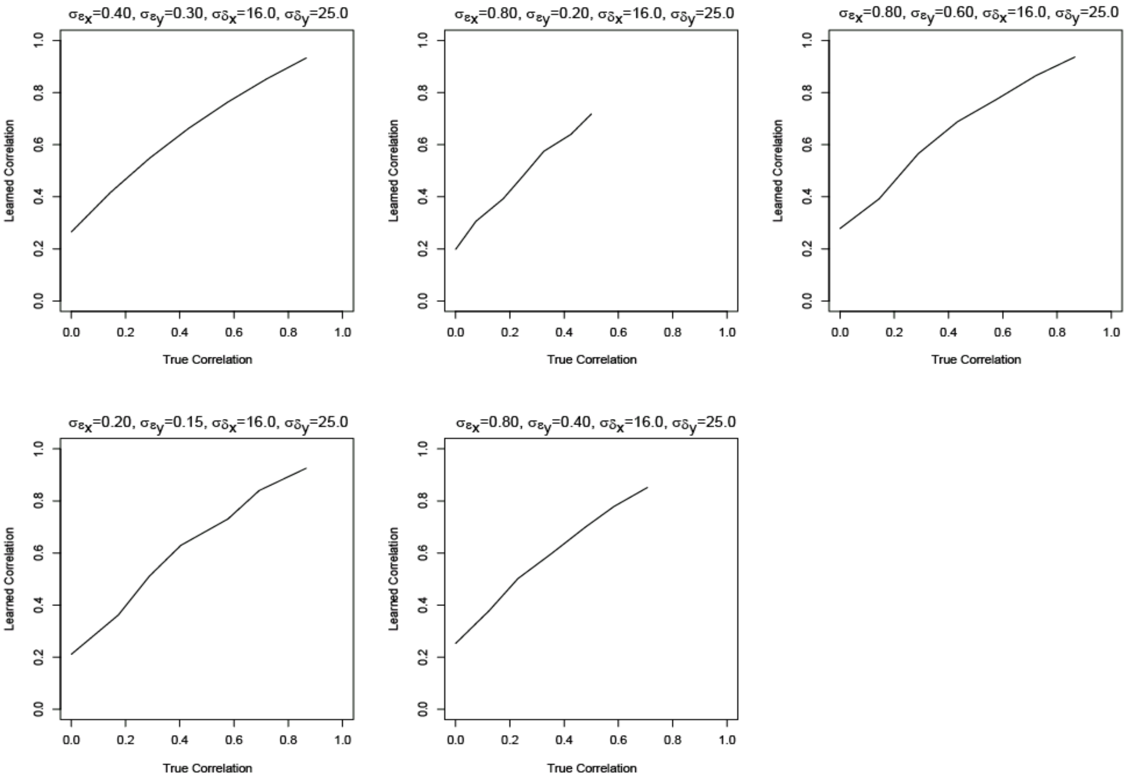
While error models such as these can mitigate the effect of gene-specific dye bias, it would always be preferable to remove or reduce such bias if possible. Having identified nucleotide content as one contributing factor, this information might be useful in the future design of arrays. For example, probes might be chosen so as to minimize variation in adenine nucleotide content. An alternative might be to use a mix of labeled nucleotides during first strand cDNA synthesis.

In the exploratory phase of this work, we used the average ratio values determined from control experiments as an estimate of gene-specific dye bias. Only later was this bias modeled explicitly in the context of a probabilistic framework incorporating other errors. However, this raises an important question. Is an error modeling process required at all? Alternatively, one could simply estimate bias values from replicated controls and directly apply these estimates to future experimental results. One problem with this simpler approach is that not all genes are highly expressed under control conditions. The signals associated with low intensity genes would still be dominated by error, especially when these genes become highly expressed in some other (non-control) condition. In addition, Rosenzweig et al.[90] noted that the gene-specific dye bias can be somewhat variable between experiments. Therefore, the values learned in a control experiment may be inapplicable, whereas the maximum-likelihood model is custom-fit to each experimental data set.

In a properly balanced microarray experiment, the influence of gene-specific dye bias on the production of false-positive measurements is mitigated, if not eliminated. As Dobbin et al.[88] noted, the predominant effect is the generation of more false negatives. In addition, gene-specific effects can alter the ordering of significant genes, which many statistical methods rely upon. How important is it then to correct for gene-specific dye bias? This is a question that cannot be addressed in a universal manner. As shown by our experiments with different labeled nucleotides, the magnitude of gene-specific dye bias is apparently platform specific, and its impact depends critically on this magnitude in relation to the magnitude of the expression changes occurring in the biological system. Certainly, if the reliable identification of subtle differential expression changes is desired, then correcting for this systematic bias is crucial.

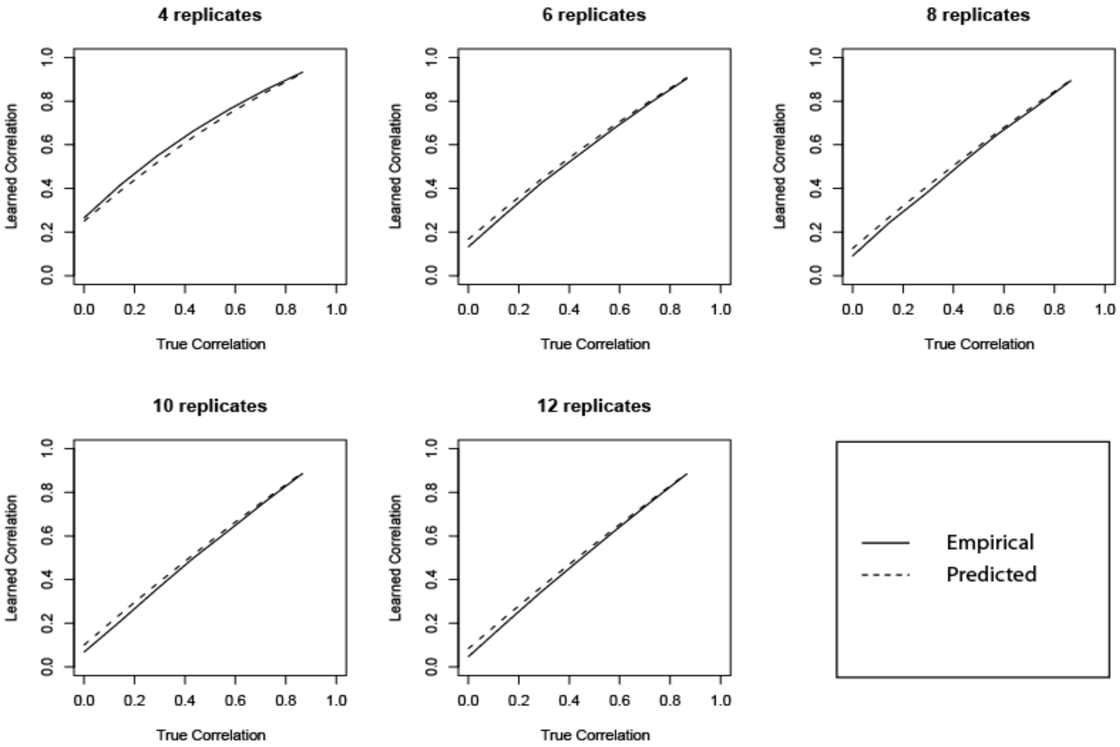
In summary, we have presented a method for correcting gene-specific dye bias with a maximum likelihood model and test for differential expression. This method can effectively learn the parameters of the systematic bias without the need for additional control microarray experiments. An implementation of this algorithm is freely available at http://cellcircuits.org/VERA/.

Supplemental Figures



Supplemental Figure 4.1

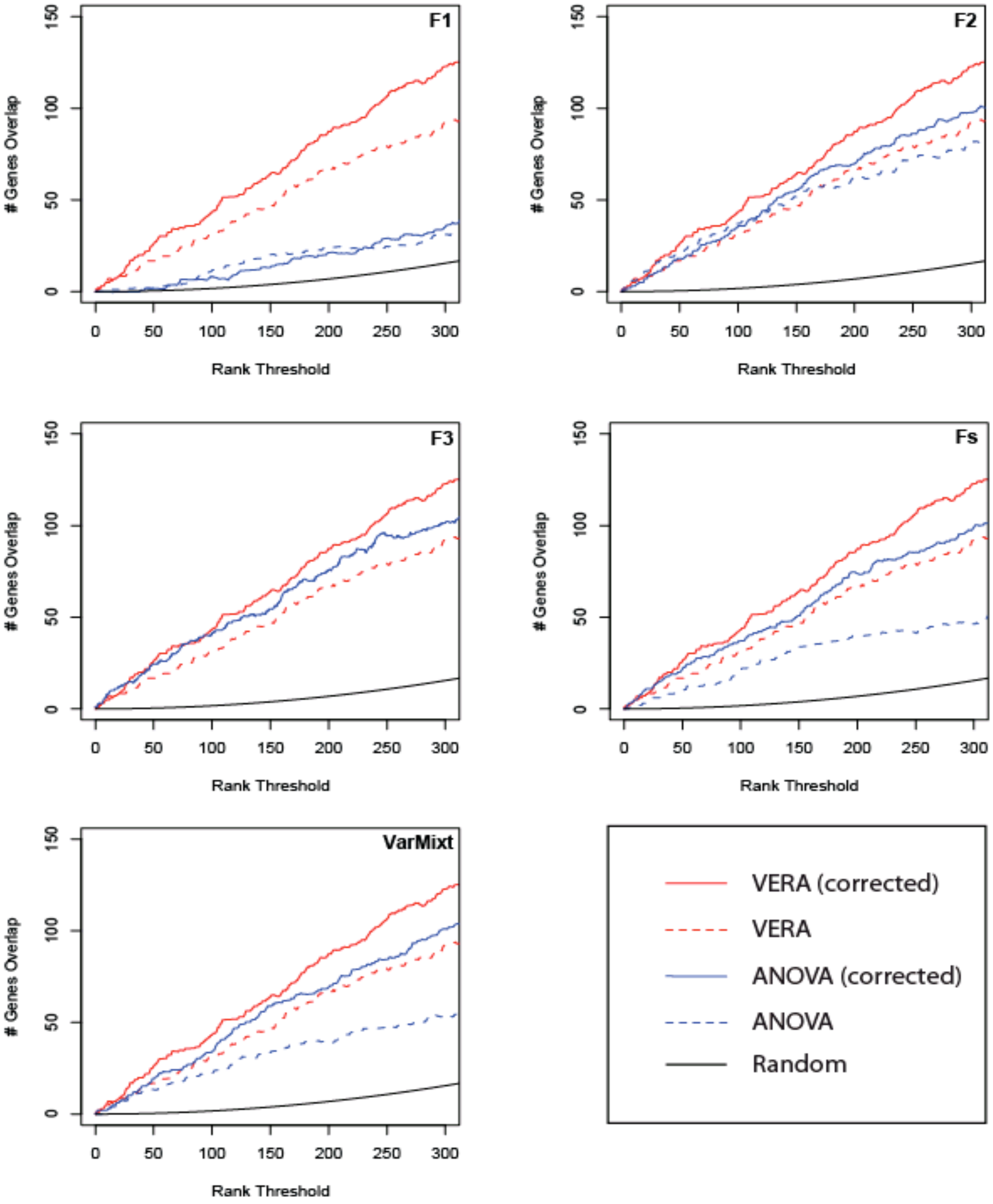
As is often true of maximum-likelihood procedures, parameter estimation is biased due to small sample size effects. Specifically, the correlation of multiplicative error is overestimated. Using simulated data, the relationship between true and learned correlation is plotted for several different combinations of variance parameters (as indicated in the individual plot titles). For all plots, data are simulated for four array replicates. Relative to changes in the sample size (Supplementary Figure S2), the relationship between estimated and true correlation is largely invariant for different selections of variance parameters.



Supplemental Figure 4.2

The relationship between estimated and true correlation is dependent only on the number of array replicates. Using simulated data, we determined this relationship for various number of array replicates. From these results, we can see that the relationship is well-approximated by the equation (1), where *n* is the number of replicate arrays. Solid lines represent the empirically determined relationship, while the approximation of equation (1) is shown with a dashed line.

$$\rho_{estimated} = -\frac{1}{n}\rho_{real}^{2} + \rho_{real} + \frac{1}{n} \qquad (1)$$



Supplemental Figure 4.3

The dCTP- versus dUTP-labeled expression data are compared for different analysis methods. Since the true number of differentially expressed genes is unknown, the calculation is performed over a range of values (x-axis). The y-axis shows the number of genes which pass the rank threshold in both labeling approaches after correcting for any bias in the method. For the ANOVA comparison, the test statistic used to determine differential expression is indicated in the upper right hand corner

Acknowledgements

We thank Dr. John Quackenbush for motivating discussions. We gratefully acknowledge the NSF (0425926) and NIH (GM070743) for funding this work. TI is a fellow of the David and Lucille Packard Foundation.

Chapter 4, in full, is a reprint of the following work,

Kelley, R., Feizi H., Ideker T. Correcting for gene-specific dye bias in DNA microarrays using the method of maximum likelihood. **Bioinformatics** (2007).

The dissertation author was the sole first author on this paper, responsible for designing, performing, and analyzing experiments and algorithms.

Chapter 5. Integrating phenotypic and expression profiles to map arsenic-response networks

Abstract

Background

Arsenic is a nonmutagenic carcinogen affecting millions of people. The cellular impact of this metalloid in *Saccharomyces cerevisiae* was determined by profiling global gene expression and sensitivity phenotypes. These data were then mapped to a metabolic network composed of all known biochemical reactions in yeast, as well as the yeast network of 20,985 protein-protein/protein-DNA interactions.

Results

While the expression data unveiled no significant nodes in the metabolic network, the regulatory network revealed several important nodes as centers of arsenic-induced activity. The highest-scoring proteins included Fhl1, Msn2, Msn4, Yap1, Cad1 (Yap2), Pre1, Hsf1 and Met31. Contrary to the gene-expression analyses, the phenotypic-profiling data mapped to the metabolic network. The two significant metabolic networks unveiled were shikimate, and serine, threonine and glutamate biosynthesis. We also carried out transcriptional profiling of specific deletion strains, confirming that the transcription factors Yap1, Arr1 (Yap8), and Rpn4 strongly mediate the cell's adaptation to arsenic-induced stress but that Cad1 has negligible impact.

Conclusions

By integrating phenotypic and transcriptional profiling and mapping the data onto

the metabolic and regulatory networks, we have shown that arsenic is likely to channel sulfur into glutathione for detoxification, leads to indirect oxidative stress by depleting glutathione pools, and alters protein turnover via arsenation of sulfhydryl groups on proteins. Furthermore, we show that phenotypically sensitive pathways are upstream of differentially expressed ones, indicating that transcriptional and phenotypic profiling implicate distinct, but related, pathways.

Background

Global technologies in the budding yeast Saccharomyces cerevisiae have changed the face of biological study from the investigation of individual genes and proteins to a systems-biology approach involving integration of global gene expression with protein-protein and protein-DNA information[103]. These data, when combined with phenotypic profiling of the deletion mutant library of nonessential genes, allow an unparalleled assessment of the responses of yeast to environmental stressors[104-106]. In this study, we used these two genomic approaches to study the response of yeast to arsenic, a toxicant present worldwide, affecting millions of people[107].

Arsenic, a ubiquitous environmental pollutant found in drinking water, is a metalloid and human carcinogen affecting the skin and other internal organs[108]. It is also implicated in vascular disorders, neuropathy, diabetes and as a teratogen[109]. Furthermore, arsenic compounds are also used in the treatment of acute promyelocytic leukemia[110-112]. Consequently, the potential for future secondary tumors resulting from such therapy necessitates an understanding of the mechanisms of arsenic-mediated toxicity and carcinogenicity. However, even though a number of arsenic-related genes and processes related to defective DNA repair, increased cell proliferation and oxidative stress have been described, the exact mechanisms of arsenic-related disease remain elusive[112-121]. This is, in part, due to the lack of an acceptable animal model that faithfully recapitulates human disease[115].

A number of proteins involved in metalloid detoxification have been described in different organisms, including *Saccharomyces cerevisiae*. Bobrowicz *et al.*[122] found that Arr1 (also known as Yap8 and which is a member of the YAP family that shares a conserved bZIP DNA-binding domain) confers resistance to arsenic by directly or indirectly regulating the expression of the plasma membrane pump Arr3 (also known as Acr3), another mechanism for arsenite detoxification of yeast in addition to the transporter gene, YCF1[123]. Arr3 is 37% identical to a *Bacillus subtilis* putative arsenic-resistance protein and encodes a small (46 kilodalton (kDa)) efflux transporter that extrudes arsenite from the cytosol[124, 125]. Ycf1, on the other hand, is an ATP-binding cassette protein that mediates uptake of glutathione-conjugates of AsIII into the vacuole[123, 124]. Until recently, very little was known about arsenic-specific transcriptional regulation of detoxification genes. Wysocki *et al.*[126]found that Yap1 and Arr1 (called Yap8 in their paper) are not only required for arsenic resistance, but that Arr1 enhances the expression of Arr2 and Arr3 while Yap1 stimulates an antioxidant response to the metalloid. Menezes *et al.*[127], on the other hand, found that arsenite-induced expression of Arr2 and Arr3, as well as Ycf1, is likely to be regulated by both Arr1 (called Yap 8 in their paper) and Yap1.

Although Arr1 and Yap1 seem specifically suited for arsenic tolerance, the other seven YAP-family proteins are still worthy of investigation in light of the fact that each one regulates a specific set of genes involved in multidrug resistance with overlaps in downstream targets. One such interesting protein is Cad1 (Yap2). Although Yap1 and Cad1 are nearly identical in their DNA-binding domains, Yap1 controls a set of genes (including Ycf1) involved in detoxifying the effects of reactive oxygen species, whereas Cad1 controls genes that are over-represented for the function of stabilizing proteins in an oxidant environment[128]. However, Cad1 also has a role in cadmium resistance. As arsenic has metal properties, it is conceivable that Cad1 might play a greater part in arsenic tolerance and perhaps more so than the oxidative-stress response gene, YAP1.

Understanding the role of AP-1-like proteins (such as YAP family members) in metalloid tolerance was one of the goals in this study within the realm of the larger objective - using an integrative experimental and computational approach to combine gene expression and phenotypic profiles (multiplexed competitive growth assay) with existing high-throughput molecular interaction networks for yeast. As a consequence we uncovered the pathways that influence the recovery and detoxification of eukaryotic cells after exposure to arsenic. Networks were analyzed to identify particular network regions that showed significant changes in gene expression or systematic phenotype. For each data type, independent searches were performed against two networks: the network of yeast protein-protein and protein-DNA interactions, corresponding to signaling and regulatory effects (the regulatory network); and the network of all known biochemical reactions in yeast (the metabolic network). For the gene-expression analysis, we found several significant regions in the regulatory network, suggesting that Yap1 and Cad1 have an important role. However, no significant regions in the metabolic network were found. In order to test the functional significance of Yap1 and Cad1, we used targeted gene deletions of these and other genes, to test a specific model of transcriptional control of arsenic responses.

In contrast to the gene-expression data, the phenotypic profile analysis revealed no significant regions in the regulatory network, but two significant metabolic networks. Furthermore, we found that phenotypically sensitive pathways are upstream of differentially expressed ones, indicating that metabolic pathway associations can be discerned between phenotypic and transcriptional profiling. This is the first study to show a relationship between transcriptional and phenotypic profiles in the response to an environmental stress.

Results and discussion

Transcript profiling reveals that arsenic affects glutathione, methionine, sulfur, selenoamino-acid metabolism, cell communication and heat-shock response

Before gene-expression analysis of arsenic responses in *S. cerevisiae*, we performed a series of dose-response studies. We found that treatment of wild type cells with 100 μM and 1 mM AsIII had a negligible effect on growth, but that these cells still exhibited a pronounced transcriptional response (see Supplemental Figure 5.1 and Supplemental Figure 5.2). Microarray analysis of biological replicates (four chips per replicate experiment) of the high-dose treated cells (1 mM AsIII) clustered extremely well together when using Treeview (see Materials and methods, and Supplemental Figure 5.2). The lower dose time-course (100 μM AsIII) showed the beginning of gene-expression changes at 30 minutes, with the robust changes occurring at 2 hours, or one cell division (see Supplemental Figure 5.2). The 2 hour, 100 μM dose clustered together with the 30 minute, 1 mM biological replicates and was in fact so similar to them that an experiment of one set of four chips for the 2 hour lower dose was deemed sufficient. Furthermore, when combining the three datasets (2 hour, 100 μM AsIII and each 30 minute, 1 mM AsIII replicate data) and using a 95% confidence interval (see Materials and methods) we found 271 genes that were not only statistically significant in at least 75% of the total data (9 out of 12 chips), but also that the direction and level of expression of these genes were similar between the datasets. The lower dose time-course also included a 4 hour treatment, or two cell divisions. This experiment demonstrated the greatest degree of variability, indicating either a cycling effect or the cell's return to homeostasis, which was further exemplified by a decrease in the transcriptional response (see Supplemental Figure 5.2).

Genes were categorized by Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Simplified Gene Ontology (biological process, cellular component and molecular function) (Table 5.1). In total, 829 genes out of 6,240 had significantly altered expression (see Materials and methods) in at least one experimental condition. The categories significantly enriched for differentially expressed genes in the KEGG pathways were glutathione, methionine, sulfur and selenoamino-acid metabolism, and in the Simplified Gene Ontology (biological process), cell communication and heat-shock response (Table 5.1).

Network mapping of transcript profiling data finds a stress-response network involving transcriptional activation and protein degradation

We used the Cytoscape network visualization and modeling environment together with the ActiveModules network search plug-in to carry out a comprehensive search of the regulatory and metabolic networks[112, 129]. The former consists of the complete yeast-interaction network of 20,985 interactions, in which 5,453 proteins are connected into circuits of protein-protein or protein-DNA interactions[130, 131]. For each protein in this network, we defined a network neighborhood containing the protein and all its directly interacting partners. In the metabolic network, based on a reconstruction by Forster *et al.*[132] with 2,210 metabolic reactions and 584 metabolites, nodes represent individual reactions and edges represent metabolites. A shared metabolite links two reactions. We searched for sequences of related reactions governed by sensitive proteins (enzymes) in the phenotypic profiling data. To aid visualization, these sequences of reactions were combined to create metabolic pathways. We then identified the neighborhoods associated with significant changes in expression using the ActiveModules plug-in. This process resulted in the identification of seven significant neighborhoods in the regulatory network, centered on nodes Fhl1, Pre1, Yap1, Cad1, Hsf1, Msn2 and Msn4 (Figure 5.1). Together these neighborhoods narrow the significant data to 20% of the genes with the most significant changes in expression across one or more arsenic conditions (see Materials and methods and Supplemental Figure 5.2). We did not find the emergence of any significant neighborhoods in the metabolic network.

The highest-scoring regulatory network neighborhood was defined by the transcription factor Fhl1 (Figure 5.1a). Its expression did not change significantly, but it was the highest-scoring node as judged by the significant expression changes observed for its surrounding neighborhood. Fhl1 controls a group of proteins important for nucleotide and RNA synthesis, as well as the synthesis and assembly of ribosomal proteins[133] which, from our data, are downregulated by arsenic exposure. Downregulation of ribosomal proteins in response to environmental stress has been reported previously[134, 135], but to our knowledge this is the first association of Fhl1 as a key control element in this process. It seems likely that the repression of de novo protein synthesis in response to arsenic allows energy to be diverted to the increased expression of genes involved in stress responses and protection of the cell.

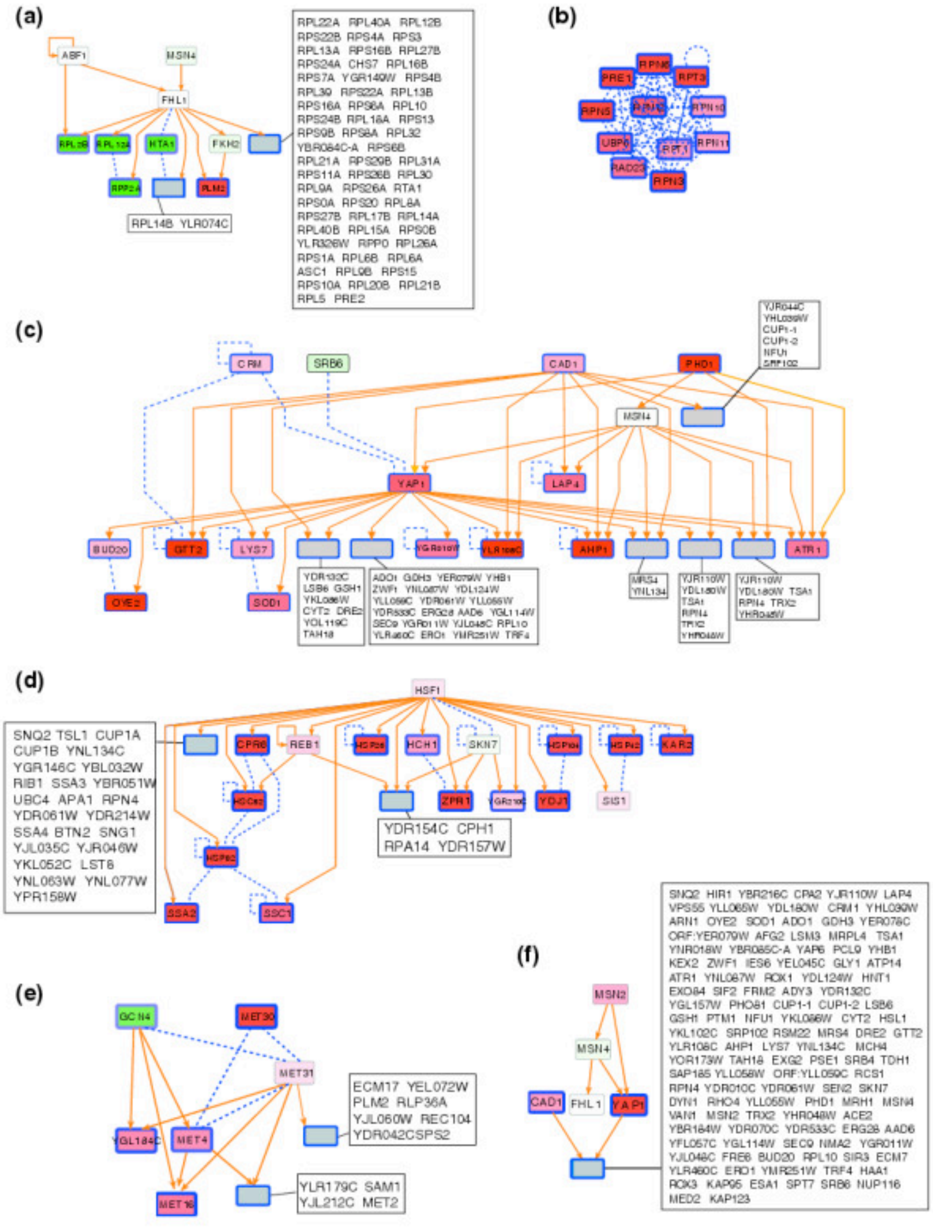


Figure 5.1. Arsenic-induced signaling and regulatory mechanisms involve transcriptional activators and the proteasome.

(a-d) Significant network neighborhoods ($p < 0.005$) uncovered by the ActiveModules algorithm, with the search performed at depth 1 (all nodes in the network are the nearest neighbors of one central node): (a) FHL1 center; (b) PRE1 center and proteasome complex; (c) YAP1 and CAD1 centers; (d) HSF1 center. (e) An additional network centered on MET31 with functional relevance to the arsenic response, which, however, did not reach significance in this analysis, $p < 0.11$. (f) An overview of the network relationships between major arsenic-responsive transcription factors. Shades of red, induced; shades of green, repressed; blue boxed outline, significant expression; orange arrows, protein-DNA interaction; blue dashed lines, protein-protein interactions. The 2 h, 100 μM AsIII condition was used for the visual mappings. Many of the genes mapped to the network neighborhoods and displayed in this figure are boxed for the sake of clarity and space, but are mostly significantly differentially expressed.

One such pathway may involve sulfur metabolism, which leads to glutathione synthesis. In fact, included in Figure 5.1 is Met31 (Figure 5.1e), a transcriptional regulator of methionine metabolism, which interacts with Met4, an important activator of the sulfur-assimilation pathway that is probably involved in the glutathione-requiring detoxification process. While the differential expression of this neighborhood was not strictly significant according to ActiveModules (see Materials and methods), it has high biological relevance in light of the statistically significant alteration in expression categorized using KEGG pathways (Table 5.1).

Another high-scoring neighborhood comprises part of the proteasome protein complex (Figure 5.1b). The components of the proteasome are likely to be upregulated to meet the increased demand for protein degradation brought about by the binding of AsIII to the sulfhydryl groups on proteins and/or glutathione that subsequently interfere with numerous enzyme systems such as cellular respiration[109, 115]. In this paper, we will propose that this occurs through indirect oxidative stress as a result of the depletion of glutathione.

The role of transcription factors Yap1 and Cad1 and the metalloid stress response

Many of the central proteins in the significant neighborhoods uncovered by ActiveModules were transcription factors (Figure 5.1a,c-f). Although some of these proteins were not differentially expressed themselves, they were still high-scoring nodes because of the highly significant expression of their targets. This is also important to keep in mind as we discuss later which genes might be sensitive to arsenic, but not necessarily differentially expressed, and why many genes that are differentially expressed do not display sensitive phenotypes when deleted.

Transcription factors Msn2, Yap1, Msn4, Cad1 and Hsf1 were the central proteins for many of the significant neighborhoods found (Figure 5.1c,d,f). Together with several genes previously implicated in oxidative-stress responses, these neighborhoods compose a stress-response network[126, 128, 136-140]. Of particular interest are Yap1 and Cad1, because of the high number of shared downstream targets (Figure 5.1c,f).

When overexpressed, Yap1 confers resistance to several toxic agents, and Yap1 mutants are hypersensitive to oxidants[134, 141-145]. Conversely, Cad1 responds strongly to cadmium, but not to hydrogen peroxide (H_2O_2)[128, 138]. Following arsenic exposure, Yap1 is induced at least fourfold, with many of its downstream targets showing high levels of induction. Several of its targets are among the most highly upregulated genes (as high as 178-fold for OYE3 (encoding a NADPH dehydrogenase)). Moreover, Yap1 regulates GSH1, which encodes γ-glutamylcysteine synthetase (an enzyme involved in the biosynthesis of antioxidant glutathione), TRX2 (the antioxidant thioredoxin), GLR1 (glutathione reductase) and drug-efflux pumps ATR1 and FLR1[138, 146-151]. It should be noted that GSH1 and ATR1 are examples of several genes also targeted by Cad1. All of these specified Yap1 targets are induced after arsenic exposure, recapitulating the toxicant's role as a likely oxidant. During the course of this work, Wysocki *et al.*[126] also implicated Yap1 in arsenic tolerance.

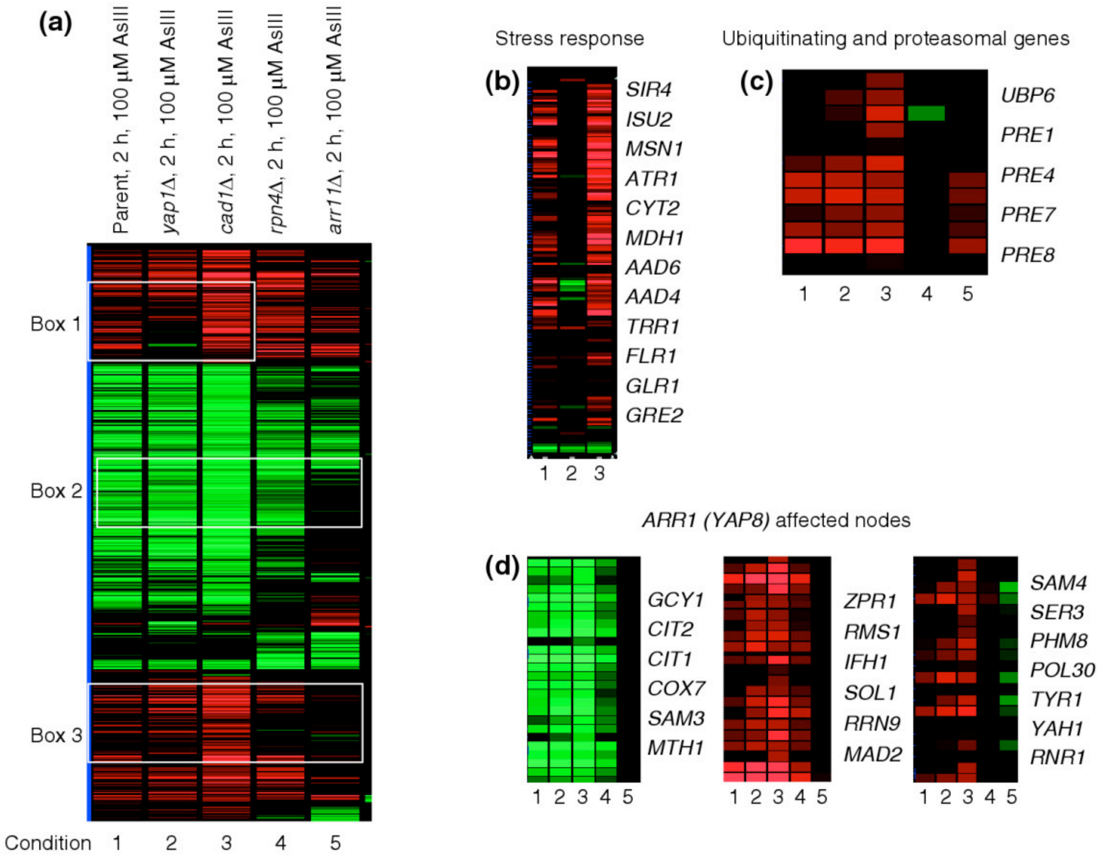


Figure 5.2. Yap1 but not Cad1 is important for mediating the cell's adaptation to arsenic.

(a) Self-organized heat map (dendograms were removed and boxes 1-3 indicate specific clusters) of 6,172 genes selected from the various indicated conditions. AsIII-treated parent wild type strain with normalized data values that are greater or less than those in condition(s) knocked-out Yap1, Cad1, Rpn4, or Arr1 treated with AsIII, by a factor of twofold. All knockouts tested revealed altered profiles compared to the wild type, except for cad1Δ. (b) yap1Δ (condition 2) loses induced expression of stress response genes found in box 1, such as SIR4, ISU2, MSN1, ATR1, CYT2, MDH1, AAD6, AAD4, TRR1, FLR1, GLR1 and GRE2. (c) rpn4Δ (condition 4) loses induced expression of ubiquitinating and proteasomal genes found in box 3 - UBP6, PRE8, PRE4, PRE7 and PRE1. (d) arr1Δ (condition 5) loses repressed expression of sulfur amino-acid metabolism gene SAM3 and glutamate biosynthesis gene CIT2, among others (box 2). arr1Δ also loses induced expression of serine biosynthesis gene SER3, sulfur amino-acid metabolism gene SAM4, cell-cycle regulator ZPR1, spindle-checkpoint subunit MAD2, ribonucleotide reductase RNR1and RNA polymerase I transcription factor RRN9, to name a few (box 3). Red, induced; green, repressed.

As Cad1 and Yap1 share many downstream targets, the genes defined by these transcription factors are very similar. To determine which transcription factor is playing the most active role in the high level of differential expression for this group (see Figure 5.1c,f), we tested the roles of both activators by treatment of yap1Δ and cad1Δ deletion strains with 100 µM AsIII for 2 hours. Surprisingly, we did not find that Cad1 was involved in regulation in response to arsenic-mediated stress. The yap1Δ strain was not only sensitive to AsIII by phenotypic profiling but also defective in the induction of several downstream enzymes with antioxidant properties (Figure 5.2a,b). Conversely, the cad1Δ strain displayed an almost identical profile to wild type, eliminating it as a strong factor in the arsenic response (Figure 5.2a,b).

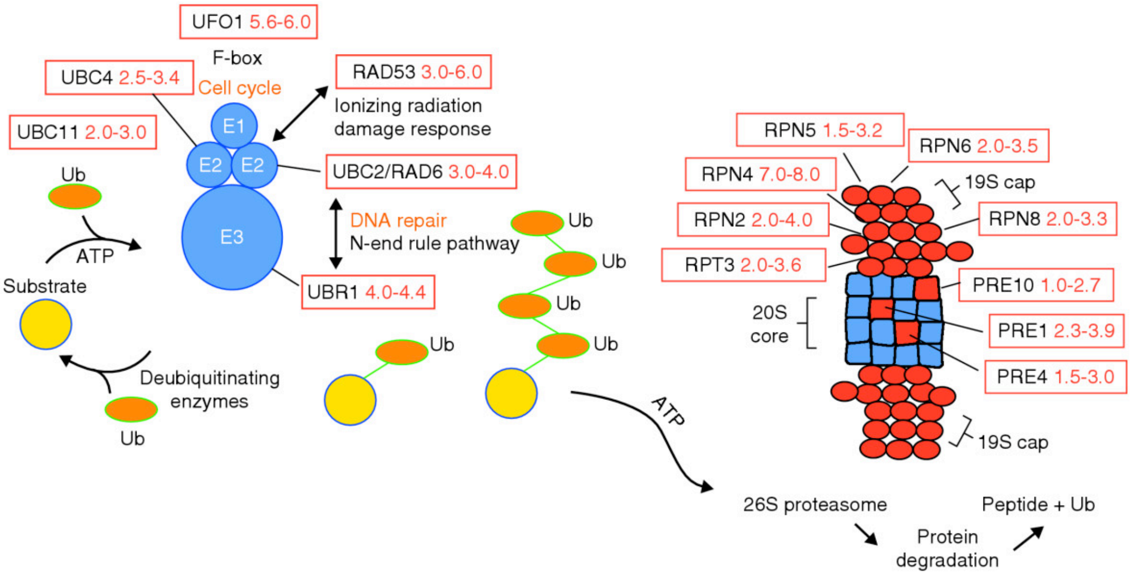


Figure 5.3. The ubiquitin (Ub) and proteasome system responds to arsenic-mediated toxicity.

S. cerevisiae ubiquitin and proteasome pathways show differential expression in a number of key genes, including that for the proteasomal activator RPN4. Induction is denoted by red boxes with fold-change ranges representing the 2 h, 100 µM AsIII and 0.5 h, 1 mM AsIII experiments, respectively.

The proteasome responds to arsenic, and Rpn4 mediates a transcriptional role

Treatment of yeast with as little as 100 µM AsIII for 2 hours resulted in the induction of at least 14 ubiquitin-related and proteasome gene products (Figure 5.1b and Figure 5.3). The eukaryotic proteasome consists of a 20S protease core and a 19S regulator complex, which includes six AAA-ATPases known as regulatory particle triple-

A proteins (RPT1-6p)[152, 153]. Proteins are targeted for degradation by the proteasome via the covalent attachment of ubiquitin to a lysine side chain on the target protein (Figure 5.3). Conjugating enzymes then function together with ubiquitin-ligase enzymes to adhere to the target protein, and are tailored to carry out specific protein degradation in DNA repair, growth control, cell-cycle regulation, receptor function and stress response, to name a few[154, 155]. The apparent importance of Yap1 in response to possible oxidative damage by arsenic indicated a potential role for Rpn4 (induced eightfold, Figure 5.3). This is a 19S proteasome cap subunit, which also acts as a transcriptional activator of the ubiquitin-proteasome pathway and a variety of base-excision and nucleotide-excision DNA repair genes[130, 135, 156].

Rpn4 is required for tolerance to cytotoxic compounds and may regulate multidrug resistance via the proteasome[157]. Moreover, Owsianik *et al.*[157] identified an YRE (Yap-response element) site present in the RPN4 promoter. This YRE was found to be functional and important for the transactivation of RPN4 by Yap1 in response to oxidative compounds, such as H_2O_2. However, we also located the Rpn4-binding sequence, TTTTGCCACC, 47 bases distant from the open reading frame (ORF) of YAP1, indicating that Yap1 not only activates Rpn4, but that Rpn4 may in fact activate Yap1[158]. In support of this hypothesis we found that relative to wild type, the level of Yap1 induction was lower in the rpn4Δ strain under arsenic stress conditions, whereas Rpn4 was equally induced in the yap1Δ strain (Supplemental Figure 5.3).

With respect to wild type, the profile of rpn4Δ after treatment with arsenic was the most dramatically altered, save for arr1Δ (Figure 5.2). These data suggest that arsenic modification of sulfhydryl groups on proteins leads to protein inactivation and therefore degradation via the 26S proteasome. Another scenario is that the proteasome, and/or its proteases, is sensitive to arsenic-related events, leading to dysfunctional protein turnover and an increased requirement for 26S proteasome subunits. A similar idea was proposed for the direct methylating agent, methylmethane sulfonate[135].

ARR1 transcriptional responses

Arr1 is structurally related to Yap1 and Cad1[122, 126]. However, little is known about how Arr1 may be involved in oxidative stress and/or multidrug resistance. Furthermore, Arr1 is not well represented by the interactions present in the yeast regulatory network. However, studies by Bobrowicz *et al.*[122, 159] show that the transcriptional activation of Arr3 requires the presence of the Arr1 gene product. Moreover, a report by Bouganim *et al.*[160] supports our finding that Yap1 also is important for arsenic resistance. They show that overproduction of Yap1 blocks the ability of Arr1 to fully activate Arr3 expression at high doses of arsenite, suggesting that Yap1 can compete for binding to the promoter of the Arr1 target gene, ARR3. While this paper was being written, Tamas and co-workers[126] showed that Arr1 transcriptionally controls Arr2 and Arr3 expression from a plasmid containing their promoters fused to the lacZ gene and measuring β-galactosidase activities. This was done by growing the cells for 20 hours with a low dose of metalloid and spiking the concentration to 1 mM AsIII for the last 2 hours of incubation. These experiments showed that ARR1 deletion resulted in complete loss of Arr3-lacZ induction, whereas YAP1 deletion did not significantly affect induction. Similar results were obtained for the Arr2-lacZ induction assay and the authors concluded that Yap1 has a role in metalloid-dependent activation of oxidative stress response genes, whereas the main function of Arr1 seems linked to the control of Arr2 and Arr3. Interestingly, this study was shortly followed by another from Menezes *et al.*[127] which found contrasting results when looking at mRNA and Northern-blot analysis. In this study, the induction of Arr2 and Arr3, after treatment with 2 mM AsIII for up to 90 minutes, did not occur in either the ARR1-deleted strain or the YAP1-deleted strain. These authors conclude that the requirement for both YAP1 and ARR1 is vital to yeast in the function of regulating and inducing genes important for arsenic detoxification. Finally, transcription profiling experiments presented here show that the arsenic transport proteins Arr2 and Arr3 are still expressed (2.9-fold induction for Arr2 and 1.8-fold for Arr3, respectively) in the ARR1 mutant, but show defective induction in the yap1Δ strain treated in parallel (Supplemental Figure 5.3). These results indicate that Yap1 may control Arr2 and Arr3 when yeast is subjected to 100 µM AsIII for 2 hours.

Our results and those of Menezes *et al.*[127], in contrast to the results of Tamas and colleagues[126], might be explained by the following. Our and Menezes *et al.*'s studies looked at genes in the normal chromosome context rather than genes ectopically expressed from a plasmid; in addition, in our study, we treated the yeast with 100 µM AsIII while Wysocki *et al.*[126] started with a low dose, but spiked the concentration to 1 mM AsIII in the last 2 hours of incubation. However, Menezes *et al.*[127] used an even higher dose (2 mM AsIII for a time-course ending at 90 minutes) and obtained more similar results to ours, with the exception that their Northern-blot analysis, which can sometimes miss relatively small changes, indicated an apparent lack of induction of ARR2 or ARR3 in either the ARR1- or YAP1-deleted strains. Taken together, these data indicate that both ARR1 and YAP1 are important genes involved in the process of arsenite detoxification in the yeast cell, but because of the different strains and treatment protocols used between these three studies, further experiments are warranted to resolve the differences.

Other interesting results from our transcription profiling of the arr1Δ and parent strains after arsenic treatment (Figure 5.2a,d), included large differences in expression as a whole and in particular the inability of arr1Δ to induce serine biosynthesis-related genes such as SER3, and sulfur and methionine amino-acid metabolism genes including SAM4. Conversely, arr1Δ failed to repress SAM3, as well as CIT2, a glutamate biosynthesis gene, when compared to the parent profile.

These observations indicate that Arr1 may regulate sulfur-assimilation enzymes that are necessary for arsenic detoxification. This is particularly interesting considering that the ActiveModules algorithm identified the node Met31 (Figure 5.1e), the transcriptional regulator of methionine metabolism which interacts with Met4, an important activator of the sulfur-assimilation pathway that is likely to be involved in the glutathione-requiring detoxification process. Sulfur metabolism was also a functional category in the Simplified Gene Ontology found to be significantly enriched by the hypergeometric statistical test (see Materials and methods) (Table 5.1). Furthermore, phenotypic profiling results discussed later show the importance of serine and glutamate metabolism in the sensitivity response to arsenic. Lastly, it is important to note that arr1Δ also displays loss of expression of a number of ubiquitin-proteasome-related gene products, sharing similar expression patterns with rpn4Δ and suggesting that it may have a role in protein degradation as well.

Arsenic treatment stimulates cysteine and glutathione biosynthesis and leads to indirect oxidative stress

Our arsenic-treatment experiments revealed the strong induction of over 20 enzymes in the KEGG sulfur amino acid and glutathione biosynthesis pathways (Table 5.1). This is consistent with the hypothesis that glutathione acts as a first line of defense against arsenic by sequestering and forming complexes with the toxic metalloid[123].

Dormer *et al.*[161] showed that GSH1 induction by cadmium is dependent on the presence of Met4, Met31, Met32 and Cbf1 in the transcriptional complex of MET genes. Met4 and Met32 are also differentially expressed in response to arsenic and interact with Met31, which defines a network neighborhood as shown in Figure 5.1e. The biological impact of the sulfur-related stress response was further exemplified by comparisons of our arsenic profiles to H2O2 profiles (400 μM H_2O_2) from Causton *et al.*[162] (Table 5.2). Although we found many expected similarities between arsenic and H_2O_2 gene-expression profiles in regard to oxidative-stress response genes, sulfur and methionine metabolism genes, in response to H_2O_2, were either repressed or did not change (Table 5.2). Furthermore, a study by Fauchon *et al.*[163] showed that yeast cells treated for 1 hour with 1 mM of the metal Cd2+, responded by converting most of the sulfur assimilated by the cells into glutathione, thus reducing the availability of sulfur for protein synthesis. Our arsenic profile showed a similar response to the sulfur-assimilation profile seen with Cd2+ (Table 5.2). As a consequence, arsenic may be conferring indirect rather than direct oxidative stress mediated by the depletion of glutathione, thus inhibiting the breakdown of increasing amounts of H2O2 by glutathione peroxidase (GPX2, up 13-fold) (Figure 5.4)[123, 164].

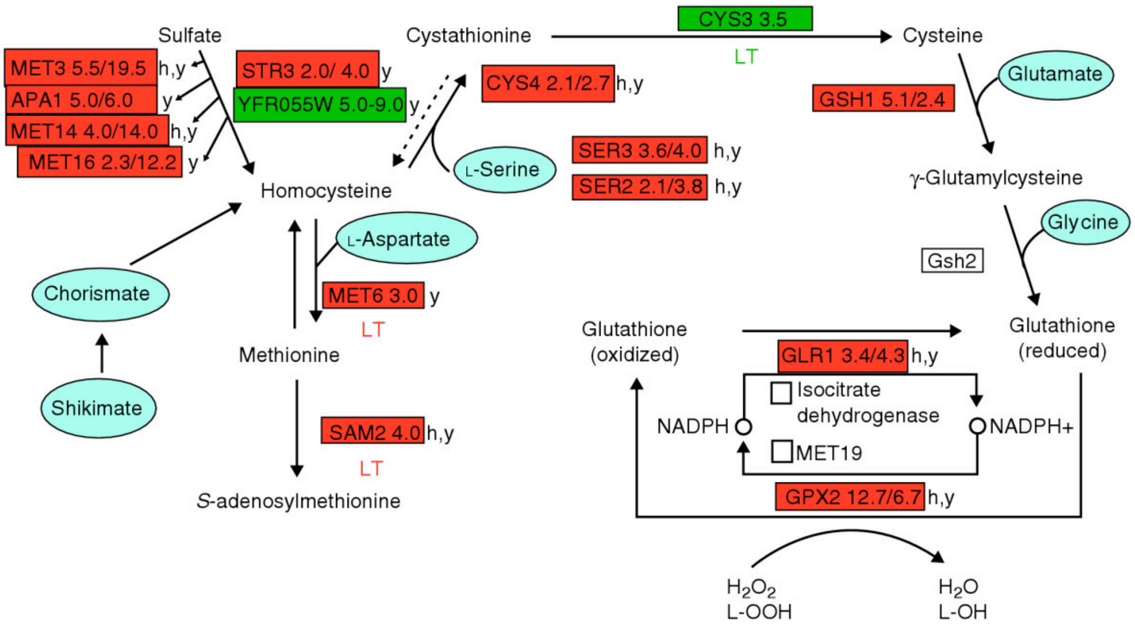


Figure 5.4. Gene-expression profiling links sulfur assimilation, methionine and glutathione pathways.

Selected genes in these pathways are represented as red for induced (2 h, 100 µM AsIII and 0.5 h, 1 mM AsIII, respectively) and green for repressed. Genes in white boxes are not differentially expressed. The pathways in the blue ovals are upstream of methionine, cysteine and glutathione, and are sensitive to arsenic. The downstream pathways employ numerous redundant enzymes that are differentially expressed, but are not sensitive. LT, late time-point, 4 h, 100 µM AsIII experiment; h, human; y, yeast.

Phenotypic profiling defines arsenic-sensitive strains and maps to the metabolic network

To identify genes and pathways that confer sensitivity to arsenic, we identified deletion mutants with increased sensitivity to growth inhibition using a deletion mutant library of nonessential genes (4,650 homozygous diploid strains)[35, 165]. Each strain contains two unique 20-bp sequences (UPTAG and DOWNTAG) enabling their growth to be analyzed en masse and the fitness contribution of each gene to be quantitatively assayed by hybridization to high-density oligonucleotide arrays. The top 50 sensitive deletion strains included: THR4, SER1, SER2, CPA2, CPA1, HOM2, HOM3, HOM6, ARG1, YAP1, CDC26, ARR3, CIN2, ARO1, ARO2 and ARO7.

Only 10% of the top 50 sensitive mutant strains were significantly differentially expressed in the transcript profile. This lack of direct correlation between gene expression and fitness data is consistent with data from our own and other laboratories[104, 105, 165]. At least three factors may contribute to this discrepancy. First, some highly expressed genes when deleted are nonviable (around 1,000 genes) and are therefore unable to be scored for fitness. Some examples of highly expressed, yet nonviable, genes under arsenic stress are ERO1 (7- to 10-fold induced), HCA4 (5- to 9-fold induced), and DCP1 (9- to 22-fold induced). Second, there are redundant pathways mediated by multiple genes, such that deletion of one does not lead to sensitivity. OYE2, OYE3, and a large number of reductases fall into this category. Finally, gene products that do not change significantly, mediate important biological responses and thus when deleted could sensitize the cell to a specific stressor. ARO1, ARO2, THR4 and HOM2 are examples of genes that are not differentially expressed but are very sensitive to arsenic.

Like the gene-expression data, the phenotypic data was subjected to searches performed against the regulatory network of yeast protein-protein and protein-DNA interactions as well as the metabolic network of all known biochemical reactions in yeast. Unlike the transcription profile, the phenotypic data analysis revealed no significant regions in the regulatory network, but did map to two statistically significant metabolic networks. The first significant pathway was amino acid synthesis/degradation with the terminal products being L-threonine and L-homoserine, beginning with precursors such as L-arginine, fumarate and oxaloacetate (Figure 5.5a).

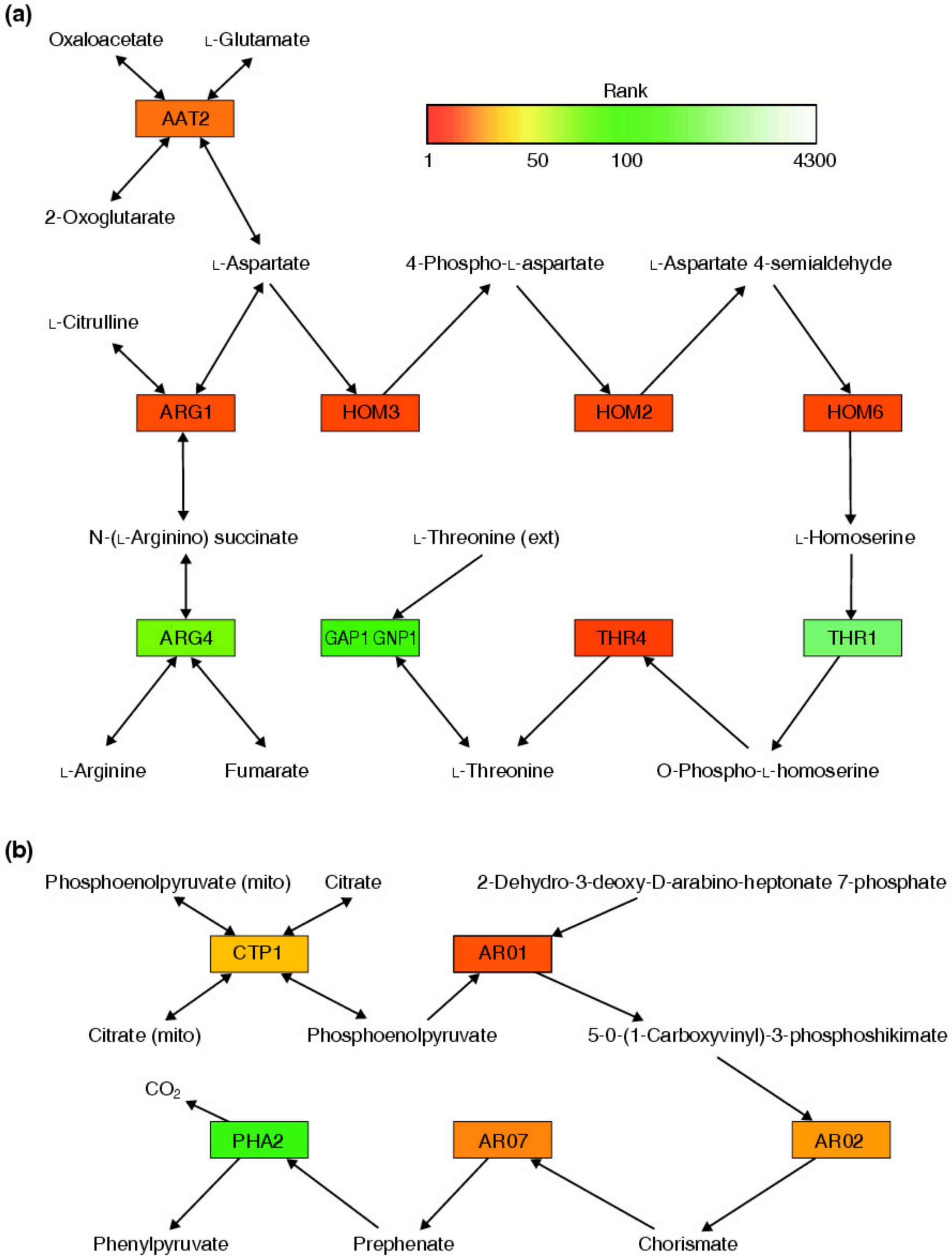


Figure 5.5. LinearActivePaths analysis finds that virtually all genes in active metabolic networks confer sensitivity to arsenic when deleted.

(a) Serine, threonine, glutamate amino-acid synthetic pathways; (b) the shikimate pathway. The paths that compose these networks all have individual p-values of < 0.05. The coloration for these figures is based on red for any gene ranked in the top 50 significant genes, yellow for 51-100, and green for >101.

These products function in serine, threonine and glutamate metabolism. The second network indicated the importance of the shikimate pathway, which is essential for the production of aromatic compounds in plants, bacteria and fungi (Figure 5.5b). The shikimate pathway operates in the cytosol of yeast and utilizes phosphoenol pyruvate and erythrose 4-phosphate to produce chorismate through seven catalytic steps. It is a pathway with multiple branches, with chorismate representing the main branch point, and various branches giving rise to many end products. Interestingly, chorismate is also used for the production of ubiquinone, p-aminobenzoic acid (PABA) and folates, which are donors to homocysteine[166-168].

Relationship between gene-expression and phenotypic profiles

Combining transcript profiling and phenotypic profiling provides deeper insights into the biology of arsenic responses. Until now there has been a lack of correlation between the differential expression of genes and sensitivity of deletion mutants[104, 106, 165] and this was the case in the present study. However, by mapping each dataset to the regulatory and metabolic networks, we have uncovered the likely reason for this lack of congruence. Our data show that many of the most sensitive genes are involved in serine and threonine metabolism, glutamate, aspartate and arginine metabolism, or shikimate metabolism, which are pathways upstream of the differentially expressed sulfur, methionine and homocysteine metabolic pathways, respectively. These downstream pathways are important for the conversion to glutathione, necessary for the cell's defense from arsenic (Figure 5.4, Figure 5.5a, Figure 5.6 and Table 5.1). This overlap of sensitive upstream pathways and differentially expressed downstream pathways provides the link between transcriptional and phenotypic profiling data (Figure 5.4 and Figure 5.6).

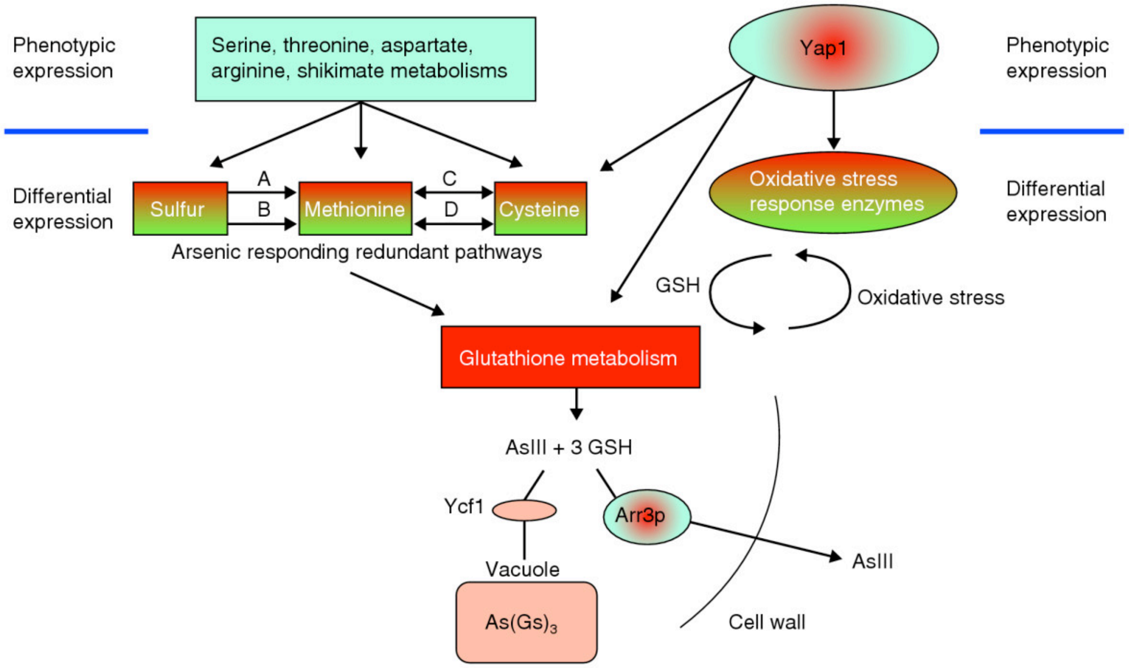


Figure 5.6. Global model of the arsenic response: combining phenotypic data with gene-expression profiles reveals synergistic pathways leading to yeast detoxification mechanisms.

Serine, threonine, aspartate and arginine, as well as shikimate metabolisms, in light blue, represent pathways that are judged as sensitive by phenotypic profiling. Yap1, colored light blue and red, is an example of a transcription factor that is both sensitive and confers induced gene expression. Deletion analysis confirms its role in arsenic-mediated control of the stress response. Red and green represent pathways or genes that are differentially expressed but not sensitive by phenotypic profiling. This schematic diagram demonstrates how the deletion of an individual gene leads to a change in sensitivity if the protein product of that gene is important in a biological process for adaptation to arsenic. On the other hand, expression profiling shows the end product of the cell's response to arsenic. Many of these downstream targets share redundant functions and are not vulnerable in the phenotypic profiling. The expression changes lead to the cell's response to indirect oxidative stress and mechanisms for detoxification. The arrows A, B, C and D represent the multiple branchpoints between redundant pathways. Note that the transport protein, Arr3, which extrudes AsIII out of the cell, is both sensitive and highly differentially expressed.

Thus, we believe our work shows that the deletion of an individual gene can lead to a change in sensitivity to an agent only if the protein product of that gene is important for some process (for example, amino-acid synthesis or a transcription factor required for the increased expression of genes needed to protect against the agent). On the other hand, expression profiling shows the end product of the cell's response to arsenic. Therefore, an agent such as arsenic might cause a transcription factor (Yap1, for example) to increase the expression of as many as 50 genes, 20 of which might help to protect against the agent. However, deletion of any of the 50 would not be expected to have an effect on the response to arsenic. The effect of gene deletion would be on the transcription factor itself (whose expression might not be affected by the agent). Thus, in the case of arsenic exposure, we conclude that phenotypic profiling interrogates genes upstream of the genes that ultimately protect against arsenic toxicity and that the downstream targets that demonstrate differential expression probably share redundant functions and are not vulnerable in the phenotypic profiling (Figure 5.6).

Conclusions

Systems biology represents an important set of methods for understanding stress responses to environmental toxicants, such as arsenic. In this study we have catalogued the centers of activity associated with arsenic exposure in yeast, identifying the key neighborhoods of activity in the regulatory and metabolic networks using the visualization tools and algorithms in Cytoscape. The transcriptional profile mapped to the regulatory network, revealing several important nodes (Fhl1, Msn2, Msn4, Yap1, Cad1, Pre1, Hsf1 and Met31) as centers of arsenic-induced activity. From these results we can conclude that arsenic detoxification in yeast focuses around: nucleotide and RNA synthesis; methionine metabolism and sulfur assimilation; protein degradation; and transcriptional regulation by proteins that form a stress-response network. In summary, protein synthesis in response to arsenic allows energy to be diverted toward the genes channeling sulfur into glutathione, which then leads to indirect oxidative stress by depleting glutathione pools and alters protein turnover. These processes require regulation by transcription factors, the understanding of which we refined by analysis of specific knockout strains. Our experiments, in fact, confirmed that the transcription factors Yap1, Arr1 and Rpn4 strongly mediate the cell's adaptation to arsenic-induced stress but that Cad1 has negligible impact. Finally, contrary to the gene-expression analyses, the phenotypic profiling data mapped to the metabolic network. The two significant metabolic networks unveiled were shikimate and serine, threonine and glutamate biosynthesis. Our goal was to integrate the computational identification of these important pathways found via transcript and phenotypic profiling by regulatory and metabolic network mapping. In doing so, we have shown that genes that confer sensitivity to arsenic are in pathways that are upstream of the genes that are transcriptionally controlled by arsenic and share redundant functions.

Materials and methods

Strains, media and growth conditions

S. cerevisiae strain BY4741 (MATa, his3Δ, leu2Δ0, met15Δ0, uraΔ0) was used and grown in synthetic complete medium at 30°C. Cells were grown to a density of 1 × 107 cells per ml. Cultures were split into two; NaAsO2 (100 µM and 1 mM in two biological repeats) was added to one culture, and both were incubated at 30°C for 0.5, 2 or 4 h. Cells were pelleted and washed in distilled water before RNA extraction. Deletion strains (yap1Δ, cad1Δ, arr1Δ and rpn4Δ) of the same background were obtained from Research Genetics, confirmed and treated the same way, for 2 h and 100 µM NaAsO2.

RNA extraction

For the cDNA hybridization experiments, total RNA was isolated using an acid-phenol method. Pellets were resuspended in 4 ml lysis buffer (10 mM Tris-HCL pH 7.5, 10 mM EDTA, 0.5% SDS). Four milliliters of acid (water-saturated, low pH) phenol was added followed by vortexing. The lysing cell solutions were incubated at 65°C for 1 h with occasional vigorous vortexing and then placed on ice for 10 min before centrifuging at 4°C for 10 min. The aqueous layers were re-extracted with phenol (room temperature, no incubation) and extracted once with chloroform. Sodium acetate was then added to 0.3 M with 2 volumes of absolute ethanol, placed at -20°C for 30 min, and then spun. Pellets were washed two or three times with 70% ethanol followed by Qiagen Poly(A)+ RNA purification with the Oligotex oligo (dT) selection step. Total RNA for the specific knockout strains and parent experiment was isolated by enzymatic reaction, following the RNeasy yeast protocol (Qiagen).

Microarray hybridizations and analyses

A cDNA yeast chip, developed in-house at National Institute of Environmental Health Sciences (NIEHS), was used for gene-expression profiling experiments. A complete listing of the ORFs on this chip is available at http://dir.niehs.nih.gov/microarray/chips.htm. cDNA microarray chips were prepared as previously described[169, 170]. The cDNA was spotted as described[171]. Each poly(A) RNA sample (2 µg) was labeled with Cy3- or Cy5-conjugated dUTP (Amersham) by a reverse transcription reaction using the reverse transcriptase SuperScript (Invitrogen), and the primer oligo(dT) (Amersham). The hybridizations and analysis were performed as described Hewitt *et al.*[172] except that genes having normalized ratio intensity values outside of a 95% confidence interval were considered significantly differentially expressed. Lists of differentially expressed genes were deposited into the NIEHS MAPS database[173]. Genes that were differentially expressed in at least three of the four replicate experiments were compiled and subsequently clustered using the Cluster/Treeview software[174]. GeneSpring (Silicon Genetics) and Cytoscape[39] were used to further analyze and visualize the data.

The knockout experiments were conducted on an Agilent yeast oligo array platform. Samples of 10 μg total RNA were labeled using the Agilent fluorescent direct label kit protocol and hybridizations were performed for 16 h in a rotating hybridization oven using the Agilent 60-mer oligo microarray-processing protocol. Slides were washed as indicated and scanned with an Agilent scanner. Data was gathered using the Agilent feature extraction software, using defaults for all parameters, save the ratio terms. To account for the use of the direct label protocol, error terms were changed to: Cy5 multiplicative error = 0.15; Cy3 multiplicative error = 0.25; Cy5 additive error = 20; Cy3 additive error = 20.

GEML files and images were exported from the Agilent feature extraction software and deposited into Rosetta Resolver (version 3.2, build 3.2.2.0.33) (Rosetta Biosoftware). Two arrays for each sample pair, including a fluor reversal, were combined into ratio experiments in Rosetta Resolver. Intensity plots were generated for each ratio experiment and genes were considered 'signature genes' if the p-value was less than 0.001. p-values were calculated using the Rosetta Resolver error model with Agilent error terms. The signature genes were analyzed with GeneSpring.

Ontology enrichment

Genes have previously been categorized into various ontologies and pathways. If a particular pathway is enriched for genes that are significantly expressed in response to a process, we conclude that the pathway is likely to be involved in this process. In total, 829 genes out of 6,240 had a significant alteration in expression in at least one experimental condition. Along with the size of each functional category, a statistical measure for the significance of the enrichment was calculated by using a hypergeometric test. The level of significance for this test was determined using the Bonferroni correction, where the α value was set at 0.05 and the number of tests conducted for KEGG pathway and Simplified Gene Ontology (biological process) were 27 and 11, respectively.

Network searches

The ActiveModules algorithm was used to identify neighborhoods in the regulatory network corresponding to significant levels of differential expression. In this search, if a protein has many neighbors, it is likely that at random a few will show significant changes in expression and these could be selected as a significant sub-network. Neighborhood scoring is a method we used to correct for this bias. In this scheme, a significant sub-network must contain either all or none of the neighbors of each protein. The significance then represents an aggregate over all neighbors of a protein. This prevents the biased selection of a few top-scoring proteins out of a large neighborhood in the search for significant sub-networks. For an in-depth description of this algorithm see Ideker *et al.*[103].

In defining the network used in the metabolic analysis, edges corresponding to metabolites linking more than 175 reactions were eliminated. This excludes metabolic cofactors such as ATP, NADH and H2O from the search. Scores for each ORF were generated by mapping the fitness significance value to a Z-score. To assign scores to the individual reactions, Förster's mapping from ORF to reaction was used to generate a list of ORFs for each reaction. The Z-scores of these ORFs were then aggregated into a single score for that reaction using the following equation:

$$Z_{reaction} = \frac{1}{\sqrt{n}} \prod_{i=1}^{n} Z_{ORF_i}$$

We used a dynamic programming algorithm adapted from Kelley *et al.*[175] to identify high-scoring paths in this network. Briefly, the highest-scoring path of length (n) ending at each node is determined by combining the scores of the individual node and the highest-scoring path of length (n - 1) ending at a neighbor node using the following formula:

$$Z_n = \frac{Z_{n-1}\sqrt{n-1} + Z_{reaction}}{\sqrt{n}}$$

Since a node with many neighbors is more likely to belong to a high-scoring path by random chance, the score of the neighboring path is corrected against the extreme-value statistic with the number of observations equal to the number of neighbors.

The significances of the top-scoring networks were determined by comparison to a distribution of the top-scoring networks from random data (reaction scores randomized with respect to the nodes of the network). After running the path finding/scoring algorithm, the score of the single highest-scoring path was added to the null distribution.

This process was repeated for 10,000 interactions. This null distribution was then used to determine an empirical p-value, which represents the null hypothesis that there is no significant correlation between the topology of the metabolic network and the assignment of significance values to nodes in that network.

Specific deletion experiment filter on fold-change comparisons

The intensity plots were generated from each experiment in Rosetta Resolver. A gene was considered a signature gene if the p-value was less than 0.001 and if the fold-change value was greater than or equal to twofold. Signature genes were then broadcasted on the intensity plot and exported as text files. Lists were imported into GeneSpring. The 'Filter on Fold Change' function was used to compare the parent control vs. parent AsIII experiment with each deletion (AsIII) experiment. The gene list selected for each filter on fold change analysis was a combination of the parent signature gene list and the signature gene list of the AsIII-treated deletion being analyzed at the time. For example, if the comparison was being done between parent (AsIII-treated) and Yap1 (AsIII-treated), the list used in the analysis was the combination of the parent signature genes and the Yap1 signature genes. The filter on fold change function reports genes that were selected from the one condition (parent) that had normalized data values that were greater or less than those in the other condition (deletion under investigation) by a factor of twofold. Each resulting gene list was saved. All the resulting gene lists were combined and an annotated gene list was exported for use in Eisen's Cluster/Treeview package (described earlier). The format of the exported data was the natural log. The gene tree generated for the paper was generated in GeneSpring. Each filter on fold change was saved as an annotated gene list.

Generation of specific deletion experiment 'minus' lists

Signature gene lists were generated in Rosetta Resolver from intensity plots as described above. Each signature gene list was saved as a 'Bioset' in Resolver. The parent Bioset was compared to each deletion Bioset using the 'Minus' function. This function finds those members in Bioset group 1 (parent) that do not exist in Bioset group 2 (deletion). Each of the resulting lists was saved as a new Bioset. The new 'minus' Bioset was broadcasted on its corresponding intensity plot and exported as a text file. This was repeated for each experiment with fine-tuning of the data using GeneSpring.

Phenotypic profiling

Homozygous diploid deletion strains and pooling of the strains were done as described[35]. Aliquots were grown until logarithmic phase, diluted to OD600 0.05-0.1, split into tubes and treated with arsenic for 1-2 h at 1 mM, 2 mM and 5 mM concentrations. Similar responses were observed at each concentration, so the results were pooled. These cultures and a mock-treated sample were maintained in logarithmic phase growth by periodic dilution for 16-18 h. UPTAG and DOWNTAG sequences were separately amplified from genomic DNA of the drug and mock-treated samples by PCR using biotin-labeled primers as described previously[35]. The amplification products were combined and hybridized to Tags3 arrays (Affymetrix). Procedures for PCR amplification, hybridization and scanning were done as described[35], and according to the manufacturer's recommendation when applicable. The images were quantified by using the Affymetrix Microarray Suite software. UPTAG and DOWNTAG values were separately normalized, ratioed (treated sample signal/control) and filtered for intensities above background[176].

Tables

Table 5.1. Pathways enriched for genes significantly expressed in response to arsenic.

Transcript profiling reveals that arsenic affects glutathione, methionine, sulfur, selenoamino-acid metabolism, cell communication and heat-shock response. Genes were categorized by KEGG pathway and Simplified Gene Ontology. In total, 829 genes out of 6,240 had a significant alteration in expression in at least one experimental condition. Along with the size of each functional category, a statistical measure for the significance of the enrichment was calculated by using a hypergeometric test. The level of significance for this test (True-shown in bold, False) was determined using the Bonferroni correction, where the α value is set at 0.05 and 27 and 11 tests were done for KEGG pathway and Simplified Gene Ontology, respectively.

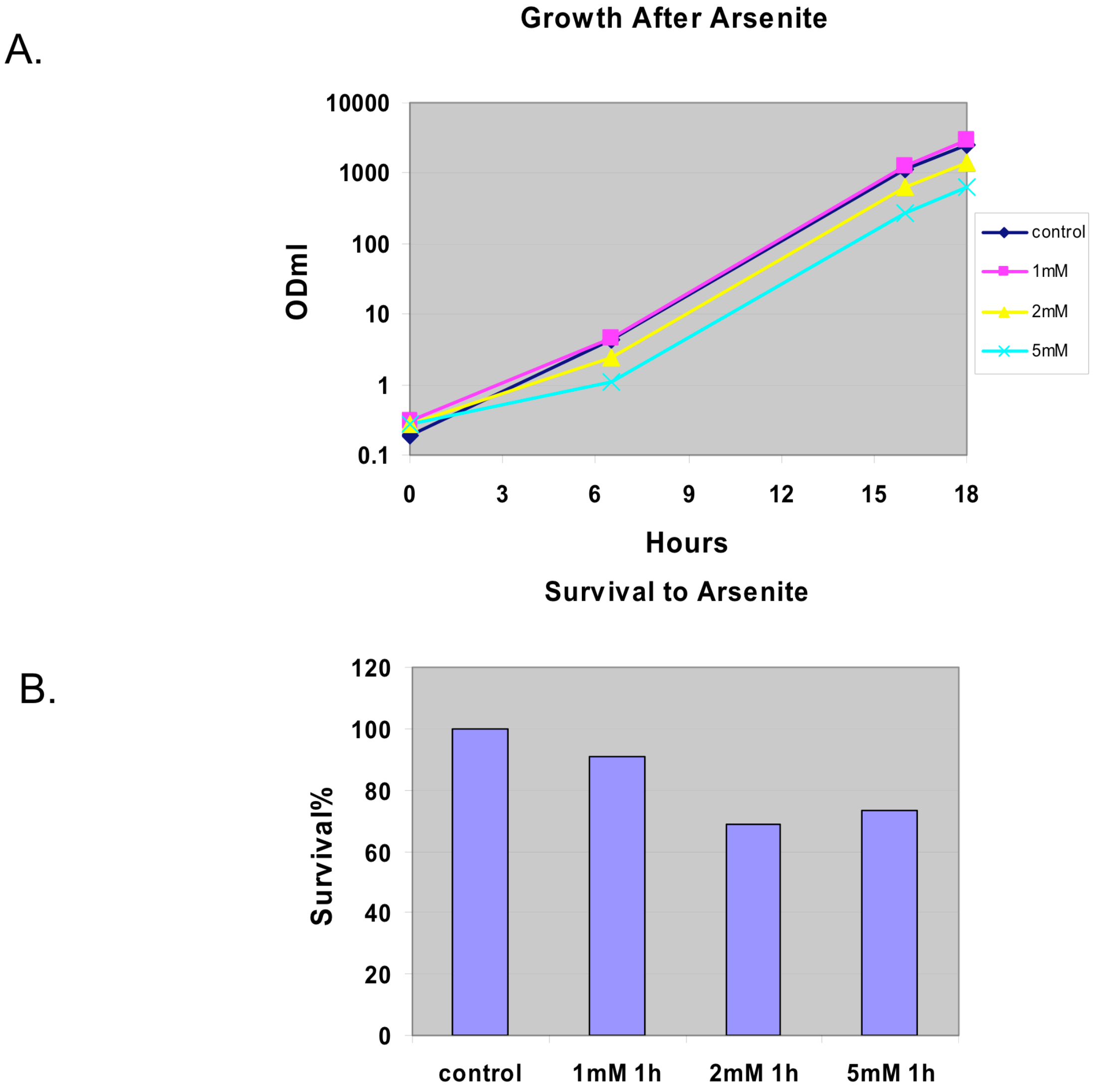
Category	Differentially expressed genes	Pathway size	p-value	Significant
KEGG pathway				
Cell cycle reference pathway	8	87	0.9072	FALSE
Galcatose metabolism	5	15	0.0391	FALSE
Glutathione metabolism	6	11	0.0014	TRUE
MAPK signaling pathway	7	55	0.609	FALSE
Methionine metabolism	8	11	1.07E-05	TRUE
Proteasome	9	30	0.0127	FALSE
Purine metabolism	14	139	0.8991	FALSE
Pyrmidine metabolism	8	80	0.8515	FALSE
Sulfur metabolism	7	7	7.15E-07	TRUE
Serine, threonine and glycine metabolism	8	25	0.0125	FALSE
Citrate cycle	4	22	0.3345	FALSE
Starch and sucrose	9	31	0.0159	FALSE
Pyruvate	4	25	0.4292	FALSE
Reductive carboxylate	5	16	0.0508	FALSE
Second messenger signaling	3	19	0.472	FALSE
Valine, leucine, isoleucine	2	13	0.5313	FALSE
Circadian rhythm	2	19	0.7398	FALSE
Porphyrin and chlorophyll metabolism	7	74	0.8782	FALSE
Selenoamino-acid metabolism	10	12	8.36E-08	TRUE
Ubiquitin-mediated proteolysis	2	29	0.9133	FALSE
Cysteine metabolism	2	4	0.088	FALSE
Fructose and mannose	6	15	0.0093	FALSE
Carbon fixation	3	15	0.3207	FALSE
Alanine and aspartate	2	24	0.8477	FALSE
Glutamate	3	19	0.472	FALSE
Methane	2	4	0.088	FALSE
Gene Ontology (biological process)				
Biological process	72	436	0.0244	FALSE
Cell communication	72	270	<1.00E-008	TRUE
Cell growth and maintenance	47	268	0.0231	FALSE
Cell surface linked signal transduction	14	91	0.3197	FALSE
Developmental processes	5	32	0.4233	FALSE
Heat-shock response	14	22	5.40E-08	TRUE
Intracellular signaling	9	47	0.1635	FALSE
Serine threonine kinase signaling	5	38	0.5815	FALSE
Signal transduction	26	172	0.2656	FALSE
ATPase	3	78	0.9988	FALSE
Cyclin	4	29	0.5499	FALSE

Table 5.2. Genes affiliated to sulfur metabolism.

Arsenic treatment stimulates a sulfur response in yeast. Gene expression data comparisons between arsenic, cadmium, and H2O2-treated Saccharomyces cerevisiae reveal arsenic and cadmium mediated sulfur responses, but none with hydrogen peroxide. AsIII column, 2 h, 100 µM and 0.5 h, 1 mM (combined biological replicates), unless noted; cadmium column, 1 h, 1 mM; H2O2 column, 10, 20, 40, 60, 120 min, 400 µM [62]. Numbers in ordinary typeface denote induction; (-) and italicized numbers denote repression; NC, no change.

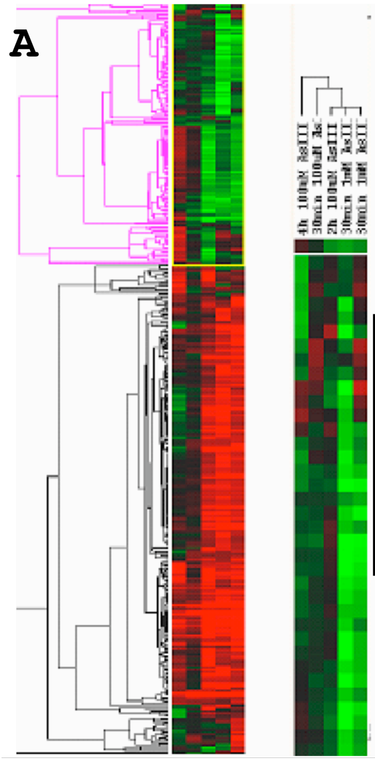
Gene	ORF	AsIII	Cd2+	H2O2	Enzyme
		Fold change			
Sulfate assimilation					
MET3	YJR010W	5.0-20.0	10.8	-4	ATP sulfurylase
MET14	YKL001C	2.0-14.0	9.6	-5	APS kinase
MET16	YPR167C	2.0-12.0	9.1	NC	PAPS reductase
MET22 (0.5 h,	YOL064C	3.3	2.5	NC	Diphosphonucleoside phosphohydrolase
MET10	YFR030W	3.0-5.0	5.2	NC	Sulfite reductase alpha
MET5/ECM17	YJR137C	2.0-5.0	4.5	-6	Sulfite reductase beta
MET1/20	YKR069W	NC	6	NC	Uroporphyrinogen III methylase
MET8	YBR213W	NC	7	-6	Siroheme synthase
Sulfide incorporation and transulfuration pathways					
MET2	YNL277W	4.0-3.0	4.6	NC	Homoserine transacetylase
MET25/17	YLR303W	NC	4.8	NC	O-Acetylhomoserine sulfhydrylase
STR4/CYS4	YGR155W	2.0-2.7	2.5	-5	Cystathionine B-synthase
STR1/CYS3 (4	YAL012W	-3.5	13.4	NC	Cystathionine-lyase
STR3	YGL184C	2.2-3.9	13.5	2.0-4.0	Cystathionine G-synthase
YFR055W	YFR055W	-14	-1.1	NC	Cystathionine-lyase
Methionine and AdoMet biosynthesis					
MET6	YER091C	1.0-3.5	NC	-7	N5-Methyltetrahydrofolate homocysteine transferase
MET7	YOR241W	NC	-1.6	NC	Tetrahydrofolyl polyglutamate synthase
MET13	YGL125W	NC	NC	-4	Methylene tetrahydrofolate reductase
SAM1 (4 h, 10	YLR180W	2.5	NC	-11	AdoMet synthetase
SAM2 (4 h, 10	YDR502C	3.8	NC	-6	AdoMet synthetase
MHT1	YLL062C	5.0-2.8	10.6	NC	S-methylmethionine: homocysteine S-methyltransferase
Sulfur compound uptake					
SUL1	YBR294W	5.4-2.4	20	NC	Sulfate transporter
SUL2	YLR092W	2.5-2.8	3	NC	Sulfate transporter
MUP1 (2 h anc	YGR055W	5.0-14.0	2	-5	Methionine permease, high affinity
MUP3	YHL036W	8.0-7.0	7		Methionine permease, low affinity
Regulatory genes					
MET4 (2 h, 10	YNL103W	2	1.5	NC	bZIP
MET28	YIR017C	NC	5	NC	bZIP
CBF1	YJR060W	NC	-2	NC	bHLH
MET30	YIL046W	5.0-1.6	7	NC	WD40 repeats F box
MET31	YPL039W	NC	1	-8	Zinc finger
MET32	YDR253C	6.0-3.6	14	-5	Zinc finger

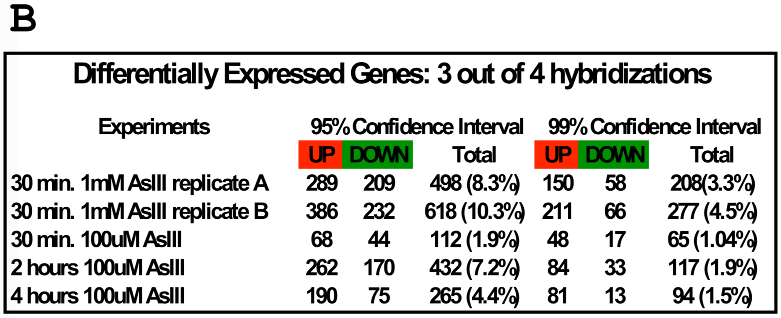
Supplemental Figures



Supplemental Figure 5.1. The dose-response curve of *S. cerevisiae* strain, BY4741.

Treatment with 1 mM, 2 mM and 5 mM AsIII resulted in a negligible effect on growth (after 18 h) and survival (1 h treatment followed by plating and colony formation counting), but still exhibited a pronounced transcriptional response



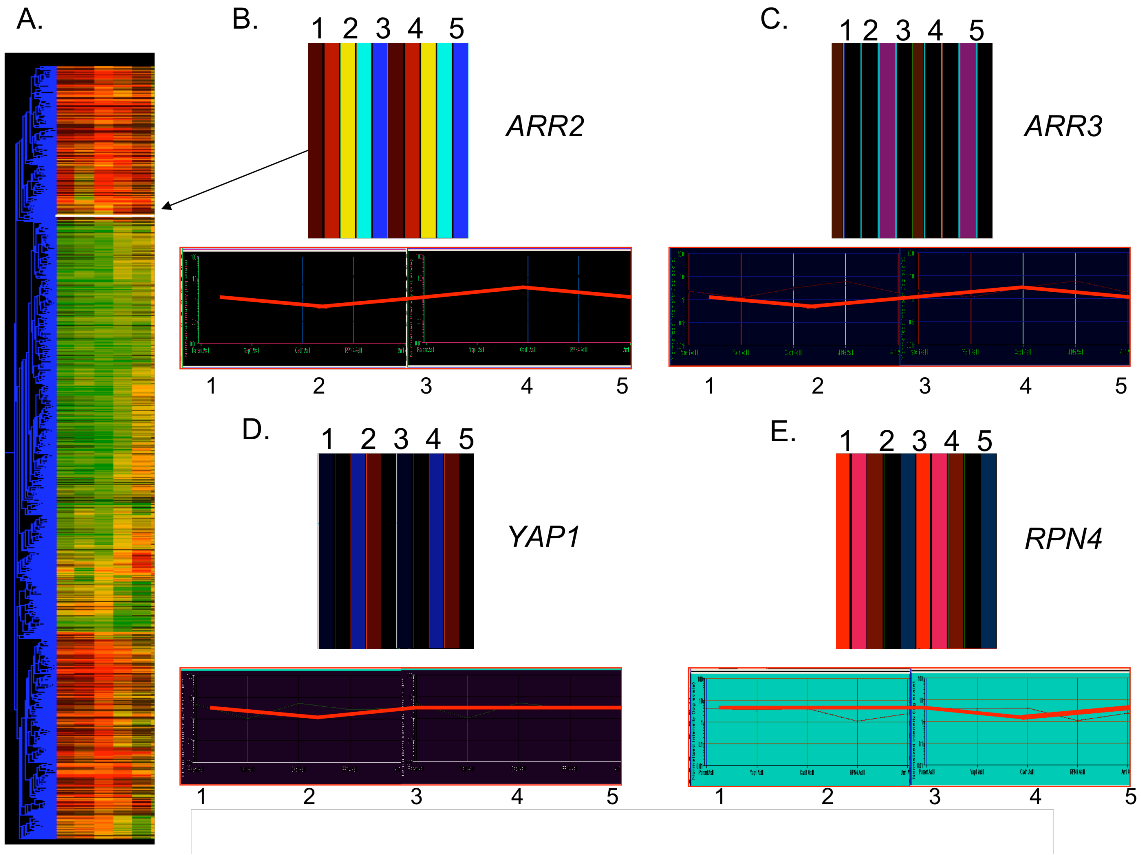


Differentially Expressed Genes: 3 out of 4 hybridizations

Experiments	95% Confidence Interval			99% Confidence Interval		
	UP	DOWN	Total	UP	DOWN	Total
30 min. 1mM AsIII replicate A	289	209	498 (8.3%)	150	58	208(3.3%)
30 min. 1mM AsIII replicate B	386	232	618 (10.3%)	211	66	277 (4.5%)
30 min. 100uM AsIII	68	44	112 (1.9%)	48	17	65 (1.04%)
2 hours 100uM AsIII	262	170	432 (7.2%)	84	33	117 (1.9%)
4 hours 100uM AsIII	190	75	265 (4.4%)	81	13	94 (1.5%)

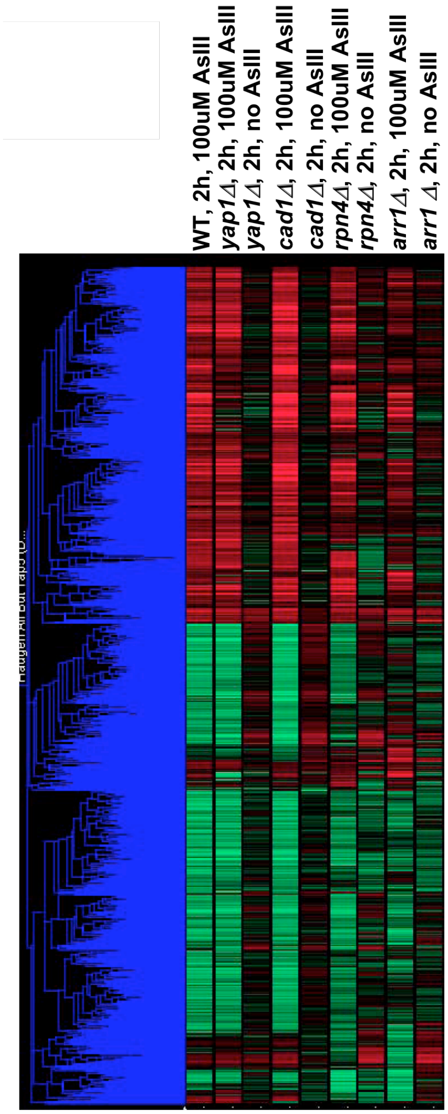
Supplemental Figure 5.2. A self-organized tree of arsenite treated yeast experiments and a table depicting the numbers of significant genes.

All genes found to be significant by MAPS analysis (see Materials and methods) were compiled across the four arrays, averaged and subsequently clustered with Cluster/Treeview software. The dendogram highlighted in pink depicts the zoomed in region shown to the right of the entire tree. Genes in red are induced and genes in green are repressed. A table depicts the numbers of genes changing in each experiment at both the 95% and 99% confidence intervals (see Materials and methods).



Supplemental Figure 5.3. Under arsenite-treated conditions, Yap1 might regulate Arr2 and Arr3, and does not regulate Rpn4.

Yap1 is likely to regulate Arr2 and Arr3 after 2 h 100 μM AsIII but it does not regulate Rpn4 under arsenic-induced stress. The self-organized heat map labeling and conditions in this figure are the same as for Figure 5.2. (a) The Yap1 knockout strain fails completely to induce Arr2 (0.834 average fold-change) whereas the Arr1 knock-out induces Arr2 (2.90 average fold-change). (b) The Arr1 knockout induction is more elevated compared to the Yap1 knock-out (1.8 and 1.1 average fold-change, respectively). (c) Yap1 is induced 2.7 fold in the Rpn4 knock-out. (d) The wild type parent strain shows an averaged induction of 4.7 fold. (e) Rpn4 is induced 3.7 fold in the Yap1 knock-out compared to 4.1 fold induction in the wild type parent strain. In the presence of arsenic, Yap1 does not appear to regulate Rpn4.



Supplemental Figure 5.4. Self-organized clustering of deletion strains with AsIII treatment and parent strain vs. deletion strains without arsenic.

Self-organized clustering of specific deletion and parent strain experiments (yap1Δ vs. yap1Δ 2 h 100 µM AsIII, cad1Δ vs. cad1Δ 2 h 100 µM AsIII, rpn4Δ vs. rpn4Δ 2 h 100 µM AsIII, arr1Δ vs. arr1Δ 2 h 100 µM AsIII, parent vs. parent with 2 h 100 µM AsIII, as well as the parent strain vs. each deletion strain without arsenic).

Acknowledgements

We thank Sherry Grissom, Eric Steele, Dmitry Gordenin and Gopalakrishnan Karthikeyan for technical assistance, James Brown for help with the analysis of the phenotypic profiling, and Rick Paules and Jennifer Fostel for critical review of the manuscript. This work was in part supported by grant CA 67166 (J.M.B.) from the US National Cancer Institute.

Chapter 5, in full, is a reprint of the following work,

Haugen AC, Kelley R, Collins JB, Tucker CJ, Deng C, Afshari CA, Brown JM, Ideker T, Van Houten B. *Integrating phenotypic and expression profiles to map arsenic-response networks*. **Genome Biology** 2004;5(12):R95.

The dissertation author was the second author on this work, responsible for designing and implementing network analysis algorithms.

Chapter 6. Genome-wide fitness and expression profiling implicate Mga2 in adaptation to hydrogen peroxide

Abstract

Caloric restriction extends lifespan, an effect once thought to involve attenuation of reactive oxygen species (ROS) generated by aerobic metabolism. However, recent evidence suggests that caloric restriction may in fact raise ROS levels, which in turn provides protection from acute doses of oxidant through a process called adaptation. To shed light on the molecular mechanisms of adaptation, we designed a series of genome-wide deletion fitness and mRNA expression screens to identify genes involved in adaptation to hydrogen peroxide. These were integrated with databases of known transcriptional interactions to build a genome-scale model of adaptation to oxidative stress. This model supports Yap1 and Skn7 as central transcriptional regulators of both the adaptive and acute oxidative responses. It also underscores the importance of the transcription factors Mga2 and Rox1 exclusively in adaptation, which is striking because these factors have been thought to control the response to hypoxic, not oxidative, conditions. Expression profiling of *mga1Δ* and *rox1Δ* knockouts confirms that these factors most strongly regulate targets in ergosterol, fatty-acid, and zinc metabolic pathways. Direct quantitation of ergosterol shows that its basal concentration indeed depends on Mga2 and Rox1, but that these factors are not required for the decrease in ergosterol observed during adaptation.

Introduction

Oxidative stress is caused by a number of reactive oxygen species (ROS) generated as a result of aerobic metabolism or chemical exposure. These compounds damage a variety of cellular products, including DNA, proteins, and lipid membranes, and are associated with a number of human pathologies. For example, in cardiovascular disease, oxidation of low-density lipoprotein signals an inflammatory response[177]. The sensitivity of neurons to oxidative stress implicates ROS in neurodegenerative diseases, such as Parkinson's and Alzheimer's[12-14].

A continuing source of controversy is the role of oxidative stress in aging. Caloric restriction has been shown to extend lifespan in a number of species[178]. Initially, it was hypothesized that the effect on lifespan occurs primarily because caloric restriction reduces the level of aerobic respiration, a major source of ROS[179]. Newer evidence is challenging this hypothesis, since caloric restriction paradoxically increases respiration[180]. Increased respiration, in turn, can generate mild levels of ROS which protect against high doses of oxidant[181]. This process is known as adaptation or hormesis[11] and is widely conserved among eukaryotes[181-184]. One hypothesis is that adaptation to oxidative stress is the basis for the lifespan-extending effect of caloric restriction[185, 186]. Thus, further efforts to understand the process of adaptation may have broad implications on models of aging and disease.

In one model of adaptation, the cell increases the activity of the enzymes and pathways required to rid the cells of ROS, leaving it better equipped to process acute dosages of oxidant when they arise. Under this model, genes involved in the adaptive response are expected to be a subset of those that become active in the acute response[187].

Many such candidates have been identified, including a variety of biosynthetic enzymes which produce small molecular compounds or proteins with reduction potential, such as glutathione (GSH), thioredoxin, NADPH, and trehalose[188-192]. Different enzymes facilitate this process for different ROS, including catalases and peroxidases (which deal with peroxide radicals)[193, 194] and superoxide dismutases (which deal with superoxide radicals)[195, 196]. Additional proteins serve to repair the damage caused by oxidative stress. Heat shock proteins act as chaperones within the cell, allowing damaged proteins to fold properly or preparing them for disposal[197]. DNA repair genes are also vital, as oxidative stress can damage both nucleotides and the phosphodiester DNA backbone[198]. Several studies have implicated classical oxidative stress proteins and pathways in adaptation, including the transcription factor Yap1[139] and glutathione synthesis[199-201].

In contrast to this model, a second body of evidence suggests that adaptation may be governed by novel pathways not directly involved in the response to acute oxidation. In a study of adaptation to the oxidant linoleic acid, Alic *et al.* found that adaptation can occur without induction of oxidative or general stress response genes following pretreatment[202]. Instead, various metabolic processes were activated and protein synthesis was inhibited. Moreover, machinery with a central role in the acute response, such as the mitochondria[11, 203]or the Msn2/4 environmental stress response factors, are not required for adaptation[139, 204].

Nonetheless, expression studies of acute oxidative damage have helped to identify a set of genes involved in the common environmental stress response (ESR) and implicated the Msn2/4 transcription factors in control of this gene set[134, 162, 205]. In fitness studies of yeast deletion strains, Thorpe *et al.* identified a set of genes required for the response to hydrogen peroxide, mainly dealing with the proper functioning of the mitochondria. However, to-date these genome-scale approaches have focused on the acute, rather than the adaptive, response. The one study to date that has screened for adaptive genes focused on a set of 268 genes selected based on previous literature[206].

Here, we use the rich functional genomics toolbox of yeast to identify pathways involved in adaptation to oxidants. To accomplish this goal, we use barcode arrays to screen the *Saccharomyces cerevisiae* gene deletion collection[207] for genes required in the acute and adaptive responses, and we couple these data with genome-wide mRNA expression profiles to build a system-wide model of adaptation.

Results & Discussion

A genetic screen to identify genes functioning in adaptation

As shown in Figure 6.1A, we elicited adaptation using a protocol consisting of a mild pretreatment of hydrogen peroxide (0.1 mM H_2O_2 for 45 min) followed by a later high dose (0.4 mM H_2O_2 for 1 hr). For purposes of comparison, we also conducted an acute protocol which exposed cells to the high dose only (0.4 mM H_2O_2 for 1 hr). Consistent with previous findings[11], we observed that yeast cells undergoing the adaptation protocol exhibited a smaller reduction in growth rate compared to cells exposed to the acute treatment protocol (Figure 6.1B).

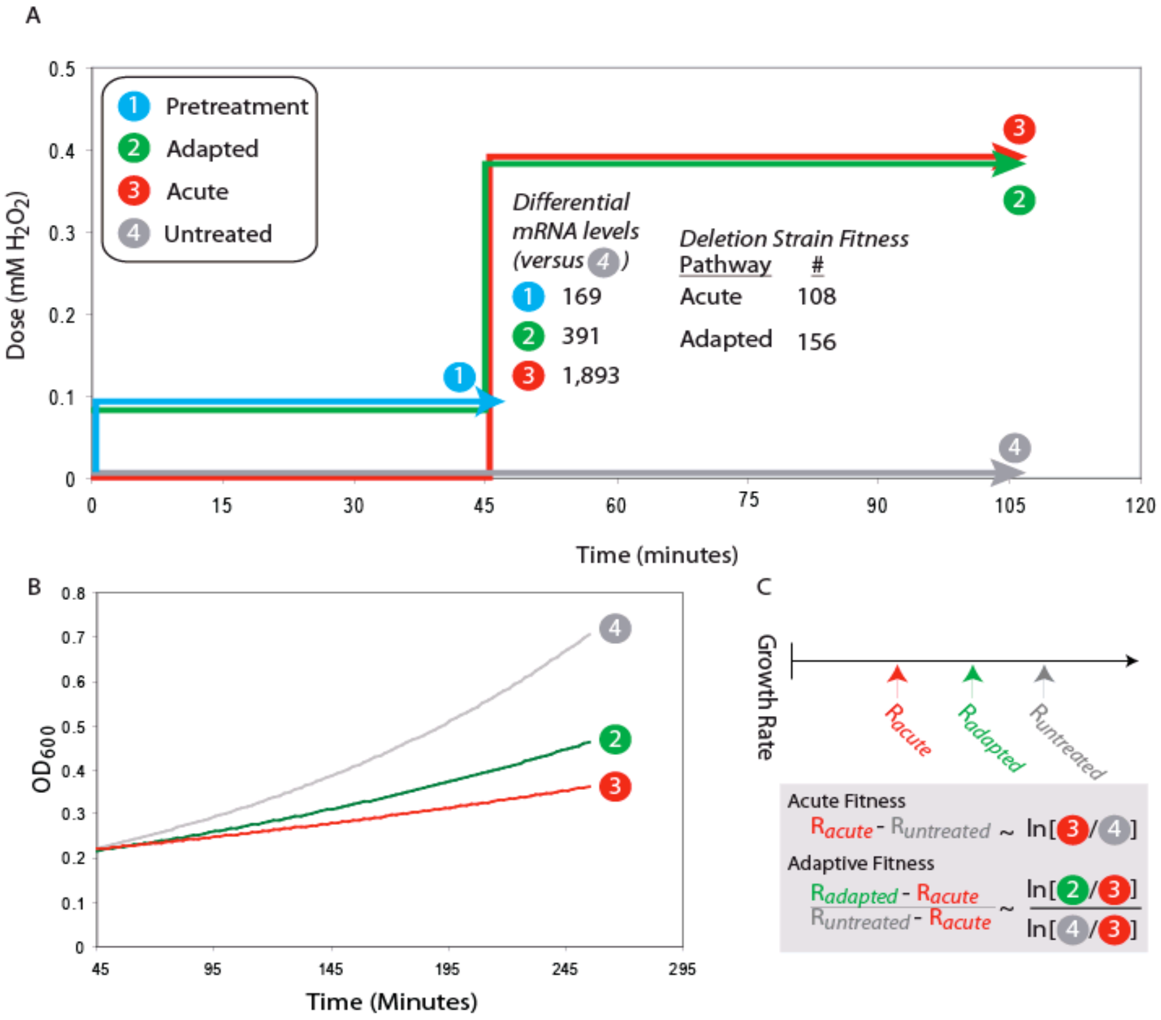


Figure 6.1. Study Design.

A. Yeast cells were collected following each of four hydrogen peroxide treatment conditions (pretreated, adapted, acute, and untreated, labeled 1-4). Competitive growth experiments were performed between gene deletion pools grown in adapted versus acute conditions (to identify genes required specifically for adaptation) and between pools grown in acute versus untreated conditions (to identify genes required for the acute response). Gene expression profiling was performed in either adapted or acute conditions versus untreated cells. **B.** Pretreatment with mild hydrogen peroxide (green) leads to improved growth compared to no pretreatment (red) following a high dose of hydrogen peroxide. **C.** For an individual gene deletion, the acute sensitivity is defined as the difference between the acute and untreated growth rates. The adapted sensitivity is the fraction of that difference that is recovered by mild pretreatment

Given these protocols, we designed a series of yeast genome-wide phenotyping experiments using the publicly available pool of 4,831 viable single-gene deletion strains[35]. Each strain in the pool incorporates a pair of unique oligonucleotide barcode tags, which allow the relative growth rates of all strains to be tracked in competitive growth experiments by hybridization of pooled genomic DNA to a barcode microarray. In a first experiment, two identical pools of deletion mutants were treated with the adapted or acute protocol, respectively, and directly compared on a barcode array (with multiple biological replicates; see Methods). In a second experiment, a pool subjected to the acute treatment was compared against an untreated pool.

These experiments were used to identify genes required for adaptation or for the acute response, as shown in Figure 6.1C. Fitness in the acute response was defined as the difference in growth rate between the acute and untreated conditions (determined from the log ratio of intensities measured in the direct comparison of the acute and untreated pools, see Methods). Adaptive fitness was defined as the difference in growth rate between the acute and adapted conditions, normalized by the magnitude of the acute effect (Figure 6.1C).

Adaptive sensitive deletions are not enriched for the response to oxidative stress

A total of 156 versus 108 genes were found to be required for the adaptive versus the acute responses ($p<0.005$) (Figure 6.2A). These sets overlapped by 88 genes, including *YAP1* and *SKN7*, genes encoding transcription factors with known involvement in the response to oxidative stress. *YAP1* and *SKN7* were previously identified as adaptive-sensitive in the restricted screen conducted by Ng *et al.*[206]. Given the large degree of overlap, it was not surprising that both the adaptive and acute gene sets were also enriched for similar functional categories, such as the mitochondrial ribosome and aerobic respiration (Figure 6.2C). The identification of these functions is puzzling in light of an earlier finding that yeast with defective mitochondria (rho$^-$ mutants) adapt to oxidative stress[11, 203].

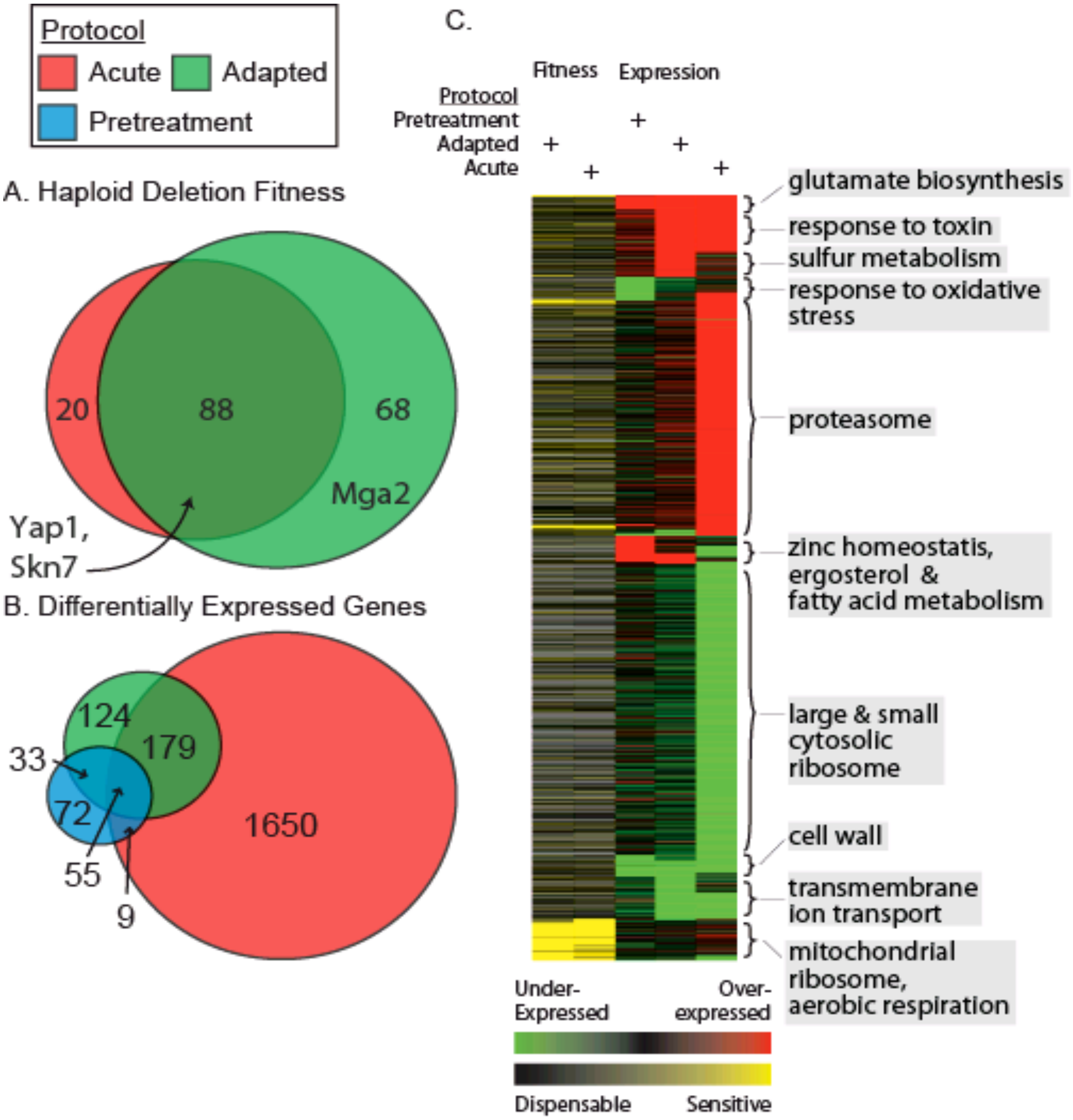


Figure 6.2. Fitness and Expression Profiling Overview.

A. Numbers and overlap of gene deletions that are sensitive in the adaptive (green) and acute (red) treatment protocols. **B.** Numbers and overlap of differentially expressed genes identified in each of the three expression treatment protocols. **C.** Hierarchical clustering of the differentially expressed or sensitive genes from each screen. Clusters are annotated at right with over-represented functional groups.

In these studies, a milder high dose was required to demonstrate adaptation; therefore, the observed deficiency in adaptation of mitochondrial mutants in our screen may be due to increased sensitivity to the high dose. Surprisingly, neither set was enriched for genes involved in the response to oxidative stress (GO Biological Process 0006979) which may be due to the ability of this response to compensate for the loss of single gene activities, confirming earlier observations regarding the acute response by Thorpe *et al.* (Supplemental Table 6.1)[208]. Despite the strong overlap of these two sets of sensitive genes, a number of gene deletions were sensitive only in the adaptive screen, including the transcription factor *mga2Δ*.

Unique sets of genes are expressed during the adaptive response

Next, we performed mRNA expression profiling on each of the three treatment protocols (pretreated, adapted, acute, see Figure 6.1A) in comparison to untreated conditions. These profiles were analyzed to identify two types of adaptive response genes: early versus late. Early adaptive genes were defined as those that were differentially expressed after the 45 min. pretreatment relative to untreated conditions (169 genes at $p<10^{-5}$, see Methods). Late adaptive genes were defined as those that were differentially expressed after the 1 hr. high dose following pretreatment (391 genes). In comparison, a much larger set of 1,893 genes was differentially expressed in response to the high dose in the absence of pretreatment. Thus, the numbers of differentially expressed genes increases with the severity of treatment (pretreatment, adapted, acute). The sizes of these gene sets roughly correlated with the reduction in growth rate associated with each treatment.

The overlap of the acute expression response with either the early or late adapted responses was significant (p=0.02 versus p=7×10^{-36} by hypergeometric test, respectively); nonetheless the overlap with the early response was much less than with the late adapted response (38% versus 60%, see Figure 6.2B). In addition, 26 genes that would be expected to be increasing in expression based on the acute expression data were decreasing in expression during adaptation, such as genes involved in the response to oxidative stress (GO Biological Process 0006979) (Figure 6.2C). Other sets of genes were expressed uniquely during early and late adaptation, including ergosterol metabolism, fatty acid synthesis, and zinc homeostasis (GO Biological Processes 0008204, 0006631, 0055069, respectively) (Figure 6.2C). Unlike the fitness profiling, oxidative stress genes were strongly implicated in the acute expression response (as also found by others; Supplemental Table 6.2 and Supplemental Table 6.3**).**

Centrality of transcription factors Mga2, Rox1, and Yap1 during pretreatment

To map the transcriptional program underlying adaptation, we computed the activity of each yeast transcription factor based on the significance of differential expression among its set of known targets (Figure 6.3). Lists of targets for each factor were drawn from YeastRACT, a database of literature-curated regulatory interactions[209] (Methods). Application of this method to the acute treatment protocol identified Msn2/4, Yap1, and Skn7 as key factors, all of which had been previously associated with the acute response to oxidative stress. All of these factors were also moderately active during pretreatment and become more so after transitioning to the high dose (Figure 6.3). Other factors exhibiting this behavior include Adr1, Hsf1, and Pdr1/3.

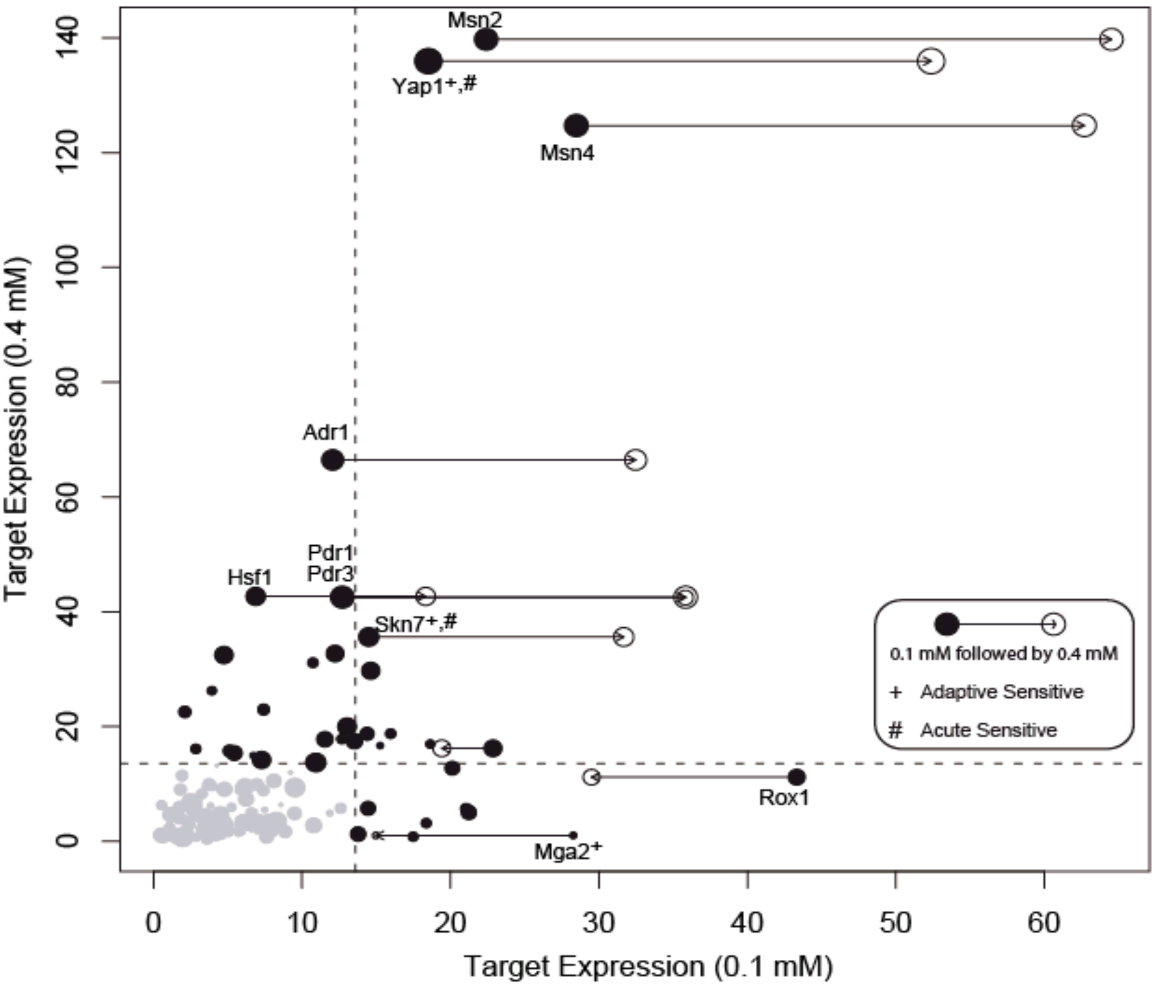


Figure 6.3. Dynamics of transcription factor target expression in mild and acute conditions.

For each transcription factor, we compute a score based on a hypergeometric test representing the significance of increased expression (relative to untreated) of known targets (see Methods) following either pretreatment (0.1 mM H_2O_2, x-axis) or acute treatment (0.4 mM H_2O_2, y-axis). For a limited set of transcription factors with the most significant activity following acute treatment or pretreatment, the activity following adaptive treatment (0.1 mM followed by 0.4 mM H_2O_2) is also displayed on the x-axis with an open circle. The size of each point corresponds to the number of known targets of that transcription factor. The dotted lines indicate a threshold for significance determined by a randomization procedure (see Methods). Although there is significant overlap in the set of expressed genes following mild and acute treatment, examination of specific transcription factors reveals those with unique behavior in each condition. Transcription factors identified in the deletion fitness analysis of the acute and adaptive treatments are indicated with "#" and "+" symbols, respectively.

On the other hand, Mga2 and Rox1 exhibited highly significant activity during pretreatment, but not during the acute response (Figure 6.3). These factors had previously been associated with the hypoxic, not oxidative, stress response[210, 211]. Thus, our analysis appears to classify transcription factors into two categories: early response factors activated by mild doses of oxidant during pretreatment only (Rox1, Mga2), and late damage response factors whose level of activation responds in proportion to treatment dose (Msn2/4, Yap1, Skn7).

A model of adaptation to oxidative stress

When considered together and in light of the previous literature, our results suggest the model of adaptation shown in Figure 6.4. When the cell is exposed to a high dose of hydrogen peroxide, Yap1 and Skn7 up-regulate the expression of genes involved in redox homeostasis, preventing cellular damage by increasing the degradation of ROS. This is the likely mechanism of the observed requirement for both Yap1 and Skn7 in the acute response. In addition, Yap1 and Skn7 are also activated during the pretreatment and are required for adaptation (Figure 6.3). In response to mild pretreatment with hydrogen peroxide, Mga2 and Rox1 activate targets involved in ergosterol metabolism, fatty acid biosynthesis, and zinc homeostasis. Previous literature suggests possible roles for each of these processes in oxidative adaptation. Ergosterol is a cholesterol-like component of the plasma membrane with diverse effects on its function[212]. Branco *et al.* observed that adaptation is associated with an increase in membrane rigidity, an effect which is abrogated in the ergosterol-deficient *erg3*Δ and *erg6*Δ strains[213]. We hypothesize that an increase in ergosterol biosynthesis may inhibit diffusion of H_2O_2 across the plasma membrane by reducing membrane permeability. Zinc homeostasis genes may play a similar role, as these genes also influence ergosterol metabolism[214]. Conversely, Tafforeau *et al.* observed a decrease of both squalene synthase (Erg9) activity and

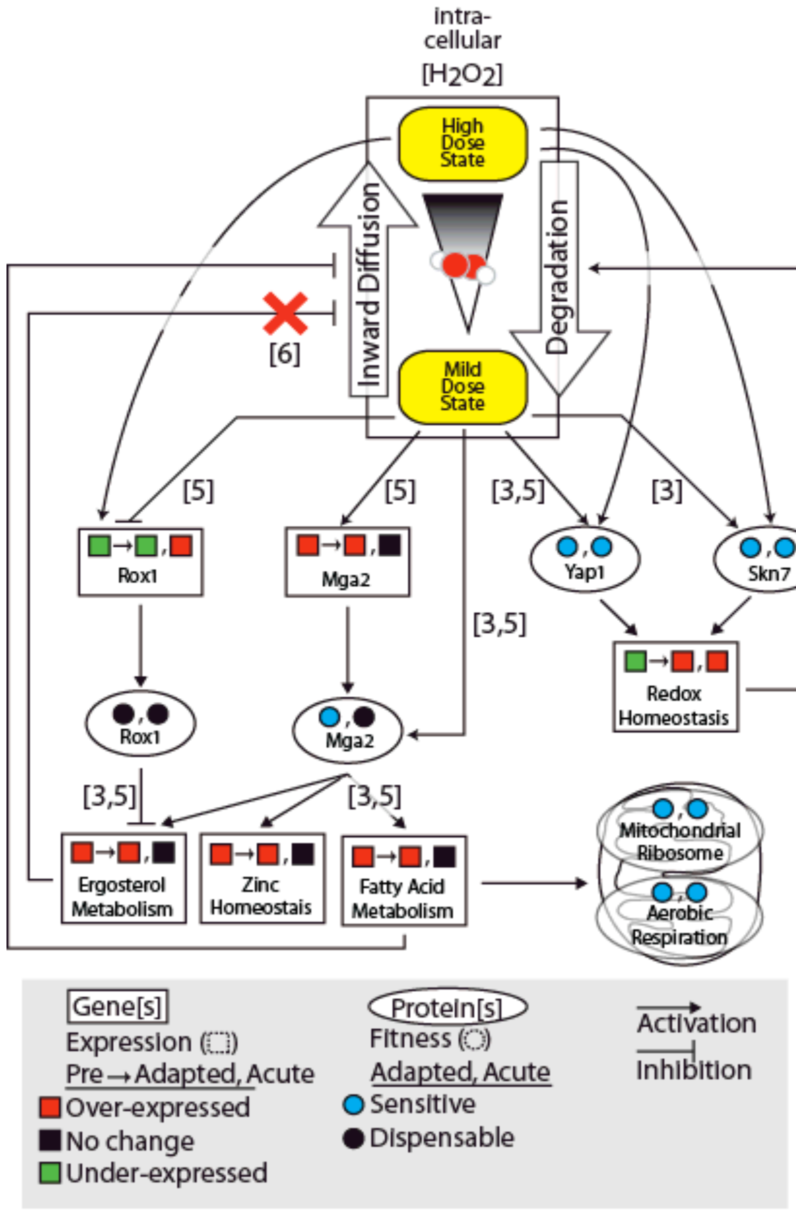


Figure 6.4. Model of the adaptive response.

Results and hypotheses regarding transcriptional regulators and functional categories identified in this study are summarized. The influence of hydrogen peroxide is determined by its concentration within the cell. In addition to treatment dose, several cellular processes affect the level of H_2O_2. In order to enter the cell, hydrogen peroxide must first diffuse across the plasma membrane. Inside the cell, peroxide levels are reduced by degradation into oxygen and water. Squares denote the expression of genes or gene sets (rectangles) following each of the three treatment protocols (pretreatment, adapted, and acute). Conversely, circles denote the sensitivity of the corresponding gene deletion for a particular protein or protein set (oval) in the adapted and acute treatment protocol. Arrows between different objects indicate either an activating (triangular arrowhead) or inhibitory (flat arrowhead) influence. The figure number(s) which provides support for each link are shown in brackets. A red “X” denotes a hypothesis which is later refuted by experimental observation.

ergosterol content during adaptation in *S. pombe*, highlighting the complex relationship between ergosterol and membrane permeability[215]. Although the activity of fatty acid synthetic enzymes could influence the stability and permeability of the plasma membrane[216], these enzymes also influence the composition of membranes throughout the cell. Thus, an additional possibility is that these enzymes influence the activity of enzymes in the mitochondrial membrane[217]. Mutations in *OLE1* are known to influence mitochondrial morphology and inheritance, ostensibly through altering the properties of the mitochondrial membrane[218].

Deletion studies confirm the activation of genes by Mga2 and Rox1

The involvement of Mga2 in early adaptation is supported by its requirement for adaptive growth in the deletion profiling experiments (Figure 6.2) and the striking behavior of its targets in the expression profiling experiments (Figure 6.3). To further confirm the activity of Mga2, pretreatment with hydrogen peroxide was repeated in an *mga2*Δ background and gene expression was profiled versus wildtype cells using quadruplicate whole-genome microarrays. In this experiment, the number of up-regulated Mga2 targets was significantly decreased (Figure 6.5A, p=0.012 by Fisher's Exact Test), supporting its role in the transcriptional program leading to adaptation. Moreover, the *MGA2* gene is itself up-regulated following pretreatment and the transition to the high dose ($p=1.4\times10^{-3}$ and 5.3×10^{-5}, respectively).

Rox1 (<u>R</u>epressor of Hyp<u>ox</u>ic Genes) is a repressor under transcriptional control of Hap1[219]. The decrease in expression of the *ROX1* gene following both the pretreatment and adapted treatment protocols ($p= 3.6\times10^{-11}$ and 1.4×10^{-7}, respectively) suggests that this repressor is deactivated in the process of adaptation. To confirm this observation we profiled a *rox1*Δ strain and found that the number of Rox1 targets with increased expression following pretreatment falls significantly (p=0.046 by Fisher's Exact Test) indicating reduced activation of the genes that it is known to repress (Figure 6.5B).

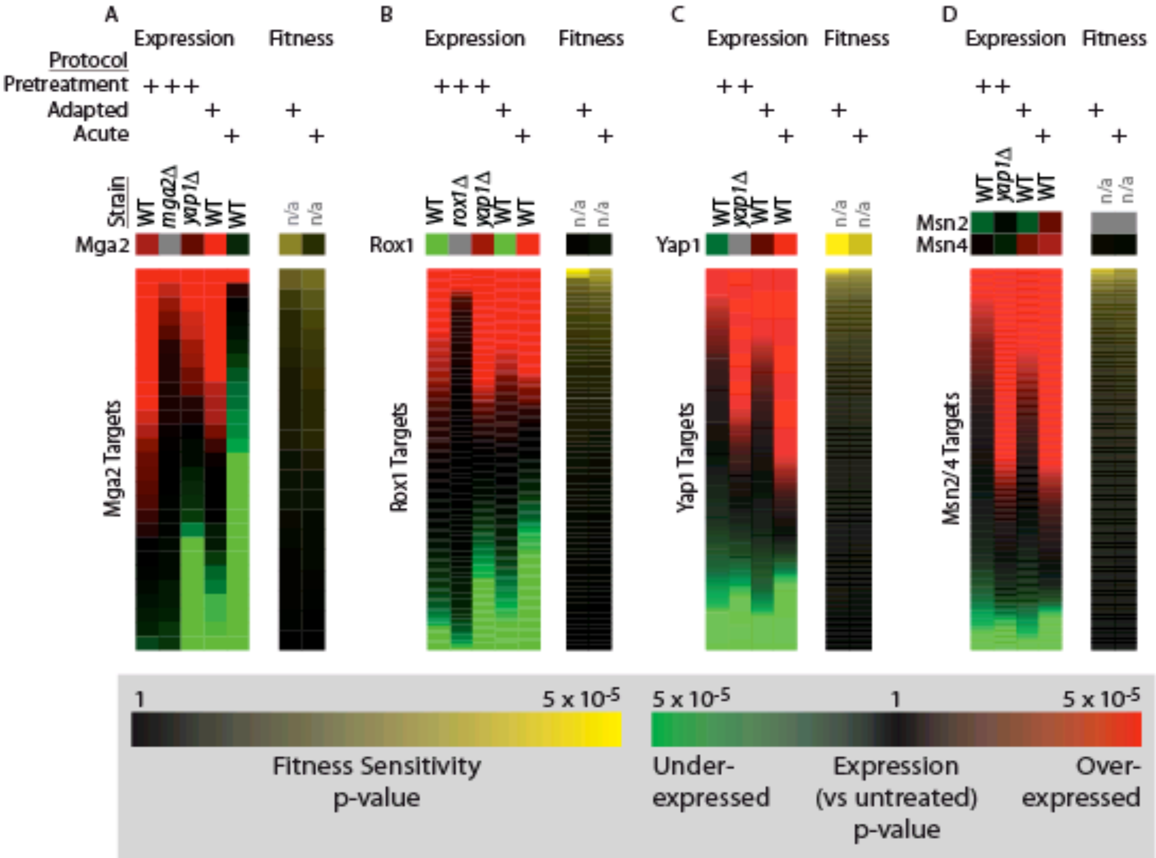


Figure 6.5. Expression analysis of deletion mutants validates the activation of key transcription factors in response to H_2O_2 pretreatment.

Panels A-D detail the behavior of the transcription factors Mga2, Rox1, Yap1, and Msn2/4 and their target sets, respectively. Each column represents the expression or fitness values in sorted order for a specific set of genes.

The mechanism by which Mga2 and Rox1 can be activated by mild pretreatment with oxidants is unknown, but several lines of evidence suggest that the mechanism is shared with the hypoxic response. Rox1 is expressed in a heme-dependent manner[220]. While falling heme levels typically signal hypoxic conditions[221], hydrogen peroxide may also reduce heme levels via degradation[222]. Dirmeier *et al.* found that ROS levels transiently increase following exposure to anoxic conditions, suggesting that this could signal the expression of hypoxic genes[223]. They did not believe the activation of hypoxic genes could be replicated with exogenously supplied ROS, based on the H_2O_2 expression profiling data of Causton *et al*.[162]. We contradict this earlier hypothesis with the observation of increased expression of hypoxic genes as a result of treatment with H_2O_2. The apparent discrepancy may be a result of the higher dose of H_2O_2 used by Causton *et al*[162].

Yap1 is required for expression changes in response to mild pretreatment

To validate the observed requirement of Yap1 during adaptation, we profiled the expression response of a *yap1*Δ strain versus wildtype cells under the pretreatment protocol. This experiment revealed widespread changes in patterns of expression (Figure 6.5). For all sets of transcription factor targets examined in Figure 6.5 (including not only the direct targets of Yap1 but also the targets of Mga2, Rox1, and Msn2/4), their expression responses in the *yap1*Δ strain most closely resembled their expression responses in the wild type following acute treatment. One explanation for this result is that Yap1 acts during pretreatment to promote H_2O_2 degradation and prevent oxidative damage, which otherwise interferes with the adaptive response of many downstream factors (Figure 6.4).

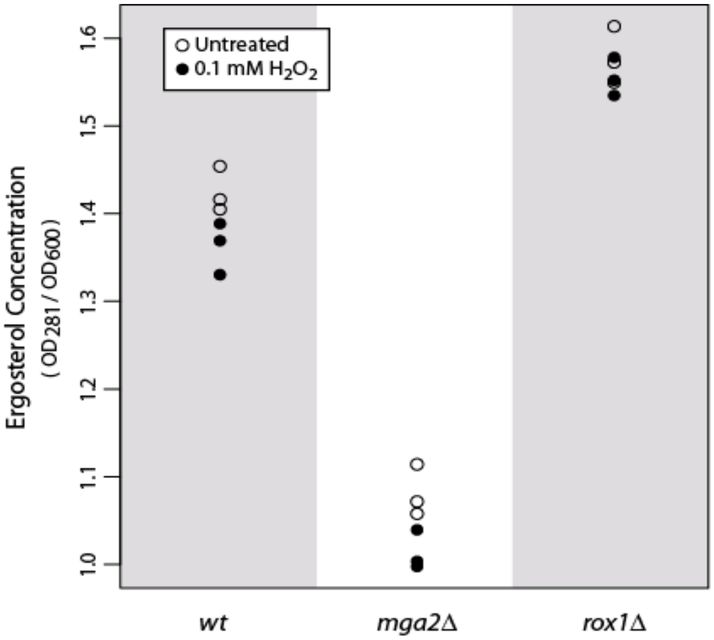


Figure 6.6. Dynamics of ergosterol following mild treatment with hydrogen peroxide.

Following an n-heptane extraction (see Methods), the presence of ergosterol is detected at 281 nm. The ergosterol concentration (relative to the number of cells [OD_{600} value] in the original culture) is reported for wild type, *mga2Δ*, and *rox1Δ* strains with and without mild hydrogen peroxide pretreatment.

Mga2 and Rox1 do not mediate the role of ergosterol in adaptation

To elucidate the role of ergosterol biosynthesis in adaptation, we profiled ergosterol concentration in both untreated and adaptive conditions in *wt*, *mga2*Δ, and *rox1*Δ strains (see Methods). Relative to wild type, the basal concentration of ergosterol was lower in the *mga2*Δ strain and higher in the *rox1*Δ strain (Figure 6.6). This finding agrees with the regulatory roles of Mga2 and Rox1 as an activator and repressor of ergosterol biosynthesis genes, respectively. In each strain, ergosterol content decreased significantly following mild pretreatment with hydrogen peroxide (p=0.014, 0.005, and 0.031 for wild type, *mga2Δ*, *rox1Δ* strains, respectively using a paired t-test). This supports the earlier work of Tafforeau *et al.*[215] but is surprising given the increased expression of ergosterol biosynthetic genes relative to untreated conditions. Nonetheless, the change in ergosterol content between pretreated and untreated conditions was persistent across all strains, suggesting that this change is not due to the regulatory activity of either Mga2 or Rox1. Therefore, we conclude that transcriptional regulation of ergosterol biosynthesis by Mga2 and Rox1 is not a primary mechanism of adaptation. This conclusion is further supported by the observation that deletion of either *MGA2* or *ROX1* attenuates the expression response of ergosterol synthesis genes to mild pretreatment; however, only the *mga2*Δ strain exhibits defects in adaptation (as confirmed in Figure 6.7).

Mga2 may influence adaptation via regulation of fatty acid synthesis

One of the most highly expressed genes following mild pretreatment with hydrogen peroxide was *OLE1,* a gene involved in fatty acid biosynthesis. We found that the high expression of *OLE1* was maintained in a *rox1*Δ background but was greatly reduced in a *mga2*Δ strain ($p=8.3*10^{-3}$), suggesting that the key role of Mga2 in oxidative adaptation might be its regulation of fatty acid synthesis.

Previous work by Matias *et al.* noted repression of Fatty Acid Synthetase (*FAS1*) during adaptation and an inverse correlation between Fas activity and resistance to H_2O_2[216]. In comparison, we observed increased expression of *OLE1*, *ELO1*, *FAS1*, and *FAS2* during mild pretreatment, demonstrating that adaptation occurs in the presence of increased *FAS1* expression. In the study of Matias *et al.*, 0.15 mM H_2O_2 was used to stimulate adaptation compared to 0.10 mM for this study. At higher concentrations (0.4 mM), we also observed a decrease in *FAS1* expression, suggesting that a difference in dosage can explain the discrepancy.

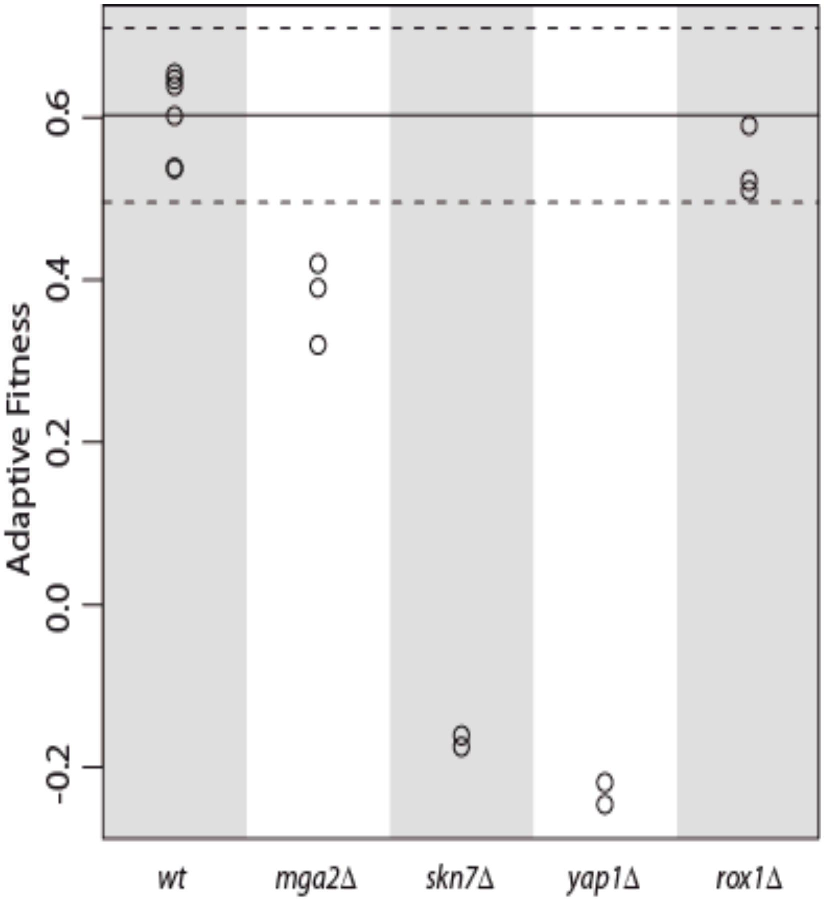


Figure 6.7. Sensitivity of mutant strains in adaptation to hydrogen peroxide.

Adaptive fitness was measured for each of seven deletion or *DAmP* strains over replicate cultures starting from single-cell colonies (Methods). Smaller values indicate a strain defective in adaptation. The adaptation defects measured with the barcode array (Fig. 2A,C) are confirmed for *yap1*, *skn7*, *mga2*, and *rox1* deletion strains. Wild type (*wt*) adaptive fitness is provided as a control, with horizontal lines indicating the *wt* mean (solid line) ± 2 standard deviations (dotted).

In conclusion, we have completed the first genome-wide scan for genes required for the adaptive response to oxidative stress. By integrating these data with results from expression profiling, we have identified pathways with novel involvement in the response to oxidative stress, including the hypoxic response factors Mga2 and Rox1. The activation of Rox1 and Mga2 under adaptive conditions provides additional information about the sensing mechanism of the hypoxic response, given that we have demonstrated this response can be initiated by exogenous oxidative stress. Future studies can interrogate the manner in which the homologs of these genes are necessary for adaptation in higher organisms and explore their role in disease-related oxidative stress.

Methods

Determination of treatment protocols

The high dose of 0.4 mM H_2O_2 was selected to be comparable to other previous expression studies of acute hydrogen peroxide exposure (0.4 mM, 0.24 mM, 0.32 mM, for Causton, Shapira, Gasch, respectively)[134, 162, 205]. This dose resulted in a reduction of growth rate by approximately two thirds as measured by OD_{600}. The pretreatment dose was selected as the largest dose that did not result in impaired growth or viability. This criteria and the length of pretreatment (45 minutes) were selected in accordance with previous studies of adaptation to oxidative stress[11, 204, 224].

Sample growth and treatment for mRNA profiling

We profiled the response to three hydrogen peroxide treatment protocols (pretreatment, adapted, and acute) over a series of microarray experiments. Each series consisted of four biological replicates. For each replicate in the acute treatment protocol, a single colony of BY4741 (ATCC, Manassas, Virginia, USA) was used to inoculate 10 ml of YPD media. Following overnight growth at 30˚ C, this culture was resuspended in 100 ml of YPD media at an OD_{600} of 0.1 and placed in an orbital shaker at 30˚ C. At $OD_{600} = 0.6$ cells were split into two 50 mL portions. In the acute treatment protocol growth continued for 45 minutes, at which point a high dose of hydrogen peroxide (final concentration in media: 0.4 mM H_2O_2) was administered to one member of the pair (with the other receiving a sham treatment of 100 mM phosphate buffer). Treatment continued for 1 hour at which point cells were harvested by centrifugation at 3000 rpm for 5 min. Pellets were immediately frozen in liquid nitrogen and stored at -80˚ C. The pretreatment protocol was identical except for the final concentration of hydrogen peroxide (0.1 mM). For the adapted treatment, a pretreatment dose of hydrogen peroxide (0.1 mM) and corresponding sham treatment were administered directly after splitting the culture, but otherwise the treatment was identical to the acute protocol.

Strain construction

All single deletions were obtained from the complete yeast deletion collection in the BY4741 background (ATCC, #2013888) and verified by PCR (http://www-sequence.stanford.edu/group/yeast_deletion_project/single_tube_protocol.html).

mRNA expression analysis

RNA from each sample was isolated via phenol extraction followed by mRNA purification [Poly(A)Purist, Ambion, Catalog # 1916]. Purified mRNA from the control experiments was labeled with dUTP incorporating either Cy3 or Cy5 dye (CyScribe First-Strand cDNA labeling kit, Amersham Biosciences). Cy3 and Cy5 labelings were alternated between replicates to create a balanced design. Complementary labelings (Cy3 versus Cy5) were hybridized to Agilent expression arrays (Catalog # G4140B).

Arrays were scanned using a GenePix 4000A or PerkinElmer Scanarray Lite microarray scanner and quantified with the GenePix 6.0 software package. Data from each array were subjected to background and quantile normalization[100]. The VERA software package was used with dye bias correction[225] to assign a significance value λ of differential expression to each gene. In a negative control experiment (quadruplicate untreated vs. untreated arrays), the distribution of significance values λ over all genes was fit parametrically as 1.7 * $\chi^2_{1,}$ where χ^2_1 is the chi square distribution with one degree of freedom. This null distribution was used for assignment of p-values.

Sample growth and treatment for haploid deletion fitness profiling experiments

A pool of the 4,831 viable haploid deletion strains was created from individual collections kept in glycerol stock and divided into 1 mL aliquots stored at -80˚ C. Two separate types of treatment protocols (acute and adapted) were studied consisting of four and six replicate arrays, respectively. For each replicate, a single aliquot of pooled deletion strains was diluted in 15 mL YPD media and grown in a rotating wheel at 30˚ C to OD_{600} = 0.6. The sample was then split into two 6.5 mL portions. In the adapted treatment protocol, one member of the paired samples was immediately treated with a mild dose of oxidant (final concentration in media: 0.1 mM) and the other received a sham treatment. After 45 minutes of continued growth at 30˚ C, a high dose was administered (final concentration in media: 0.4 mM) to both samples. After 1 hour of treatment, the cells were harvested by centrifugation at 3000 rpm for 5 min and resuspended in 50 mL of YPD media. After 5 hours of growth, the cells were once again harvested by centrifugation and the pellets were immediately frozen in liquid nitrogen and stored at 80˚ C. The acute treatment protocol was identical, except that no sample was treated with a mild pretreatment dose and only one member of the sample pair was treated with the high dose.

Deletion Fitness Analysis

Genomic DNA was extracted from cell pellets using a glass bead preparation[226]. Subsequent DNA labeling, hybridization, and microarray design followed the protocol of Yuan *et al.*[227]. Briefly, asymmetric PCR was used to amplify unique tag sequences in the genomic DNA of the deletion strains. In each PCR reaction, 1 μg of gDNA was used for labeling. Arrays were scanned and quantified in the same manner as the arrays prepared for the expression profiling experiments.

The *hoptag* package (implemented in R) was used to analyze the intensity data from the scanned arrays. Briefly, median and loess correction were performed on the intensity distributions[227], after which each deletion strain was assigned an UPTAG ratio and a DNTAG ratio for each array. The logs of these ratios were averaged to derive one measurement per gene per array. Across multiple arrays measuring the same treatment protocol comparison (acute vs. untreated or acute vs. adapted), the distribution of log ratio values was quantile normalized[100].To determine acute and adaptive fitness values, we assumed that the signal intensity for a given gene deletion strain is:

$$I_{i,treatment} = N_{treatment}\left[C_i\right]e^{tR_{i,treatment}}$$

where $I_{i,treatment}$ is the observed signal intensity for gene deletion strain *i* subject to the designated *treatment* protocol, [C$_i$] is the initial concentration of deletion strain *i,* $R_{i,treatment}$ is the specified growth rate, and *t* is time. $N_{treatment}$ is a constant factor applied to all intensities from the same treatment representing the shared effect of normalization procedures. For each gene deletion strain *i*, the log ratio of the acute and untreated signal intensities is therefore:

$$\begin{aligned}\ln\left(\frac{I_{i,acute}}{I_{i,untreated}}\right) &= \ln\left(\frac{N_{acute}[C_i]e^{tR_{i,acute}}}{N_{untreated}\left[C_i\right]e^{tR_{i,untreated}}}\right)\\ &= t\left(R_{i,acute}-R_{i,untreated}\right)+\ln\left(\frac{N_{acute}}{N_{untreated}}\right)\end{aligned}$$

Note that the acute fitness measure R_{acute}-$R_{untreated}$ is not directly equivalent to this log ratio. However, since the parameters *t, N_{acute}, and $N_{untreated}$* are shared over all gene deletions *i*, their ordering is the same. Since each intensity distribution was normalized to share the same median, the distribution of log ratios was centered on zero. In order to indentify genes which deviate significantly from this expected value, we performed a one sample *t*-test testing the difference of the mean against zero. This test was regularized to share the estimate of variance among all genes.

Similarly, the log ratio obtained from the direct comparison of the acute and adapted samples was centered on zero and related to the magnitude of the difference, $R_{adapted}$-R_{acute}. Furthermore, due to median normalization of the intensity distributions, the scales of both log ratio distributions were approximately equal. Thus, for most genes without a defect in adaptive fitness, the difference $R_{adapted}$-R_{acute} was strongly correlated to the acute fitness measure, R_{acute}-$R_{untreated}$. A gene with a large difference between the values R_{acute}-$R_{untreated}$ and $R_{adapted}$-R_{acute} indicates a deviation from the average adaptive fitness measure. A two-sample regularized t-test comparing the log ratios determined from each direct comparison was used to identify such cases.

Validation of Sensitive Targets

To verify that the identified sensitive genes are meaningful, the sensitivity of specific gene deletions was verified in small-scale experiments. In these, a colony of a specific deletion strain of interest was incubated in YPD overnight. Following dilution to OD600 0.1 in 30 mL YPD media, the culture was grown to OD_{600} 0.6 and split into three aliquots. Each aliquot was treated according to one treatment protocol (untreated, adapted, or acute). Following one hour of recovery, the optical density of each culture was measured. Optical density values were used to calculate the adaptive fitness measure in the following manner:

$$\frac{\ln\left(OD_{adapted} / OD_{acute} \right)}{\ln\left(OD_{untreated} / OD_{acute} \right)} = \frac{\ln\left(OD_{initial} e^{r_{adapted} t} / OD_{initial} e^{r_{acute} t} \right)}{\ln\left(OD_{initial} e^{r_{untreated} t} / OD_{initial} e^{r_{acute} t} \right)}$$

$$= \frac{\ln\left(e^{r_{adapted} t}\right) - \ln\left(e^{r_{acute} t}\right)}{\ln\left(e^{r_{untreated} t}\right) - \ln\left(e^{r_{acute} t}\right)}$$

$$= \frac{r_{adapted} - r_{acute}}{r_{untreated} - r_{acute}}$$

An unpaired t-test was used to determine the significance of the difference from results obtained when applying the same procedure to wild type (BY4741) colonies.

Determination of Ergosterol Concentration

The determination of ergosterol was adapted from Arthington-Skaggs *et al.*[228]. Following overnight incubation, a culture was grown in YPD to OD_{600} 0.6 and split into two aliquots of 50 mL. One of the aliquots was treated with 0.1 mM H2O2 for 1 hour, after which the OD_{600} of each aliquot was measured. Each aliquot was pelleted and washed once with water. The cleaned pellet was incubated for 1 hour at 85 C with 3 mL 25% alcoholic KOH. After cooling for 15 minutes, 1 mL water and 3 mL n-heptane were added and the mixture was vortexed for 3 minutes. The n-heptane layer was extracted and the presence of ergosterol was detected via absorbance at OD_{281}. Ergosterol concentration was reported as the ratio of OD_{600} / OD_{281}.

Enrichment Analysis of Gene Sets

We investigated the significance of enrichment for functional classes among both differentially expressed and sensitive genes. Functional classes were defined in one of two ways: (1) classes of genes with common annotation in the Gene Ontology (GO) hierarchy[229]) or (2) classes of genes targeted by the same transcription factor as recorded in the YEASTRACT online database[209]. To prevent the identification of redundant or overly general gene ontology categories, we limited the GO analysis to those categories that contained between 5 and 100 genes. Similarly, the YeastRACT database contained several transcription factors with an excessive number of annotated targets (Yap1 alone was annotated with over 1,500). To reduce the incidence of false positives, those studies which contributed over 100 targets for a given factor were discarded (on a per factor basis). While this may eliminate some true interactions, the goal is to generate a smaller set of high-confidence interactions which may be used to accurately assess the activity of given transcription factor. A hypergeometric test was used to assess the enrichment of each gene set in the lists of differentially expressed or sensitive genes.

Since the true number of differentially expressed or sensitive genes was unknown and poorly defined, we varied the cutoff for significance between 100 and 500 genes. The minimal p-value for each gene set was returned, and the activity/sensitivity of each gene set was reported as the negative log of this minimal p-value. Since the corresponding p-value was no longer strictly accurate as a consequence of multiple hypothesis testing, significance was assessed by repeated randomization trials in which the order of genes was shuffled. Every gene set was tested and the maximum significance value was retained in each trial. Only those gene sets which exceeded the 95th quantile in this set were determined to be significant.

Supplemental Tables

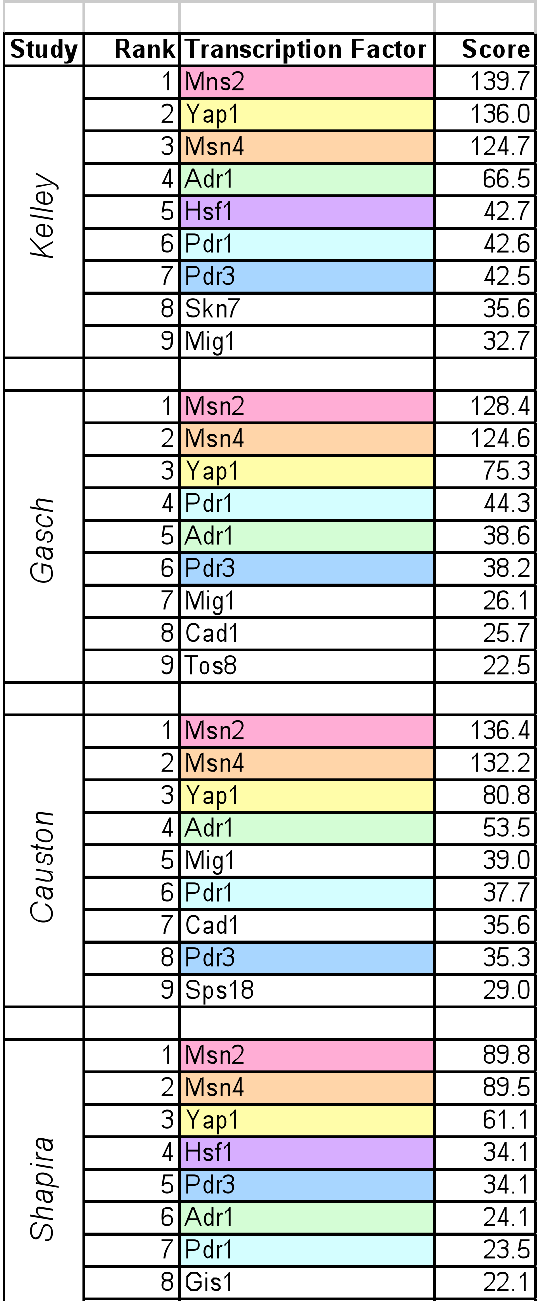
Supplemental Table 6.1. Sensitive gene ontology categories following acute hydrogen peroxide stress.

For our study and the study of Thorpe *et al.*, we determined those gene ontology categories which were enriched for sensitive gene deletions. Here we report all categories which exceed the threshold for significance.

	Sensitive Gene Ontology Categories
Thorpe *et al*	mitochondrial ribosome(C)
	aerobic respiration(P)
	mitochondrial protein processing(P)
	nucleoid(C)
	mitochondrial respiratory chain complex III(C)
	tRNA aminoacylation for protein translation(P)
	mitochondrial genome maintenance(P)
Kelley *et al*	mitochondrial ribosome(C)
	aerobic respiration(P)
	mitochondrial transport(P)
	protein processing(P)
	small ribosomal subunit(C)
	double-strand break repair via homologous recombination(P)
	mitochondrial genome maintenance(P)
	endosome membrane(C)
	tRNA aminoacylation for protein translation(P)
	double-strand break repair via single-strand annealing(P)

Supplemental Table 6.2. Up-regulated transcription factor target sets following acute hydrogen peroxide stress.

For our and previous comparable studies (Gasch 2000, Causton 2001, Shapira 2004), the set of known targets for each transcription factor was ranked based on enrichment for genes with increased expression in response to acute hydrogen peroxide stress. Here, we report the top nine sets of transcription factor targets. To facilitate comparison, frequently occurring items are high-lighted in a consistent manner.



Study	Rank	Transcription Factor	Score
Kelley	1	Mns2	139.7
	2	Yap1	136.0
	3	Msn4	124.7
	4	Adr1	66.5
	5	Hsf1	42.7
	6	Pdr1	42.6
	7	Pdr3	42.5
	8	Skn7	35.6
	9	Mig1	32.7
Gasch	1	Msn2	128.4
	2	Msn4	124.6
	3	Yap1	75.3
	4	Pdr1	44.3
	5	Adr1	38.6
	6	Pdr3	38.2
	7	Mig1	26.1
	8	Cad1	25.7
	9	Tos8	22.5
Causton	1	Msn2	136.4
	2	Msn4	132.2
	3	Yap1	80.8
	4	Adr1	53.5
	5	Mig1	39.0
	6	Pdr1	37.7
	7	Cad1	35.6
	8	Pdr3	35.3
	9	Sps18	29.0
Shapira	1	Msn2	89.8
	2	Msn4	89.5
	3	Yap1	61.1
	4	Hsf1	34.1
	5	Pdr3	34.1
	6	Adr1	24.1
	7	Pdr1	23.5
	8	Gis1	22.1

Supplemental Table 6.3. Up- and down-regulated gene ontology categories following acute hydrogen peroxide stress.

For our and previous comparable studies (Gasch 2000, Causton 2001, Shapira 2004), a pruned set of functional categories was ranked based on enrichment for genes with increased and decreased expression in response to acute hydrogen peroxide stress. In each case, we report the top five categories. To facilitate comparison, frequently occurring categories are high-lighted in a consistent manner.

	Gene Ontology Categories	
	Up-regulated	Down-regulated
Kelley	proteasome complex(C)	cytosolic large ribosomal subunit(C)
	response to toxin(P)	small subunit processome(C)
	response to oxidative stress(P)	maturation of SSU-rRNA(P)
	pentose metabolic process(P)	ribosomal large subunit biogenesis and assembly(P)
	carbohydrate catabolic process(P)	cytosolic small ribosomal subunit(C)
	alcohol catabolic process(P)	ribosome assembly(P)
	aldehyde metabolic process(P)	ribonucleoside monophosphate metabolic process(P)
	glycogen metabolic process(P)	nucleobase metabolic process(P)
	mitochondrial intermembrane space(C)	DNA-directed RNA polymerase I complex(C)
Gasch	aldehyde metabolic process(P)	cytosolic large ribosomal subunit(C)
	response to oxidative stress(P)	cytosolic small ribosomal subunit(C)
	response to toxin(P)	ribosomal subunit assembly(P)
	mitochondrial intermembrane space(C)	ribosomal large subunit biogenesis and assembly(P)
	glutathione metabolic process(P)	cell wall(C)
	pentose metabolic process(P)	amine transport(P)
	amino acid derivative catabolic process(P)	carboxylic acid transport(P)
	carbohydrate catabolic process(P)	snRNA metabolic process(P)
	vitamin biosynthetic process(P)	U4/U6 x U5 tri-snRNP complex(C)
Shapira	response to oxidative stress(P)	cytosolic large ribosomal subunit(C)
	aldehyde metabolic process(P)	cytosolic small ribosomal subunit(C)
	trehalose metabolic process(P)	ribosomal large subunit biogenesis and assembly(P)
	response to toxin(P)	ribosome assembly(P)
	siderophore transport(P)	ribosomal small subunit biogenesis and assembly(P)
	vacuolar lumen(C)	small subunit processome(C)
	amino acid derivative catabolic process(P)	ribonucleoside monophosphate biosynthetic process(P)
	proteasome complex(C)	maturation of SSU-rRNA from tricistronic rRNA transcript(P)
	sulfur metabolic process(P)	nucleolar preribosome(C)
Causton	response to oxidative stress(P)	cytosolic large ribosomal subunit(C)
	response to toxin(P)	cytosolic small ribosomal subunit(C)
	pentose metabolic process(P)	small subunit processome(C)
	carbohydrate catabolic process(P)	ribosomal large subunit biogenesis and assembly(P)
	trehalose metabolic process(P)	ribosome assembly(P)
	alcohol catabolic process(P)	maturation of SSU-rRNA(P)
	aldehyde metabolic process(P)	cleavages during rRNA processing(P)
	glutathione metabolic process(P)	maturation of 5.8S rRNA from tricistronic rRNA transcript(P)

Acknowledgements

We thank Dr. Jean Wang and Dr. Richard Kolodner for their insightful comments. This work was generously supported by grant ES014811 from the National Institute of Environmental Health Sciences (NIEHS). TI is a David and Lucille Packard Fellow.

Chapter 6, in full, is a copy of the following manuscript currently under preparation,

Kelley R, Ideker T. Genome-wide fitness and expression profiling implicate Mga2 in adaptation to hydrogen peroxide. **In preparation**.

The dissertation author is the sole first author on this work, responsible for designing and executing experiments and computational algorithms.

Chapter 7. Conclusions

In this work, we investigated several strategies for improving the analysis of high-throughput biological data sources. First, by improving our statistical analysis methods, we can mitigate the problem of "multiple hypothesis testing." In addition, we showed how high-throughput physical and genetic interaction screens can be used to uncover pathways of genes. This type of pathway information can be used to enhance the analysis of gene expression and deletion fitness profiling data collected in studies of various cellular stress responses. As technology continues to advance, additional analytical breakthroughs will need to be made to keep pace.

Genome-wide expression levels are now routinely assayed with the use of expression microarrays. However, new technologies on the horizon promise to alter the way in which we determine this information. In RNA-Seq, high-throughput sequencing technologies are used to rapidly sequence all of the RNA in a particular biological sample. Although the sequencing reads are relatively short (~50 bp), this is typically more than enough information to map a sequence to a location in the genome. Exact sequence counts can be used to precisely determine relative abundance levels of transcripts across different biological samples[230]. In *Saccharomyces cerevisiae* the benefits of this approach are modest, as there are relatively few genes and little sequence modification of those genes[4]. However, in large genomes with high prevalence of multiple splicing events, RNA-Seq has a large advantage over expression arrays in that it is difficult to design a probe to query every possible splicing event. However, statistical analysis of RNA-Seq data is still a developing field. By definition, gene-specific dye bias should not occur, as there is no differential dye labeling. However, we saw that underlying sequence-specific effects were the cause of this bias. One further avenue for RNA-Seq analysis is to investigate how sequence-specific effects might likewise affect those results.

In Chapters 2 and 3, we investigated the definition of pathway information from combined sources of protein binding and genetic interactions. In Chapter 2, the set of physical interactions contained metabolic, regulatory, as well as binding interactions. However, in the latter work, only the physical binding interactions were utilized. In order to address the problem of integrating additional physical interaction types into the algorithm of Chapter 3, a simple approach would be to perform regressions on additional types of interaction data, incorporating this additional evidence into the within-pathway score. In fact, this is quite similar to Bayesian approaches that already exist[231, 232].

However, this raises an important question: is it necessary to treat genetic interactions as a special case in our pathway determination algorithm? Alternatively, the same regression could be performed on genetic interactions, incorporating the resulting score into the within-pathway score, ignoring the entire between-pathway component. Pathways could then be defined with a simple hierarchical clustering of this combined network. However, some benefits of the current approach (such as a global map of the genetic interactions between complexes) would be lost. This highlights a compromise in the more general algorithm; we lose the ability to use specific information about particular data types. We saw an example of this in Chapter 5, where we were able to combine a specific type of pathway definition with metabolic interactions to identify genes involved in the arsenic response.

Moreover, it seems that genetic interactions themselves represent a fundamentally different type of interaction than physical interactions. Even though this is the case, it may not be necessary to treat genetic interactions as a special class in our prediction algorithm. Rather, we can evaluate the ability of other interaction types to predict the same kind of information that is present in a genetic interaction. Thus, for each type of interaction, we would run two regressions, one for the within-pathway score and one for the between-pathway score. One of the challenges in proceeding with such an approach is identifying a suitable "gold standard" of between-pathway genetic interaction pairs.

In Chapter 5, we searched for activated pathways in physical networks. One limitation is that each type of interaction network was searched separately. In order to utilize the information in each of these networks simultaneously, we can apply the same Bayesian methods mentioned previously to generate a combined network. However, as currently implemented, our analysis methods cannot be applied to such a network. The main problem is that these methods are designed to work with binary interaction networks, while a Bayesian network contains quantitative log-likelihood ratios (LLRs). While we could simply set a threshold on the LLR values to create a binary network, this would discard useful information. To utilize such a network, one possible approach is to first generate a hierarchical clustering of the interaction network. Any merge point of this hierarchical clustering defines a potential pathway in the network, which may be assessed for enrichment of differentially expressed or sensitive genes. In our analysis from Chapter 5, we learned that the definition of a pathway is important. If it is too liberal, the true signal is drowned out by spurious results, while if the definition is too strict, it will not be able to capture the true underlying pathways present in the network. This approach has the potential to strike a balance between these two restrictions. In this hierarchical method, a major unresolved issue is the identification of a suitable network-clustering algorithm. Average-link hierarchical clustering is simple and efficient, but the density of the resultant clusters is often suboptimal.

In Chapter 6, we focused on the identification of activated transcriptional networks. The general strategy was to identify specific activated transcription factors by looking for enrichment of differentially expressed genes in transcription factor targets gene sets. However, individual transcription factors rarely work alone. In reality, a specific regulon of genes is often under control of multiple transcription factors[233]. By identifying the constituents of these regulons, we may be able to more accurately identify the activated transcription factors. Unfortunately, this is a difficult task, as high-throughput transcription factor target data contains many false positives. One feature of these regulons is that they should be co-expressed across a large number of conditions. Thus, it should be possible to mine for these gene sets in the growing library of expression profiles present in a standard expression databases[234]. Although the transcription factors controlling these regulons will be unknown, databases of known transcription factor targets may be sufficient for annotating such information.

Chapters 5 and 6 are both concerned with the analysis of deletion fitness profiling information. Such a dataset can be misleading. Even if a particular gene plays a role in the response, the corresponding deletion strain may not appear sensitive if another gene is able to compensate for the loss of function. One way to address this problem is to look at paired deletion strains subjected to a specific stress, that is, condition-specific genetic interactions. However, additional computational methods are needed to analyze such experiments. Given the error rates in determination of subtle genetic interactions, special care needs to be taken to ensure that the identified interactions truly represent a condition-specific occurrence.

We have shown that using high-throughput interaction screen to learn pathway information confers clear advantages in the analysis of high-throughput genetic assays. Continuing to work on these types of analyses should provide further gains in the future. Fortunately, these advances tend to rely on incorporating data that is already publicly available. Thus, with little additional cost, it is possible to generate more knowledge out of expensive experimental data. The final challenge for bioinformaticians is to make sure that these tools are widely distributed and accessible, maximizing the utilization of both their work, and the public databases upon which they rely.

REFERENCES

1. Sanger, F. & Coulson, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94**, 441-448 (1975).

2. Alwine, J.C., Kemp, D.J. & Stark, G.R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A* **74**, 5350-5354 (1977).

3. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246 (1989).

4. Fisk, D.G. et al. Saccharomyces cerevisiae S288C genome annotation: a working hypothesis. *Yeast* **23**, 857-865 (2006).

5. Young, R.A. Biomedical discovery with DNA arrays. *Cell* **102**, 9-15 (2000).

6. Golemis, E. & Adams, P.D. Protein-protein interactions : a molecular cloning manual, Edn. 2nd. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.; 2005).

7. von Mering, C. et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403 (2002).

8. Tong, A.H. et al. Global mapping of the yeast genetic interaction network. *Science* **303**, 808-813 (2004).

9. Schuldiner, M. et al. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**, 507-519 (2005).

10. Ideker, T. et al. Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science* **292**, 929-934 (2001).

11. Jamieson, D.J. Saccharomyces cerevisiae has distinct adaptive responses to both hydrogen peroxide and menadione. *J Bacteriol* **174**, 6678-6681 (1992).

12. Jenner, P. Oxidative stress in Parkinson's disease. *Ann Neurol* **53 Suppl 3**, S26-36; discussion S36-28 (2003).

13. Markesbery, W.R. Oxidative stress hypothesis in Alzheimer's disease. *Free Radic Biol Med* **23**, 134-147 (1997).

14. Christen, Y. Oxidative stress and Alzheimer disease. *The American journal of clinical nutrition* **71**, 621S-629S (2000).

15. Avery, L. & Wasserman, S. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet* **8**, 312-316 (1992).

16. Guarente, L. Synthetic enhancement in gene interaction: a genetic tool come of age. *Trends Genet* **9**, 362-366 (1993).

17. Thomas, J.H. Thinking about genetic redundancy. *Trends Genet* **9**, 395-399 (1993).

18. Hartman, J.L., Garvik, B. & Hartwell, L. Principles for the buffering of genetic variation. *Science* **291**, 1001-1004. (2001).

19. Sham, P. Shifting paradigms in gene-mapping methodology for complex traits. *Pharmacogenomics* **2**, 195-202 (2001).

20. Dolma, S., Lessnick, S.L., Hahn, W.C. & Stockwell, B.R. Identification of genotype-selective antitumor agents using synthetic lethal chemical screening in engineered human tumor cells. *Cancer Cell* **3**, 285-296 (2003).

21. Forsburg, S.L. The art and design of genetic screens: yeast. *Nat Rev Genet* **2**, 659-668 (2001).

22. Tong, A.H. et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364-2368. (2001).

23. Ooi, S.L., Shoemaker, D.D. & Boeke, J.D. DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nat Genet* **35**, 277-286 (2003).

24. Giot, L. et al. A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727-1736 (2003).

25. Ito, T. et al. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A* **97**, 1143-1147 (2000).

26. Li, S. et al. A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540-543 (2004).

27. Rain, J.C. et al. The protein-protein interaction map of Helicobacter pylori. *Nature* **409**, 211-215. (2001).

28. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-627 (2000).

29. Gavin, A.-C., Bösche, M., Krause, R., Grandi, P. & Marzioch, M. Functional

organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147 (2002).

30. Ho, Y. et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-183 (2002).

31. Harbison, C.T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104 (2004).

32. Lee, T.I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804 (2002).

33. Ozier, O., Amin, N. & Ideker, T. Global architecture of genetic interactions on the protein network. *Nat Biotechnol* **21**, 490-491 (2003).

34. Tucker, C.L. & Fields, S. Lethal combinations. *Nat Genet* **35**, 204-205 (2003).

35. Winzeler, E.A. et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-906 (1999).

36. Mewes, H.W. et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**, 31-34. (2002).

37. Xenarios, I. et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30**, 303-305. (2002).

38. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res* **30**, 42-46 (2002).

39. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).

40. GOConsortium Creating the gene ontology resource: design and implementation. *Genome Res* **11**, 1425-1433. (2001).

41. Zar, J.H. Biostatistical Analysis, Edn. 3rd. (Prentice Hall, New Jersey; 1996).

42. Kendall, S.M., Stuart, A. & Ord, J.K. Kendall's Advanced Theory of Statistics, Edn. 5. (Oxford University Press, NY; 1987).

43. Milo, R. et al. Network motifs: simple building blocks of complex networks. *Science* **298**, 824-827 (2002).

44. Geissler, S., Siegers, K. & Schiebel, E. A novel protein complex promoting formation of functional alpha- and gamma-tubulin. *Embo J* **17**, 952-966 (1998).

45. Kahana, J.A. et al. The yeast dynactin complex is involved in partitioning the mitotic spindle between mother and daughter cells during anaphase B. *Mol Biol Cell* **9**, 1741-1756 (1998).

46. Pidoux, A.L. & Allshire, R.C. Centromeres: getting a grip of chromosomes. *Curr Opin Cell Biol* **12**, 308-319 (2000).

47. Pfeffer, S.R. Membrane transport: retromer to the rescue. *Curr Biol* **11**, R109-111 (2001).

48. Siniossoglou, S., Peak-Chew, S.Y. & Pelham, H.R. Ric1p and Rgp1p form a complex that catalyses nucleotide exchange on Ypt6p. *Embo J* **19**, 4885-4894 (2000).

49. Hwang, W.W. et al. A conserved RING finger protein required for histone H2B monoubiquitination and cell size control. *Mol Cell* **11**, 261-266 (2003).

50. Sharan, R., Ideker, T., Kelley, B.P., Shamir, R. & Karp, R.M. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology--RECOMB*, 282-289 (2004).

51. Carter, G.W. et al. Prediction of phenotype and gene expression for combinations of mutations. *Mol Syst Biol* **3**, 96 (2007).

52. Hereford, L.M. & Hartwell, L.H. Sequential gene function in the initiation of Saccharomyces cerevisiae DNA synthesis. *Journal of molecular biology* **84**, 445-461 (1974).

53. Collins, S.R. et al. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**, 806-810 (2007).

54. Collins, S.R., Schuldiner, M., Krogan, N.J. & Weissman, J.S. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol* **7**, R63 (2006).

55. Drees, B.L. et al. Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol* **6**, R38 (2005).

56. St Onge, R.P. et al. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* **39**, 199-206 (2007).

57. Segre, D., Deluna, A., Church, G.M. & Kishony, R. Modular epistasis in yeast metabolism. *Nat Genet* **37**, 77-83 (2005).

58. Beyer, A., Bandyopadhyay, S. & Ideker, T. Integrating physical and genetic

maps: from genomes to interaction networks. *Nat Rev Genet* **8**, 699-710 (2007).

59. Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* **23**, 561-566 (2005).

60. Ulitsky, I. & Shamir, R. Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. *Mol Syst Biol* **3**, 104 (2007).

61. Zhang, L.V. et al. Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. *Journal of biology* **4**, 6 (2005).

62. Phillips, P.C., Otto, S.P., Whitlock, M.C. Beyond the Average: the Evolutionary Importance of Gene Interactions and Variability of Epistatic Effects in Epistasis and Evolutionary Process. (Oxford Univ. Press, New York; 2000).

63. Collins, S.R. et al. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* **6**, 439-450 (2007).

64. Guldener, U. et al. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res* **34**, D436-441 (2006).

65. Boone, C., Bussey, H. & Andrews, B.J. Exploring genetic interactions and networks with yeast. *Nat Rev Genet* **8**, 437-449 (2007).

66. Driscoll, R., Hudson, A. & Jackson, S.P. Yeast Rtt109 promotes genome stability by acetylating histone H3 on lysine 56. *Science* **315**, 649-652 (2007).

67. Han, J. et al. Rtt109 acetylates histone H3 lysine 56 and functions in DNA replication. *Science* **315**, 653-655 (2007).

68. Otero, G. et al. Elongator, a multisubunit component of a novel RNA polymerase II holoenzyme for transcriptional elongation. *Mol Cell* **3**, 109-118 (1999).

69. Winkler, G.S., Kristjuhan, A., Erdjument-Bromage, H., Tempst, P. & Svejstrup, J.Q. Elongator is a histone H3 and H4 acetyltransferase important for normal histone acetylation levels in vivo. *Proc Natl Acad Sci U S A* **99**, 3517-3522 (2002).

70. Mitchell, P., Petfalski, E., Shevchenko, A., Mann, M. & Tollervey, D. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. *Cell* **91**, 457-466 (1997).

71. Wood, A. et al. Bre1, an E3 ubiquitin ligase required for recruitment and substrate selection of Rad6 at a promoter. *Mol Cell* **11**, 267-274 (2003).

72. Kobor, M.S. et al. A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. *PLoS Biol* **2**,

E131 (2004).

73. Krogan, N.J. et al. The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. *Mol Cell* **11**, 721-729 (2003).

74. Mizuguchi, G. et al. ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* **303**, 343-348 (2004).

75. Li, B., Carey, M. & Workman, J.L. The role of chromatin during transcription. *Cell* **128**, 707-719 (2007).

76. Dover, J. et al. Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6. *J Biol Chem* **277**, 28368-28371 (2002).

77. Sun, Z.W. & Allis, C.D. Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature* **418**, 104-108 (2002).

78. Berger, M.F. et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**, 1429-1435 (2006).

79. Ptacek, J. et al. Global analysis of protein phosphorylation in yeast. *Nature* **438**, 679-684 (2005).

80. Brem, R.B., Storey, J.D., Whittle, J. & Kruglyak, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**, 701-703 (2005).

81. Sokal , R.R., Michener C. D. A statistical method for evaluating systematic relationships. *University of Kansas Sci. Bull.* **28**, 1409-1438 (1958).

82. Benjamini, Y., Hochberg, Y Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSSB* **57**, 289-300 (1995).

83. Cline, M.S. et al. Integration of biological networks and gene expression data using Cytoscape. *Nature protocols* **2**, 2366-2382 (2007).

84. Demeter, J. et al. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res* **35**, D766-770 (2007).

85. Quackenbush, J. Microarray data normalization and transformation. *Nat Genet* **32 Suppl**, 496-501 (2002).

86. Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. & Wong, W.H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* **29**, 2549-2557 (2001).

87. Yang, Y.H. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15 (2002).

88. Dobbin, K.K., Kawasaki, E.S., Petersen, D.W. & Simon, R.M. Characterizing dye bias in microarray experiments. *Bioinformatics* **21**, 2430-2437 (2005).

89. Dombkowski, A.A., Thibodeau, B.J., Starcevic, S.L. & Novak, R.F. Gene-specific dye bias in microarray reference designs. *FEBS Lett* **560**, 120-124 (2004).

90. Rosenzweig, B.A. et al. Dye bias correction in dual-labeled cDNA microarray gene expression measurements. *Environ Health Perspect* **112**, 480-487 (2004).

91. Hekstra, D., Taussig, A.R., Magnasco, M. & Naef, F. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res* **31**, 1962-1968 (2003).

92. Naef, F. & Magnasco, M.O. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E Stat Nonlin Soft Matter Phys* **68**, 011906 (2003).

93. Martin-Magniette, M.L., Aubert, J., Cabannes, E. & Daudin, J.J. Evaluation of the gene-specific dye bias in cDNA microarray experiments. *Bioinformatics* **21**, 1995-2000 (2005).

94. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96-104 (2002).

95. Ideker, T., Thorsson, V., Siegel, A.F. & Hood, L.E. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J Comput Biol* **7**, 805-817 (2000).

96. Rocke, D.M. & Durbin, B. A model for measurement error for gene expression arrays. *J Comput Biol* **8**, 557-569 (2001).

97. Press, W.H. & Numerical Recipes Software (Firm) Numerical recipes in C, Edn. 2nd. (Cambridge University Press, Cambridge, England ; New York, N.Y.; 1997).

98. Wu H, K.K., Churchill GA in The Analysis of Gene Expression Data: An Overview of Methods and Software. (ed. G.E. Parmigiani G, Irizarry RA, Zeger SL) 313-431 (Springer, New York; 2003).

99. Cui, X., Hwang, J.T., Qiu, J., Blades, N.J. & Churchill, G.A. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics (Oxford, England)* **6**, 59-75 (2005).

100. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

101. Delmar, P., Robin, S. & Daudin, J.J. VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics* **21**, 502-508 (2005).

102. Irizarry, R.A. et al. Multiple-laboratory comparison of microarray platforms. *Nature methods* **2**, 345-350 (2005).

103. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: Systems Biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343-372 (2001).

104. Begley, T.J., Rosenbach, A.S., Ideker, T. & Samson, L.D. Damage recovery pathways in Saccharomyces cerevisiae revealed by genomic phenotyping and interactome mapping. *Mol Cancer Res* **1**, 103-112 (2002).

105. Birrell, G.W. et al. Transcriptional response of Saccharomyces cerevisiae to DNA-damaging agents does not identify the genes that protect against these agents. *Proc Natl Acad Sci U S A* **99**, 8778-8783 (2002).

106. Birrell, G.W., Giaever, G., Chu, A.M., Davis, R.W. & Brown, J.M. A genome-wide screen in Saccharomyces cerevisiae for genes affecting UV radiation sensitivity. *Proc Natl Acad Sci U S A* **98**, 12608-12613 (2001).

107. National Research Council (U.S.). Subcommittee on Arsenic in Drinking Water & ebrary Inc Arsenic in drinking water 2001 update. (National Academy Press, Washington, DC; 2001).

108. Smith, A.H. et al. Cancer risks from arsenic in drinking water. *Environ Health Perspect* **97**, 259-267 (1992).

109. United States. Dept. of Health and Human Services, United States. Public Health Service, United States. Agency for Toxic Substances and Disease Registry & Syracuse Research Corporation Toxicological profile for arsenic. (U.S. Dept. of Health and Human Services Public Health Service Agency for Toxic Substances and Disease Registry, Atlanta, Ga.; 2000).

110. Agis, H. et al. Successful treatment with arsenic trioxide of a patient with ATRA-resistant relapse of acute promyelocytic leukemia. *Annals of hematology* **78**, 329-332 (1999).

111. Shen, Z.X. 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**, 3354-3360 (1997).

112. Zhang, P. The use of arsenic trioxide (As2O3) in the treatment of acute promyelocytic leukemia. *Journal of biological regulators and homeostatic agents* **13**, 195-200 (1999).

113. Brown, J.L. & Kitchin, K.T. Arsenite, but not cadmium, induces ornithine decarboxylase and heme oxygenase activity in rat liver: relevance to arsenic carcinogenesis. *Cancer letters* **98**, 227-231 (1996).

114. Hamadeh, H.K., Trouba, K.J., Amin, R.P., Afshari, C.A. & Germolec, D. Coordination of altered DNA repair and damage pathways in arsenite-exposed keratinocytes. *Toxicol Sci* **69**, 306-316 (2002).

115. Kitchin, K.T. Recent advances in arsenic carcinogenesis: modes of action, animal model systems, and methylated arsenic metabolites. *Toxicology and applied pharmacology* **172**, 249-261 (2001).

116. Liu, S.X., Athar, M., Lippai, I., Waldren, C. & Hei, T.K. Induction of oxyradicals by arsenic: implication for mechanism of genotoxicity. *Proc Natl Acad Sci U S A* **98**, 1643-1648 (2001).

117. Lynn, S., Gurr, J.R., Lai, H.T. & Jan, K.Y. NADH oxidase activation is involved in arsenite-induced oxidative DNA damage in human vascular smooth muscle cells. *Circulation research* **86**, 514-519 (2000).

118. Matsui, M. et al. The role of oxidative DNA damage in human arsenic carcinogenesis: detection of 8-hydroxy-2'-deoxyguanosine in arsenic-related Bowen's disease. *The Journal of investigative dermatology* **113**, 26-31 (1999).

119. Shi, H., Shi, X. & Liu, K.J. Oxidative mechanism of arsenic toxicity and carcinogenesis. *Molecular and cellular biochemistry* **255**, 67-78 (2004).

120. Vogt, B.L. & Rossman, T.G. Effects of arsenite on p53, p21 and cyclin D expression in normal human fibroblasts -- a possible mechanism for arsenite's comutagenicity. *Mutation research* **478**, 159-168 (2001).

121. Wang, T.S., Kuo, C.F., Jan, K.Y. & Huang, H. Arsenite induces apoptosis in Chinese hamster ovary cells by generation of reactive oxygen species. *Journal of cellular physiology* **169**, 256-268 (1996).

122. Bobrowicz, P., Wysocki, R., Owsianik, G., Goffeau, A. & Ulaszewski, S. Isolation of three contiguous genes, ACR1, ACR2 and ACR3, involved in resistance to arsenic compounds in the yeast Saccharomyces cerevisiae. *Yeast* **13**, 819-828 (1997).

123. Rosen, B.P. Families of arsenic transporters. *Trends in microbiology* **7**, 207-212 (1999).

124. Ghosh, M., Shen, J. & Rosen, B.P. Pathways of As(III) detoxification in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **96**, 5001-5006 (1999).

125. Wysocki, R., Bobrowicz, P. & Ulaszewski, S. The Saccharomyces cerevisiae ACR3 gene encodes a putative membrane protein involved in arsenite transport. *J Biol Chem* **272**, 30061-30066 (1997).

126. Wysocki, R. et al. Transcriptional activation of metalloid tolerance genes in Saccharomyces cerevisiae requires the AP-1-like proteins Yap1p and Yap8p. *Mol Biol Cell* **15**, 2049-2060 (2004).

127. Menezes, R.A., Amaral, C., Delaunay, A., Toledano, M. & Rodrigues-Pousada, C. Yap8p activation in Saccharomyces cerevisiae under arsenic conditions. *FEBS Lett* **566**, 141-146 (2004).

128. Cohen, B.A., Pilpel, Y., Mitra, R.D. & Church, G.M. Discrimination between paralogs using microarray analysis: application to the Yap1p and Yap2p transcriptional networks. *Mol Biol Cell* **13**, 1608-1614 (2002).

129. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18 Suppl 1**, S233-240 (2002).

130. Lee, J. et al. Yap1 and Skn7 control two specialized oxidative stress response regulons in yeast. *J Biol Chem* **274**, 16040-16046 (1999).

131. Xenarios, I. et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30**, 303-305 (2002).

132. Forster, J., Famili, I., Fu, P., Palsson, B.O. & Nielsen, J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res* **13**, 244-253 (2003).

133. Hermann-Le Denmat, S., Werner, M., Sentenac, A. & Thuriaux, P. Suppression of yeast RNA polymerase III mutations by FHL1, a gene coding for a fork head protein involved in rRNA processing. *Mol Cell Biol* **14**, 2905-2913 (1994).

134. Gasch, A.P. et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**, 4241-4257 (2000).

135. Jelinsky, S.A., Estep, P., Church, G.M. & Samson, L.D. Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes. *Mol Cell Biol* **20**, 8157-8167 (2000).

136. Amoros, M. & Estruch, F. Hsf1p and Msn2/4p cooperate in the expression of

Saccharomyces cerevisiae genes HSP26 and HSP104 in a gene- and stress type-dependent manner. *Molecular microbiology* **39**, 1523-1532 (2001).

137. Boy-Marcotte, E. et al. The heat shock response in yeast: differential regulations and contributions of the Msn2p/Msn4p and Hsf1p regulons. *Molecular microbiology* **33**, 274-283 (1999).

138. Fernandes, L., Rodrigues-Pousada, C. & Struhl, K. Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. *Mol Cell Biol* **17**, 6982-6993 (1997).

139. Hasan, R. et al. The control of the yeast H2O2 response by the Msn2/4 transcription factors. *Molecular microbiology* **45**, 233-241 (2002).

140. Raitt, D.C. et al. The Skn7 response regulator of Saccharomyces cerevisiae interacts with Hsf1 in vivo and is required for the induction of heat shock genes by oxidative stress. *Mol Biol Cell* **11**, 2335-2347 (2000).

141. Coleman, S.T., Epping, E.A., Steggerda, S.M. & Moye-Rowley, W.S. Yap1p activates gene transcription in an oxidant-specific fashion. *Mol Cell Biol* **19**, 8302-8313 (1999).

142. Estruch, F. Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. *FEMS microbiology reviews* **24**, 469-486 (2000).

143. Grey, M. & Brendel, M. Overexpression of the SNQ3/YAP1 gene confers hyper-resistance to nitrosoguanidine in Saccharomyces cerevisiae via a glutathione-independent mechanism. *Current genetics* **25**, 469-471 (1994).

144. Schnell, N. & Entian, K.D. Identification and characterization of a Saccharomyces cerevisiae gene (PAR1) conferring resistance to iron chelators. *European journal of biochemistry / FEBS* **200**, 487-493 (1991).

145. Schnell, N., Krems, B. & Entian, K.D. The PAR1 (YAP1/SNQ3) gene of Saccharomyces cerevisiae, a c-jun homologue, is involved in oxygen metabolism. *Current genetics* **21**, 269-273 (1992).

146. Alarco, A.M., Balan, I., Talibi, D., Mainville, N. & Raymond, M. AP1-mediated multidrug resistance in Saccharomyces cerevisiae requires FLR1 encoding a transporter of the major facilitator superfamily. *J Biol Chem* **272**, 19304-19313 (1997).

147. Coleman, S.T., Tseng, E. & Moye-Rowley, W.S. Saccharomyces cerevisiae basic region-leucine zipper protein regulatory networks converge at the ATR1 structural gene. *J Biol Chem* **272**, 23224-23230 (1997).

148. Grant, C.M., Maciver, F.H. & Dawes, I.W. Stationary-phase induction of GLR1

expression is mediated by the yAP-1 transcriptional regulatory protein in the yeast Saccharomyces cerevisiae. *Molecular microbiology* **22**, 739-746 (1996).

149. Kuge, S. & Jones, N. YAP1 dependent activation of TRX2 is essential for the response of Saccharomyces cerevisiae to oxidative stress by hydroperoxides. *Embo J* **13**, 655-664 (1994).

150. Kuge, S., Jones, N. & Nomoto, A. Regulation of yAP-1 nuclear localization in response to oxidative stress. *Embo J* **16**, 1710-1720 (1997).

151. Wu, A.L. & Moye-Rowley, W.S. GSH1, which encodes gamma-glutamylcysteine synthetase, is a target gene for yAP-1 transcriptional regulation. *Mol Cell Biol* **14**, 5832-5839 (1994).

152. Baumeister, W., Walz, J., Zuhl, F. & Seemuller, E. The proteasome: paradigm of a self-compartmentalizing protease. *Cell* **92**, 367-380 (1998).

153. Russell, S.J., Steger, K.A. & Johnston, S.A. Subcellular localization, stoichiometry, and protein levels of 26 S proteasome subunits in yeast. *J Biol Chem* **274**, 21943-21952 (1999).

154. Hochstrasser, M. et al. The Saccharomyces cerevisiae ubiquitin-proteasome system. *Philosophical transactions of the Royal Society of London* **354**, 1513-1522 (1999).

155. Kornitzer, D. & Ciechanover, A. Modes of regulation of ubiquitin-mediated protein degradation. *Journal of cellular physiology* **182**, 1-11 (2000).

156. Mannhaupt, G., Schnall, R., Karpov, V., Vetter, I. & Feldmann, H. Rpn4p acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26S proteasomal and other genes in yeast. *FEBS Lett* **450**, 27-34 (1999).

157. Owsianik, G., Balzi l, L. & Ghislain, M. Control of 26S proteasome expression by transcription factors regulating multidrug resistance in Saccharomyces cerevisiae. *Molecular microbiology* **43**, 1295-1308 (2002).

158. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-254 (2003).

159. Bobrowicz, P. Arsenical-induced transcriptional activation of the yeast *Saccharomyces cerevisiae ACR3* genes requires the presence of the *ACR1* gene product. *Cell Mol Biol Lett.* **3**, 13-20 (1998).

160. Bouganim, N., David, J., Wysocki, R. & Ramotar, D. Yap1 overproduction restores arsenite resistance to the ABC transporter deficient mutant ycf1 by activating ACR3 expression. *Biochemistry and cell biology = Biochimie et*

biologie cellulaire **79**, 441-448 (2001).

161. Dormer, U.H. et al. Cadmium-inducible expression of the yeast GSH1 gene requires a functional sulfur-amino acid regulatory network. *J Biol Chem* **275**, 32611-32616 (2000).

162. Causton, H.C. et al. Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell* **12**, 323-337 (2001).

163. Fauchon, M. et al. Sulfur sparing in the yeast proteome in response to sulfur demand. *Mol Cell* **9**, 713-723 (2002).

164. Cavigelli, M. et al. The tumor promoter arsenite stimulates AP-1 activity by inhibiting a JNK phosphatase. *Embo J* **15**, 6269-6279 (1996).

165. Giaever, G. et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-391 (2002).

166. Bentley, R. The shikimate pathway--a metabolic tree with many branches. *Critical reviews in biochemistry and molecular biology* **25**, 307-384 (1990).

167. Herrmann, K.M. & Weaver, L.M. The Shikimate Pathway. *Annual review of plant physiology and plant molecular biology* **50**, 473-503 (1999).

168. Roberts, C.W. et al. The shikimate pathway and its branches in apicomplexan parasites. *The Journal of infectious diseases* **185 Suppl 1**, S25-36 (2002).

169. DeRisi, J. et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14**, 457-460 (1996).

170. Hauser, N.C. et al. Transcriptional profiling on all open reading frames of Saccharomyces cerevisiae. *Yeast* **14**, 1209-1221 (1998).

171. Shcherbakova, P.V. et al. Inactivation of DNA mismatch repair by increased expression of yeast MLH1. *Mol Cell Biol* **21**, 940-951 (2001).

172. Hewitt, S.C. et al. Estrogen receptor-dependent genomic responses in the uterus mirror the biphasic physiological response to estrogen. *Molecular endocrinology (Baltimore, Md* **17**, 2070-2083 (2003).

173. Bushel, P.R. et al. MAPS: a microarray project system for gene expression experiment information and data validation. *Bioinformatics* **17**, 564-565 (2001).

174. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-14868 (1998).

175. Kelley, B.P. et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A* **100**, 11394-11399 (2003).

176. Wu, H.I., Brown, J.A., Dorie, M.J., Lazzeroni, L. & Brown, J.M. Genome-wide identification of genes conferring resistance to the anticancer agents cisplatin, oxaliplatin, and mitomycin C. *Cancer research* **64**, 3940-3948 (2004).

177. Lusis, A.J. Atherosclerosis. *Nature* **407**, 233-241 (2000).

178. Weindruch, R., Naylor, P.H., Goldstein, A.L. & Walford, R.L. Influences of aging and dietary restriction on serum thymosin alpha 1 levels in mice. *Journal of gerontology* **43**, B40-42 (1988).

179. Harman, D. Aging: a theory based on free radical and radiation chemistry. *Journal of gerontology* **11**, 298-300 (1956).

180. Lin, S.J. et al. Calorie restriction extends Saccharomyces cerevisiae lifespan by increasing respiration. *Nature* **418**, 344-348 (2002).

181. Schulz, T.J. et al. Glucose restriction extends Caenorhabditis elegans life span by inducing mitochondrial respiration and increasing oxidative stress. *Cell metabolism* **6**, 280-293 (2007).

182. Kim, D.K., Cho, E.S., Lee, B.R. & Um, H.D. NF-kappa B mediates the adaptation of human U937 cells to hydrogen peroxide. *Free Radic Biol Med* **30**, 563-571 (2001).

183. Lee, B.R. & Um, H.D. Hydrogen peroxide suppresses U937 cell death by two different mechanisms depending on its concentration. *Experimental cell research* **248**, 430-438 (1999).

184. Wiese, A.G., Pacifici, R.E. & Davies, K.J. Transient adaptation of oxidative stress in mammalian cells. *Arch Biochem Biophys* **318**, 231-240 (1995).

185. Masoro, E.J. Caloric restriction and aging: an update. *Experimental gerontology* **35**, 299-305 (2000).

186. Masoro, E.J. Overview of caloric restriction and ageing. *Mechanisms of ageing and development* **126**, 913-922 (2005).

187. Costa, V. & Moradas-Ferreira, P. Oxidative stress and signal transduction in Saccharomyces cerevisiae: insights into ageing, apoptosis and diseases. *Mol Aspects Med* **22**, 217-246 (2001).

188. Meister, A. Glutathione metabolism and its selective modification. *J Biol Chem* **263**, 17205-17208 (1988).

189. Muller, E.G. Thioredoxin genes in Saccharomyces cerevisiae: map positions of TRX1 and TRX2. *Yeast* **8**, 117-120 (1992).

190. Gan, Z.R. Yeast thioredoxin genes. *J Biol Chem* **266**, 1692-1696 (1991).

191. Alvarez-Peral, F.J., Zaragoza, O., Pedreno, Y. & Arguelles, J.C. Protective role of trehalose during severe oxidative stress caused by hydrogen peroxide and the adaptive oxidative stress response in Candida albicans. *Microbiology* **148**, 2599-2606 (2002).

192. Benaroudj, N., Lee, D.H. & Goldberg, A.L. Trehalose accumulation during cellular stress protects cells and cellular proteins from damage by oxygen radicals. *J Biol Chem* **276**, 24261-24267 (2001).

193. Petrova, V.Y., Drescher, D., Kujumdzieva, A.V. & Schmitt, M.J. Dual targeting of yeast catalase A to peroxisomes and mitochondria. *Biochem J* **380**, 393-400 (2004).

194. Hartig, A. & Ruis, H. Nucleotide sequence of the Saccharomyces cerevisiae CTT1 gene and deduced amino-acid sequence of yeast catalase T. *European journal of biochemistry / FEBS* **160**, 487-490 (1986).

195. Bermingham-McDonogh, O., Gralla, E.B. & Valentine, J.S. The copper, zinc-superoxide dismutase gene of Saccharomyces cerevisiae: cloning, sequencing, and biological activity. *Proc Natl Acad Sci U S A* **85**, 4789-4793 (1988).

196. Ravindranath, S.D. & Fridovich, I. Isolation and characterization of a manganese-containing superoxide dismutase from yeast. *J Biol Chem* **250**, 6107-6112 (1975).

197. Stephen, D.W., Rivers, S.L. & Jamieson, D.J. The role of the YAP1 and YAP2 genes in the regulation of the adaptive oxidative stress responses of Saccharomyces cerevisiae. *Molecular microbiology* **16**, 415-423 (1995).

198. Jackson, A.L. & Loeb, L.A. The contribution of endogenous sources of DNA damage to the multiple mutations in cancer. *Mutation research* **477**, 7-21 (2001).

199. Stephen, D.W. & Jamieson, D.J. Glutathione is an important antioxidant molecule in the yeast Saccharomyces cerevisiae. *FEMS Microbiol Lett* **141**, 207-212 (1996).

200. Kistler, M., Summer, K.H. & Eckardt, F. Isolation of glutathione-deficient mutants of the yeast Saccharomyces cerevisiae. *Mutation research* **173**, 117-120 (1986).

201. Izawa, S., Inoue, Y. & Kimura, A. Oxidative stress response in yeast: effect of glutathione on adaptation to hydrogen peroxide stress in Saccharomyces cerevisiae. *FEBS Lett* **368**, 73-76 (1995).

202. Alic, N. et al. Genome-wide transcriptional responses to a lipid hydroperoxide: adaptation occurs without induction of oxidant defenses. *Free Radic Biol Med* **37**, 23-35 (2004).

203. Grant, C.M., MacIver, F.H. & Dawes, I.W. Mitochondrial function is required for resistance to oxidative stress in the yeast Saccharomyces cerevisiae. *FEBS Lett* **410**, 219-222 (1997).

204. Flattery-O'Brien, J., Collinson, L.P. & Dawes, I.W. Saccharomyces cerevisiae has an inducible response to menadione which differs from that to hydrogen peroxide. *J Gen Microbiol* **139**, 501-507 (1993).

205. Shapira, M., Segal, E. & Botstein, D. Disruption of yeast forkhead-associated cell cycle transcription by oxidative stress. *Mol Biol Cell* **15**, 5659-5669 (2004).

206. Ng, C.H. et al. Adaptation to hydrogen peroxide in Saccharomyces cerevisiae: the role of NADPH-generating systems and the SKN7 transcription factor. *Free Radic Biol Med* **44**, 1131-1145 (2008).

207. Shoemaker, D.D., Lashkari, D.A., Morris, D., Mittmann, M. & Davis, R.W. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet* **14**, 450-456 (1996).

208. Thorpe, G.W., Fong, C.S., Alic, N., Higgins, V.J. & Dawes, I.W. Cells have distinct mechanisms to maintain protection against different reactive oxygen species: oxidative-stress-response genes. *Proc Natl Acad Sci U S A* **101**, 6564-6569 (2004).

209. Teixeira, M.C. et al. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res* **34**, D446-451 (2006).

210. Jiang, Y. et al. MGA2 is involved in the low-oxygen response element-dependent hypoxic induction of genes in Saccharomyces cerevisiae. *Mol Cell Biol* **21**, 6161-6169 (2001).

211. Zitomer, R.S. & Lowry, C.V. Regulation of gene expression by oxygen in Saccharomyces cerevisiae. *Microbiological reviews* **56**, 1-11 (1992).

212. Daum, G., Lees, N.D., Bard, M. & Dickson, R. Biochemistry, cell biology and molecular biology of lipids of Saccharomyces cerevisiae. *Yeast* **14**, 1471-1510 (1998).

213. Branco, M.R., Marinho, H.S., Cyrne, L. & Antunes, F. Decrease of H2O2 plasma membrane permeability during adaptation to H2O2 in Saccharomyces cerevisiae. *J Biol Chem* **279**, 6501-6506 (2004).

214. Lyons, T.J. et al. Metalloregulation of yeast membrane steroid receptor homologs. *Proc Natl Acad Sci U S A* **101**, 5506-5511 (2004).

215. Tafforeau, L. et al. Repression of ergosterol level during oxidative stress by fission yeast F-box protein Pof14 independently of SCF. *Embo J* **25**, 4547-4556 (2006).

216. Matias, A.C. et al. Down-regulation of fatty acid synthase increases the resistance of Saccharomyces cerevisiae cells to H2O2. *Free Radic Biol Med* **43**, 1458-1465 (2007).

217. Janki, R.M., Aithal, H.N., McMurray, W.C. & Tustanoff, E.R. The effect of altered membrane-lipid composition on enzyme activities of outer and inner mitochondrial membranes of Saccharomyces cerevisiae. *Biochem Biophys Res Commun* **56**, 1078-1085 (1974).

218. Hermann, G.J. & Shaw, J.M. Mitochondrial dynamics in yeast. *Annu Rev Cell Dev Biol* **14**, 265-303 (1998).

219. Ter Linde, J.J. & Steensma, H.Y. A microarray-assisted screen for potential Hap1 and Rox1 target genes in Saccharomyces cerevisiae. *Yeast* **19**, 825-840 (2002).

220. Keng, T. HAP1 and ROX1 form a regulatory pathway in the repression of HEM13 transcription in Saccharomyces cerevisiae. *Mol Cell Biol* **12**, 2616-2623 (1992).

221. Hon, T. et al. A mechanism of oxygen sensing in yeast. Multiple oxygen-responsive steps in the heme biosynthetic pathway affect Hap1 activity. *J Biol Chem* **278**, 50771-50780 (2003).

222. Nagababu, E. & Rifkind, J.M. Heme degradation by reactive oxygen species. *Antioxidants & redox signaling* **6**, 967-978 (2004).

223. Dirmeier, R. et al. Exposure of yeast cells to anoxia induces transient oxidative stress. Implications for the induction of hypoxic genes. *J Biol Chem* **277**, 34773-34784 (2002).

224. Davies, J.M., Lowry, C.V. & Davies, K.J. Transient adaptation to oxidative stress in yeast. *Arch Biochem Biophys* **317**, 1-6 (1995).

225. Kelley, R., Feizi, H. & Ideker, T. Correcting for gene-specific dye bias in DNA microarrays using the method of maximum likelihood. *Bioinformatics* (2007).

226. Kaiser, C., Cold Spring Harbor Laboratory. & Adams, A. Methods in yeast genetics : a Cold Spring Harbor Laboratory course manual, Edn. 1997. (Cold Spring Harbor Laboratory Press, Plainview, N.Y.; 1998).

227. Yuan, D.S. et al. Improved microarray methods for profiling the Yeast Knockout strain collection. *Nucleic Acids Res* **33**, e103 (2005).

228. Arthington-Skaggs, B.A., Jradi, H., Desai, T. & Morrison, C.J. Quantitation of ergosterol content: novel method for determination of fluconazole susceptibility of Candida albicans. *Journal of clinical microbiology* **37**, 3332-3337 (1999).

229. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

230. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628 (2008).

231. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555-1558 (2004).

232. Lee, I., Li, Z. & Marcotte, E.M. An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. *PLoS ONE* **2**, e988 (2007).

233. Segal, E. et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**, 166-176 (2003).

234. Barrett, T. et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **37**, D885-890 (2009).