Computational analysis of genetic interaction network structures and gene properties

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Elizabeth Natalie Koch

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advised by Chad L. Myers

July, 2017

Acknowledgements

I thank my advisor, Chad Myers, who has been incredibly generous with his support; I am grateful for his kind and patient guidance. I am also thankful for the support and advice of Michael Costanzo and Charlie Boone, without whom my work would not have been possible. I additionally thank my committee members, Judy Berman, Rui Kuang, and Dan Knights, for their time and suggestions.

My lab mates have made many positive contributions to my research. I particularly thank Raamesh Deshpande and Jeremy Bellay, who both have a knack for expressing insightful observations about research and who guided me when I first began my work. I also thank Benjamin VanderSluis, Robert Schaefer, Scott Simpkins, Wen Wang, Stephanie DiPrima, Trina Kuriger-Laber, Justin Nelson, Jean-Michel Michno, Roman Briskine, Colin Pesyna, Hamid Safizadeh, Maximilian Billmann, and Mahfuzur Rahman. Everyone has been generous in sharing their skills, suggestions, and friendship. I wish all of them success in future pursuits.

The completion of my research has benefited from skills I was taught before entering graduate school. I am grateful to many former teachers who emphasized the importance of precision and clarity in reasoning, reading, and writing, and I often think back to their individual varieties of advice and wisdom. In this respect, I sincerely thank my former professors at Carleton College, particularly my undergraduate advisor David Liben-Nowell, and my teachers from Lincoln Academy.

I will forever be grateful to Matt, who never failed to be the best part of each day. Most of all, I thank my parents and sister for their constant love.

Abstract

Cellular systems are responsible for many complex tasks, such as carrying out cell cycle phases, responding to intra- and extra-cellular conditions, and resolving errors. Through analysis of biological networks, researchers have begun to describe how cells coordinate these processes by means of modularity and between-process connections. However, descriptions of this network-based cellular organization often do not incorporate the diverse characteristics and individual behaviors of the genes that compose it. Knowledge of gene properties and their relationships with biological network evolution is crucial for a complete understanding of cellular function, and investigation in this area can lead to general principles of biology that apply to many species. This dissertation will describe analyses of the *Saccharomyces cerevisiae* (baker's yeast) genetic interaction network that connect gene topological behavior with various physical, functional, and evolutionary properties of genes. Genetic interactions occur between paired genes whose simultaneous mutations produce unexpected double-mutant phenotypes, which are indicative of a range of functional relationships. Because genetic interactions can be identified genome-wide in high-throughput experiments, their networks are comprehensive and unbiased representations of function to which we can apply computational methods that search for structure-function relationships.

We begin by exploring the association between a set of gene properties and gene genetic interaction (GI) degree. Here, we build a decision tree model that sorts genes based on a set of properties, each of which has a correlation with GI degree, and accurately predicts GI degree. We show that our model, trained on *S. cerevisiae*, is also accurate for a very distant yeast species, *Schizosaccharomyces pombe*, demonstrating that the rules governing gene connectivity are well conserved. Finally, we used predictions from the model to identify gene modules that differ between the two yeast species.

Next, we further characterize hub genes through an investigation of pleiotropy, the phenomenon of a single genetic locus with multiple phenotypic effects. Pleiotropy has typically been described by counting organism-level phenotypes, but a characterization based on genetic interactions can capture details about cellular processes that are buffered by the cell and never manifest in single mutant cellular phenotypes. For this analysis, we use frequent item set mining to discover GI modules,

which we annotate with high-level processes, and use entropy to measure the functional diversity of each gene's set of containing modules, thus distinguishing between genes whose functional influence is limited to very few bioprocesses and those whose roles are important for varied cellular functions. We identified a number of gene and protein characteristics that differed between genes with high and low pleiotropy and discuss the implications of these results regarding the nature and evolution of pleiotropy.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The cellular processes that support all forms of biological life are dependent on networks of physical and functional relationships of genes and proteins. Structures in biological networks reflect the mechanisms by which cells create highly complex, yet resilient, systems that are persistent throughout evolution. The most salient network structures reveal the primary cellular organization of sets of genes working as modules to carry out cellular functions (Hartwell et al., 1999; Ravasz et al., 2002). This modular structure is hierarchical and contains links between modules to ensure that temporally and physically distinct processes are coordinated. Network structure also contributes to cellular robustness (Hartman et al., 2001; Rutherford and Lindquist, 1998; Stelling et al., 2004), which means that for survival of an organism, cells must maintain phenotypic stability through appropriate responses to external environments, such as toxins, temperatures, and osmotic pressure, and internal conditions, such as mutations and stochastic events that perturb normal cell processes (Wagner, 2005). Finally, for the persistence of populations through generations of changing conditions, the network organization must allow for adaptability, the flexibility of a genome to evolve in the face of natural selection and inhabit a specific niche (Kirschner and Gerhart, 1998; Rutherford, 2003). While evidence of complex organization in biological networks has long been known, these three aspects, modularity, robustness, and adaptability, are incompletely described. In particular, these properties are highly dependent on each other and overlapping in their effects on and requirements of network structure, but their precise relationships have yet to be described.

Over the past two decades, the use of new experimental technologies to detect genome sequences and mutationally target specific genes has led to the construction of large biological networks. One such network is the genetic interaction (GI) network of the yeast *Saccharomyces cerevisiae*, in which genes are connected to each other if there is phenotypic evidence in mutant yeast strains suggesting the two genes are functionally related to each other. Although careful inspection of the network has led to discovery of new functions for genes, the size of the network, with approximately one million edges among nearly 6,000 genes, makes many valuable analyses difficult or impossible to do manually. These include comparing the network to other datasets and systematic identification of structures.

In this dissertation, we discuss the discovery of principles of gene behaviors in yeast genetic interaction network structure and how these behaviors are related to gene and network evolution. There are two broad foundations for this work that are described in this chapter. The first is biological networks, specifically the genetic interaction network of yeast, from which we can discover gene function and importance. Section 1.1 of this chapter describes the most common components of biological networks and section 1.2 describes how biological networks are thought to change through evolution. The second foundation for this dissertation is the highly-studied model organism *S. cerevisiae*, which is briefly described at the beginning of section 1.3. The remainder of section 1.3 describes the diverse set of gene characteristics that give indications of how biological systems evolve and the behaviors of network nodes that accomplish cellular tasks. Finally, the concluding section, section 1.4, reintroduces the purpose of this dissertation within the context of the background material.

1.1 Components of biological networks

Biological networks represent physical and functional associations between molecules in a cell. They comprise chemical reactions, physical structures, and even information flow between different physical locations and distinct types of molecules in the cell (Zhu et al., 2007). Consequently, there are a number of conceptually distinct biological networks that researchers study. Protein-protein interaction (PPI) networks are the most widely studied and involve physical interactions between proteins (Braun and Gingras, 2012). General PPI networks are complemented by other physical networks involving specific types of proteins and other macromolecules: regulatory networks are directed networks that regulate cell functions in response to stimuli (Pawson and Nash, 2003); transcription networks are directed networks in which transcription factors activate and suppress gene expression through DNA binding (Thieffry et al., 1998); and metabolic networks describe enzymes, metabolites, and conversions between metabolites (Hatzimanikatis et al., 2004). Lastly, some networks contain more abstract connections, such as experimentally derived genetic interaction networks (Tong et al., 2004) and computationally derived functional networks. Co-expression networks are one of the latter; these are constructed by connecting genes with similar patterns of expression. This section introduces two of the largest networks, PPI and GI networks,

which have been systematically constructed from experiments, and describes how they elucidate the functions of a cell.

1.1.1 Discovery of interactions

Protein-protein interaction networks

Protein-protein interactions are mainly detected by two types of high-throughput methods: binary assays, which detect pairwise interactions, and affinity purification followed by mass-spectrometry (AP-MS), which detects proteins in stable complexes. The interactions discovered from these methods are complementary and have quite different interpretations. The yeast two-hybrid system (Y2H) (Fields and Song, 1989), the most frequently used method to construct binary networks, consists of systematic screens that use a bait and prey set-up: two domains of a transcription factor, the DNA binding domain and the transcription activation domain, are separately attached to the two proteins of interest. These fusion proteins are then expressed in yeast cells. If the two proteins of interest bind each other, the transcription factor is reconstituted and, together, its domains activate a reporter gene that causes a growth-based phenotype in the yeast colony. There are variations to this method, such as the protein complementation assay, in which the two tested proteins are fused to fragments of a fluorescent protein. Because of the engineered systems for detection, these binary interactions are direct physical interactions that *may* naturally occur in cells of the proteins' native species, but do not necessary do so. Large-scale networks have been constructed in a number of species, including *S. cerevisiae* (Ito et al., 2001; Uetz et al., 2000; Yu et al., 2008), *Schizosaccharomyces pombe* (Vo et al., 2016), worm (Li et al., 2004), fly (Giot et al., 2003), and human (Rolland et al., 2014).

In contrast, affinity purification methods developed in yeast are designed to isolate protein complexes from cells under physiological conditions, which include normal post-translational modifications made to proteins. For these methods, reviewed in Smits and Vermeulen (2016), a protein of interest—the bait protein—is fused to an epitope tag and inserted into its original genomic position using homologous recombination. The affinity purification step is performed by preparing a cell lysate and drawing out the bait protein by catching its epitope tag with a binding (or high-affinity) protein. Any proteins that are bound to the bait protein are simultaneously captured, and

then are identified with mass spectrometry analysis. Importantly, the captured proteins may include both direct interacting partners of the bait protein and proteins that simply participate in a complex with the bait. In order to distinguish individual protein complexes from among the collection of proteins associated with a single bait protein, further data must be collected for different bait proteins. However, even with high-density data this is a difficult task and follow-up experiments, computational strategies, and dataset comparisons have been used to refine definitions of protein complexes (Gingras et al., 2007). Two landmark proteome-wide studies have been performed in yeast (Gavin et al., 2006; Krogan et al., 2006); each tagged about 2,000 bait proteins. Similar methods have subsequently been developed for high throughput screening of proteins of other model organisms (Duchaine et al., 2006; Rees et al., 2011; Veraksa et al., 2005) and human (Hein et al., 2015; Huttlin et al., 2015; Malovannaya et al., 2011; Wan et al., 2015), in which there are multiple cell lines and more challenges in achieving purification of bait proteins (Smits and Vermeulen, 2016).

Genetic interaction networks

A genetic interaction occurs between two genes if their simultaneous perturbation causes an unexpected phenotype that cannot be explained by a combination of the phenotypes measured after individually mutating the genes (Mani et al., 2008). Genetic interactions identify non-independence of genes and imply functional relationships, such as the ability of a gene to compensate for the loss of another. Although genetic interactions provide no immediate mechanistic information about protein function, their high level of abstraction means they are sensitive to complex and distant relationships between genes. Genetic interactions have been measured systematically in a number of species, including *S. pombe* (Frost et al., 2012; Roguev et al., 2008), *C. elegans* (Lehner et al., 2006), *M. drosophila* (Fischer et al., 2015; Horn et al., 2011), and human (Barbie et al., 2009; Vizeacoumar et al., 2013).

The most extensive networks of genetic interactions have been built for yeast genes using fitness (colony growth) as a quantitative phenotype (Baryshnikova et al., 2010b; Costanzo et al., 2010; Costanzo et al., 2016; Tong et al., 2004). The expected fitness of a double mutant is calculated as the product of the fitnesses of the two associated single mutants, which has been observed for the vast majority of gene pairs. An extreme example of a deviation from this is cell death in a double-mutant strain

harboring two gene deletions that individually did not have lethal effects; this is known as synthetic lethality (Mani et al., 2008). More generally, a GI score is the difference of the observed and expected double mutant fitnesses (Baryshnikova et al., 2010b). A negative genetic interaction is a fitness deviation in which the measured double-mutant fitness is significantly lower than the prediction, indicating a phenotype sicker than expected. A positive genetic interaction occurs when the measured double-mutant fitness is significantly higher than the prediction, indicating a phenotype healthier than expected. The Synthetic Genetic Array (SGA) system uses robotics to automate the construction of colonies of double-mutant yeast strains by mating single-mutant strains. The recently completed SGA network includes 90% of all genes and ~75% of all gene pairs (Costanzo et al., 2016).

The array of all GI scores for a given strain (or gene), is called a genetic interaction profile and is a high-resolution functional description of a gene. One of the most powerful uses of the GI network is the derivation of a profile correlation network, in which genes are connected by weighted edges of Pearson's correlation coefficients between their GI profiles. The profile similarity network has high accuracy in predicting genes involved in the same function and its precision is comparable to that of AP-MS PPI networks, which contain proteins interacting in complexes. This is an improvement over individual interactions: at a 50% recall of predicting shared curated annotations between genes, profile similarities have a precision approximately 50% higher than that of negative genetic interactions (Baryshnikova et al., 2010b; Costanzo et al., 2010). Inspection of uncharacterized genes in the yeast GI profile similarity network has led to validated function predictions for many genes (Costanzo et al., 2010; Costanzo et al., 2016).

1.1.2 Organization through modularity

An observed edge in a biological network can often be functionally interpreted through its network context. Frequently, dense sets of interactions among groups of genes or proteins are reflective of biological modules.

Modules are groups of functionally related genes or proteins that contribute to a task (Hartwell et al., 1999). They include pathways of consecutive physical interactions, such as signaling pathways, stable protein complexes, such as the proteasome or

nuclear pore, or proteins whose combined functions carry out a specific task, such as initiating DNA replication or docking and merging vesicles. Some modules, such as checkpoints, regulate other modules. To varying extents, these different types of modules are identifiable in many networks. For instance, it is mechanistically clear that protein complexes appear as dense clusters in Y2H or AP-MS PPI networks (e.g. (Gavin et al., 2006)). They also can be identified in GI networks because negative genetic interactions are often dense among the members of a module that can tolerate the loss of one member, but not two (Baryshnikova et al., 2010b). Co-expression networks strongly reflect complexes because co-member proteins usually have closely matching regulatory patterns (Stuart et al., 2003). Modules are not restricted to physical associations. In co-expression and regulatory networks, genes that respond as a group to a condition make up modules due to their common function (Gasch and Eisen, 2002). GI networks in particular can represent a wide variety of relationships, such as forming a structure in which many negative genetic interactions occur between two modules that functionally compensate for each other, known as the between-pathway model (Kelley and Ideker, 2005).

Modules vary greatly in size are typically not distinct: commonly, they overlap in their members and are nested within each other. This is best illustrated with the genetic interaction profile similarity network, to which Costanzo et al. (2016) applied hierarchical clustering (Figure 1.1). Genes with highly correlated GI profiles, forming the smallest clusters, correspond to protein complexes and pathways, which are contained in larger clusters reflecting broader biological classes.

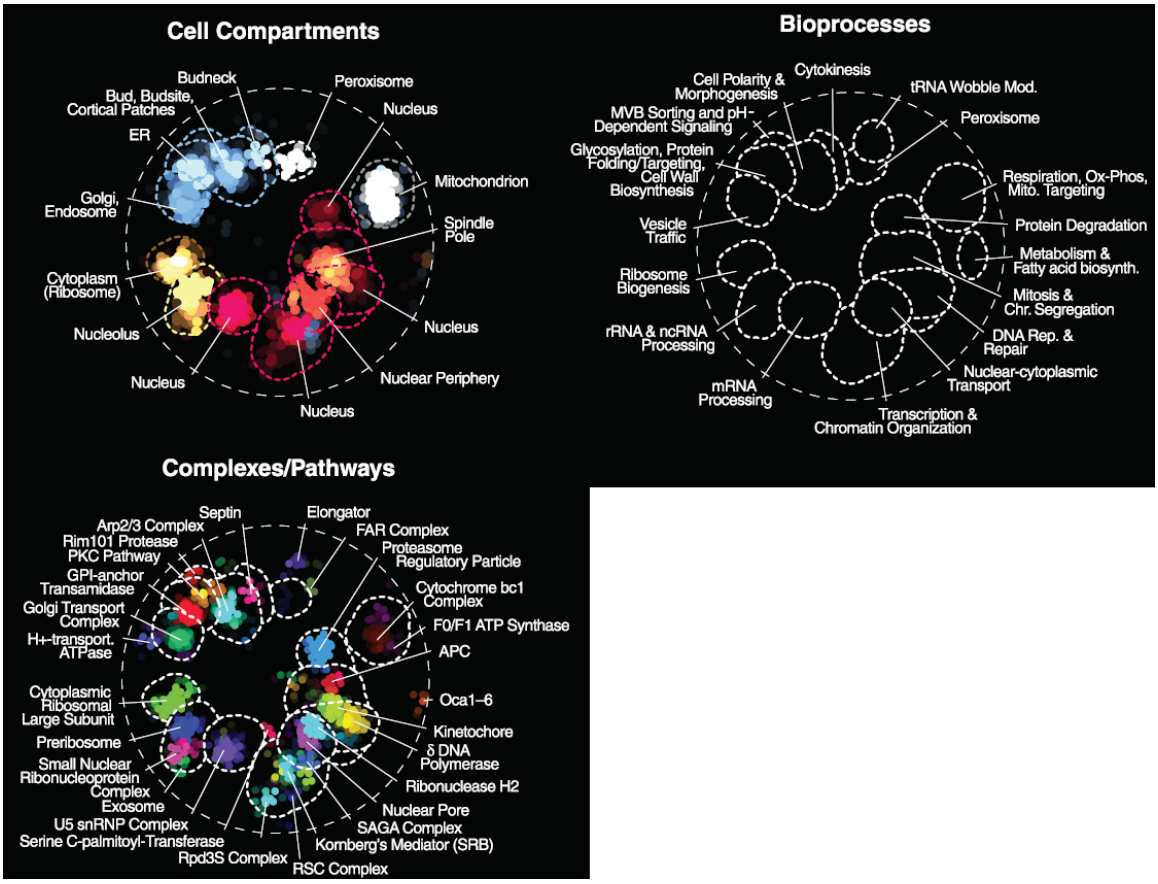


Figure 1.1. Hierarchical modularity in the yeast genetic interaction profile similarity network. Reproduced from Costanzo et al., 2016. A clustering algorithm was applied to the genetic interaction profile similarity network, determining clusters of genes and their positions as colored nodes here (network edges are not shown). Profile similarity clusters reveal groups of genes with known relationships. Large network clusters correspond to cellular compartments **(upper left)**, medium-sized clusters correspond to high-level biological processes **(upper right)**, and small tightly-connected clusters correspond to protein complexes and pathways **(lower left)**.

Within the modular landscape of gene and protein networks, there is heterogeneity of individual behaviors. Han et al. (2004) investigated the dynamic activity of hub proteins by comparing their expression profiles (expression levels measured in many conditions) to those of neighboring proteins in the yeast Y2H PPI network. The authors found a bimodal distribution across all hub proteins in their average expression correlation with neighbors, which indicates two types of hub behavior. Intramodular hubs

(also termed “party” hubs), which have high expression correlations with their physical interactors, were thought to simultaneously interact with many partners. Hub proteins with low average expression correlations with neighbors, intermodular (“date”) hubs, were thought to interact with neighbors at different times. The topology of these two behaviors had network-level effects: the simulated removal of date hub nodes from the PPI network caused higher increases in shortest path lengths than that of party hubs, which resembled removal of random nodes. Later investigations revealed that intramodular hubs have more binding sites on the protein surface than intermodular hubs (Kim et al., 2006) and confirmed the hub dichotomy using clusters in updated yeast interaction data and in other species (Chang et al., 2013; Pritykin and Singh, 2013).

Genetic interactions occur frequently between genes that have curated functions in different high-level bioprocesses, particularly involving chromatin, transcription, and Golgi-related genes (Costanzo et al., 2010). While modularity explains an organization of biological processes into units, between-process genetic interactions and the existence of intermodular hub proteins suggests an extensive coordination framework.

1.2 Evolution of biological networks

The structures and functions of biological networks are intimately linked with evolution. Broadly, some network connections are more vulnerable than others to the effects of evolution and are more likely to lead to species-specific functions and the gain or loss genetic robustness. From a practical perspective, an understanding of conservation will help determine when and how the functions of genes in one species may be applicable to the functions of orthologous genes in another species. For example, two genes interacting in a model organism might serve as evidence that their human orthologs also interact or suggest the mechanism underlying a disease phenotype (Walhout et al., 2000). The following three sections describe key aspects of our current understanding of biological network evolution. In order, they describe evolution at a local scale of genes, at an intermediate scale of modules, and finally, at a network level.

1.2.1 Mechanisms of duplicate gene divergence

Gene duplication is widely considered to be the primary source of new functions, including increasing genomic complexity and the divergence between species (Holland et al., 2017; Kimura and Ohta, 1974; Ohno, 1970). For example, Holland et al. (2017) reviewed examples in which specialized abilities of animals are closely associated with sets of duplicated genes, such as high-acuity vision in dragonflies, heat tolerance in clams during low tide, and increased neural development in cephalopods. Duplication events vary in scale, affecting single genes, multiple adjacent genes, or even the entire genome. Directly following duplication of a gene, both genes are able to perform all roles of the original, parent gene (Ohno, 1970). In some cases, multiple identical copies of a gene facilitates an advantageous increase in expression and both duplicates will be evolutionarily retained under purifying selection (Kondrashov and Kondrashov, 2006; Rapoport, 1940). More frequently, however, the selective pressure on the duplicate genes is unequal. Ohno (1970) first proposed that duplication leads to genetic redundancy and consequently different selective pressure against the two duplicate genes. Others have built on this idea, describing different scenarios in which coding-region mutations and selective pressure affect the retention of duplicate genes and the gain and loss of their functions (reviewed in (Conant and Wolfe, 2008)). To escape the loss of one gene, functional divergence of the two duplicates must occur, in which the genes acquire detrimental or adaptive mutations that lead to distinct functions (Force et al., 1999; Ohno, 1970). Models of divergence are described in terms of how these mutations affect the parent gene's functions. One possibility is neofunctionalization (Kimura and Ohta, 1974; Ohno, 1970), a process in which an adaptive mutation imbues one gene with a novel, non-parental function, leaving the other under high selective pressure to support all parental functions. Subfunctionalization (Force et al., 1999) is the process in which parental functions are partitioned between the duplicate pair, which is possible if the parent gene contained modular regions, such as domains, that carried out separable functions. If one gene receives a mutation that compromises its ability to perform a function, there will be increased selective pressure in the opposite gene to maintain this function. The isolated losses of function in the individual duplicates eventually eliminate genetic redundancy such that each parental function is only performed by one gene. The concepts of subfunctionalization and neofunctionalization

need not be exclusive: one or both genes can mature to incorporate new and old functions.

In addition to mutations within gene coding regions, regulatory changes can cause duplicate genes to diverge, even in a complementary manner (Force et al., 1999). One mechanism promoting evolution through regulatory divergence may be the development of novel protein interactions: temporal changes in protein abundance can affect the set of potential binding partners to which a protein is exposed. Experimental evidence has shown that swapping the regulatory regions of paralogous yeast genes can cause the interacting partners of the protein products to also switch, thus specifically implicating regulatory change as the mechanism of divergence (Gagnon-Arsenault et al., 2013). Various demonstrations of developmental genes being highly conserved and partially interchangeable in different metazoan phyla suggested transcriptional regulation has broad and substantial contributions to network evolution, even in the absence of coding-region sequence divergence (reviewed in Holland et al., 2017). (Adding to the apparently large impacts of transcriptional changes, it has been shown that the evolutionary rate of regulatory networks outpaces those of other biological networks (Shou et al., 2011).)

There is no consensus model describing divergence. However, a number of intriguing observations about functional relationships in duplicate gene pairs have been made along two broad conclusions. First, in yeast, whole-genome duplication (WGD) duplicates, which formed through an ancient event that caused duplication of the entire yeast genome, are functionally more similar to each other (Guan et al., 2007), more slowly evolving (Fares et al., 2013), more likely to be in protein complexes (Hakes et al., 2007), and more affected by dosage requirements (Gout and Lynch, 2015; Hakes et al., 2007) than small-scale duplication (SSD) duplicates. Second, there is a high occurrence of asymmetry between duplicates. For example, genetic interaction profiles have been used to identify divergent duplicates with significant levels of asymmetric functional importance (VanderSluis et al., 2010). Differing rates of sequence evolution have been observed in duplicate genes, likely reflecting cases in which a duplicate that evolves much more quickly than its partner is degenerate and has lost its ancestral functions (Kellis et al., 2004). Finally, duplicate genes tend to have transcriptional responses to stress conditions (Conant and Wolfe, 2006), and there is some evidence that most commonly, only one gene in a pair responds to stress (Mattenberger et al., 2017).

1.2.2 Rewiring of network modules

Patterns of network conservation between species are largely uncharacterized due to difficulties in comparing interaction networks of different species. However, there are clear evolutionary trends relating to modularity.

Protein complexes and other dense PPI modules are the most highly conserved structures in biological networks. These modules, which often represent core cellular functions, tend to have highly uniform compositions of nearly all essential genes or nearly all nonessential genes (Hart et al., 2007; Ryan et al., 2013). Based on this observation Zotenko et al. (2008) suggested that most essential genes earn their essential status through participation in modules. Consistently, multi-interface PPI hubs, which mainly represent complexes, are more likely to be essential and have a slower rate of sequence evolution as compared to all other PPI hubs (Kim et al., 2006). Phylogenetically, the genes associated with protein complexes are ubiquitous and widespread, with about two-thirds of metazoan protein complexes predicted to have ancient origins dating to at least the metazoan-fungi common ancestor (Wan et al., 2015). Evidence suggests that even in human, most complex-member proteins function in core cellular processes at far above the background rate, which may indicate conserved functions of these modules. All these observations establish the idea that evolutionary rates and statuses of genes are derived from the functions of dense PPI modules. Based this explanation, the majority of protein modules should be evolutionarily conserved, due to unchanging sequences of their constituent essential proteins (Hirsh and Fraser, 2001).

In contrast, relationships between modules are evolutionarily flexible. Kim et al. (2006) concluded that single-interface hubs, which are likely to interact with multiple modules, contribute to network growth because they are able to accommodate interactions with recently duplicated genes. By comparing genetic interaction networks of two very distantly related yeast species (*S. cerevisiae* and *S. pombe*) Roguev et al. (2008) gave an example of a set of related complexes that are individually conserved between the two species, but show different between-module interactions. The changes between the species can be partially explained by physiological differences between the two yeasts, suggesting that evolution of species can occur by new connections between

old modules. Vo et al. (2016) confirmed this idea systematically in comparisons of the complete *S. pombe* Y2H PPI network with the complete *S. cerevisiae* and genome-wide human Y2H networks. The authors used both GO biological process terms and topologically defined clusters to categorize protein interactions as within- or between-module connections. In all cases, within-module connections were much more likely to be conserved between species, with the most extreme difference occurring in the comparison between *S. pombe* and human: using network clusters, close to 90% of within-cluster interactions were conserved, but around 10% of between-cluster interactions were conserved.

The prominence of highly conserved modules and the fact that it is experimentally easier to demonstrate conservation than lack of conservation should not diminish the importance of between-module network connections. The paragraph above shows that module rewiring can enable species-specific features. Additionally, the connections between modules have substantial topological importance to biological networks. Analysis in a human AP-MS-derived PPI network showed that intermodular interactions are crucial to the connectivity and functionality of the global network (Hein et al., 2015). Interestingly, physical strength of protein interactions was shown to correlate with topological roles of the interactions in the PPI network, with an abundance of biophysically weaker interactions occurring between modules, which were composed of strong internal interactions. This is illustrated in Figure 1.2, replicated from Hein et al. (2015). The physical nature of the weak, between-module interactions may allow greater exploration of potential network rewiring, making them key elements in the adaptability of networks.

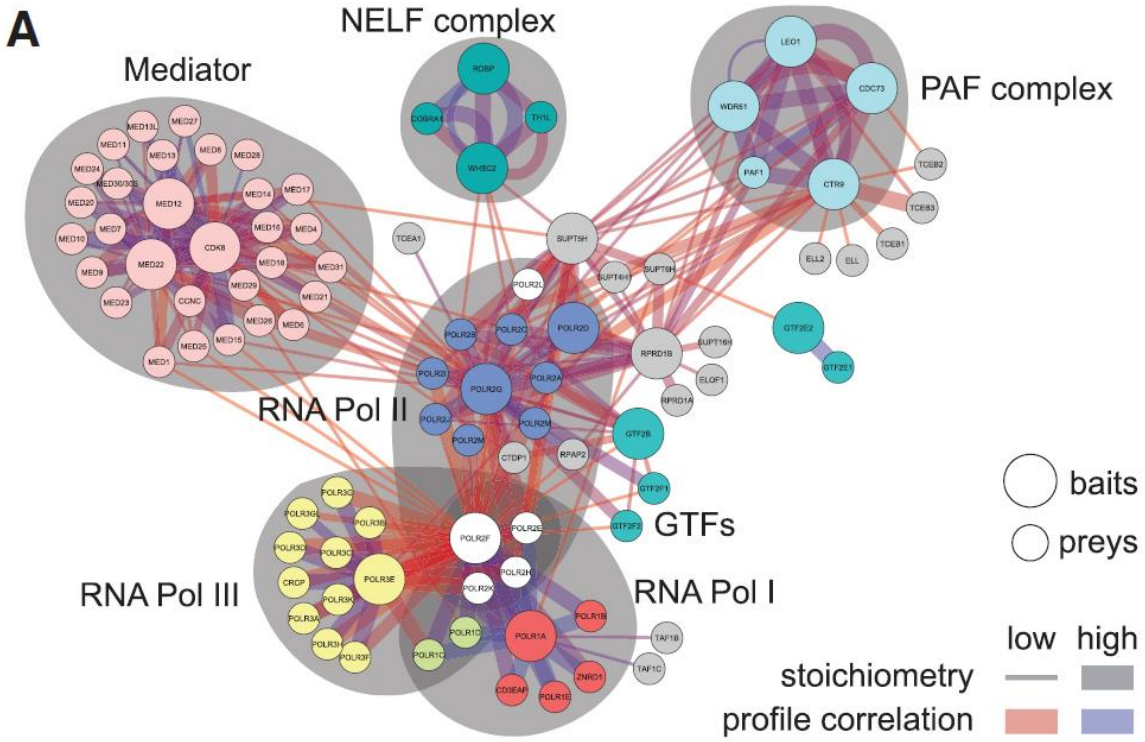


Figure 1.2. Strong and weak interactions within and between protein complex modules. Reproduced from Hein et al., 2015. The network of human protein-protein interactions identified in AP-MS experiments reveals connections between three RNA polymerases and other related proteins and complexes. Stoichiometry of a pair of proteins is highly predictive of the biophysical stability of an interaction. Thin edges indicate weak links and frequently occur between stable complexes.

1.2.3 Whole-network evolution

Despite the varied observations and anecdotal support for the divergence models of duplicate genes, and the observation of conservation within modules, explaining the evolution of PPI networks has proved to be a challenge. Gene duplication and divergence was used as the basis of many network growth models, which were developed with a central goal of producing networks with similar topology to experimentally derived networks (e.g. Middendorf et al., 2005; Pastor-Satorras et al., 2003; Rzhetsky and Gomez, 2001). In these iterative models, each growth step copies a randomly selected node and all its edges, then, with various parameter probabilities, deletes selected duplicated edges and adds new edges between the duplicate nodes

and randomly selected nodes. Additional rules incorporate biological constraints, such as requiring that at least one gene in duplicate pairs has an interaction with each neighbor of the pre-duplicated protein, mimicking the preservation of ancestral functions (Vázquez et al., 2003).

Kim and Marcotte (2008) objected to the many duplication and divergence growth models, crucially showing they produced networks in which the newest nodes have a high probability of interacting with the oldest nodes; in fact, proteins in the yeast PPI network show strong preference for interacting with proteins of similar age. Therefore they proposed the crystal growth model of network evolution, designed to produce interactions between nodes of similar age and to promote formation of network modules. The authors suggest physical justifications for both of these trends: proteins will form interactions primarily with other proteins that have available surface area, meaning that young proteins will tend to partner with each other, and proteins will form interactions with their neighbor's neighbors due to proximity. The crystal growth model is initialized with a few nodes and as the network grows, modules are continuously defined based on dense regions. When a node is added to the network, a module is randomly selected and new interactions are limited to connect only to nodes within the module, with higher probability initially given to (likely newer) nodes with low degree (with low probabilities, a node can form a new module or intermodular connections). Likely, the true manner of network growth incorporates ideas from both types of models.

1.3 *Saccharomyces cerevisiae* as a model organism

1.3.1 General description

The yeast *Saccharomyces cerevisiae* is single-celled eukaryote that has been used to study many cellular processes and structures that are fundamental to our understanding of biology. Experimental laboratory work on *S. cerevisiae* that included life cycle descriptions, strain isolation, and trait selection began in the 1930s and 1940s (Mortimer and Johnston, 1986). In a research lab, this fungus is typically grown in colonies on solid- or in liquid-nutrient mediums, and under ideal conditions cells replicate as frequently as once every 90 minutes (Duina et al., 2014). *S. cerevisiae* is somewhat flexible in its life history and metabolism: under different conditions, it can harvest energy

through fermentation and respiration, exist in both diploid and haploid states, and reproduce sexually and asexually. *S. cerevisiae* has an evolutionary distance estimated near one billion years from all metazoans (Chernikova et al., 2011) and with a genome of approximately 6,000 genes, its complexity is humble in comparison to human and metazoan model species, whose genomes contain from ~14,000 genes in fly (Adams et al., 2000) to ~19,000 genes in human and worm (C. elegans Sequencing Consortium, 1998; Ezkurdia et al., 2014).

An initial and enduring reason for *S. cerevisiae*'s popularity as a model organism is the ease with which researchers can perform targeted manipulations of gene sequences. The basis of these manipulations is homologous recombination (HR), a DNA repair pathway in which crossover of homologous sequences (e.g. from homologous chromosomes) repairs double-strand breaks. Specific gene modification is achieved through transforming cells with a plasmid containing short DNA sequences that match the targeted gene (Hinnen et al., 1978; Orr-Weaver et al., 1981; Rothstein, 1983). Through HR, yeast incorporates the plasmid DNA into its genome with high efficiency, completely replacing the targeted gene. Within the inserted sequence, selectable markers and reporter genes enable subsequent selection of mutant cells; a modified form of the target gene is often included in the insertion. The tools used in this method, including plasmid transformation and construction of effective vectors through PCR (Baudin et al., 1993; Longtine et al., 1998; Wach et al., 1994), were developed in yeast throughout the 1980s and 1990s. In 1996, yeast became the first eukaryote with a fully sequenced genome (Goffeau et al., 1996) and emerged as an ideal system for the systematic study of a complete set of genes. With the goal of determining functions for all genes, researchers created collections of single-gene mutant strains. Researchers were able to determine cellular localizations of all genes (Huh et al., 2003) by replacing genes with GFP-fusion versions, and by replacing genes with epitope-tagged versions, multiple groups identified protein complexes (Gavin et al., 2006; Ho et al., 2002; Krogan et al., 2006; Rigaut et al., 1999). A number of single-mutant strain collections were designed to compromise the function of individual genes in order to observe the resulting phenotypes. The deletion collection contains strains with selectable markers and barcode identification sequences in place of all nonessential genes (Giaever et al., 2002); the DAmP method causes depressed levels of expression by insertion of a marker in the 3' UTR of a gene, and has been applied to essential genes (Schuldiner et

al., 2005); the haploinsufficient collection contains heterozygous diploid strains, each with only one copy of an essential gene (Giaever et al., 1999); and many temperature-sensitive alleles along with selectable markers have replaced essential genes (Kofoed et al., 2015; Li et al., 2011). These *S. cerevisiae* strain collections, all constructed through HR-based methods, remain at the forefront of yeast genomics, since they are used in high-throughput endeavors, such as screening for genetic interactions (Costanzo et al., 2016) and chemical-genetic interactions (Hillenmeyer et al., 2008), and other projects. The efficiency of homologous recombination seen in *S. cerevisiae* does not exist for more complex model species, such as worm, fly, and mouse, which have species-specific complexities and low rates of homologous recombination due to cellular preference of alternative DNA repair pathways (Hardy et al., 2010).

Research in yeast has had many contributions the understanding of the cellular biology of higher eukaryotes. Firstly, many discoveries in yeast elucidate cellular structures and pathways that are highly conserved among eukaryotes and therefore applicable to understanding all eukaryotic cells at a fundamental level. Some of the most famous discoveries made in yeast, and those which lead to Nobel prizes, are Leland Hartwell's description of genes that advance the cell cycle and checkpoint genes that delay it (Pulverer, 2001), Roger Kornberg's structural elucidation of the molecular components that carry out transcription (Service, 2006), and Randy Schekman's identification of genes controlling vesicle trafficking and secretion (Ferro-Novick and Brose, 2013). However, it is impossible to quantify the influence that yeast has had on our understanding of eukaryotic biology since virtually all aspects of yeast cellular biology have some level of conservation in other eukaryotes. Secondly, important experimental technologies have been developed in yeast and applied to research of other organisms, such as the yeast two-hybrid method for detecting protein interactions (Fields and Song, 1989) and protein array technology (Ptacek et al., 2005; Zhu et al., 2001).

With the recent sequencing of many genomes and appreciation for the influence of genetic diversity within species, understanding genomes in an evolutionary context is currently of great importance. *S. cerevisiae* is well-poised for the study of comparative genomics. Dozens of *S. cerevisiae* strains that have evolved in widely varying natural and domesticated environments in countries all over the world have been sequenced (e.g. Liti et al., 2009), which lays a foundation for studying intraspecies genetic variation

(Peter and Schacherer, 2016). At a broader evolutionary scope, there is active genomics research on yeast species at many evolutionary distances (Zarin and Moses, 2014). Lastly, the important model yeast species *Schizosaccharomyces pombe*, which diverged from *S. cerevisiae* approximately 500 million years ago (Rhind et al., 2011), makes a particularly powerful comparison to *S. cerevisiae* because a collection of single-mutant deletion strains has been constructed for its nonessential genes. This means the fitness defects caused by gene loss can be measured as colony size with very similar methods to those used in *S. cerevisiae* (Baryshnikova et al., 2010a). Available for comparison, there are many characteristics of *S. pombe* that are present in higher eukaryotes but absent from *S. cerevisiae*, including RNA interference, a high proportion of genes containing introns, alternative splicing, and repetitive centromeres (Rhind et al., 2011).

1.3.2 Gene and protein characteristics for genomic analyses

There are many diverse ways to quantitatively describe genes in terms of their protein products, evolutionary histories, functional and phenotypic behaviors, and other descriptors. These gene characteristics are integral to much of the work presented in this dissertation, so we introduce them here, organized by the general methods used to collect them. Some characteristics have been analyzed and collected in many species; many *could* be obtained for any species, with varying amounts of experimental and computational effort. Therefore, the descriptions below discuss both generic methods and analyses that are specific to *S. cerevisiae*. Table 1.1 summarizes the gene characteristics in advance.

Table 1.1. Gene characteristics organized by type of measurement. All gene characteristics used in this document are represented by the entries of this table, though in some cases multiple precise characteristics are described by one item here. Most, but not all, items are described in the following section. Abbreviations: dN/dS, normalized rate of nonsynonymous mutations; PPI, protein-protein interaction; SSD, small-scale duplicate; WGD, whole-genome duplicate.

Sequence based	Genome-wide experimental data
Codon usage bias	Expression level
Protein domains	Expression variance
Protein disorder	Co-expression
Distance from telomere	Chemical genetic degree
dN/dS	Protein abundance
Single nucleotide polymorphisms	Phenotypic capacitance
Copy number	PPI degree
Duplicate status (SSD or WGD)	Single mutant fitness

Phylogenetic	Curated data
Age	Complex membership
Conservation	Phenotypes
Copy number volatility	GO terms

Sequence-based characteristics

Functional and evolutionary qualities of genes are, in some cases, detectable solely from sequence analysis. Codon usage bias is one such phenomenon in which genes preferentially contain specific codons instead of uniformly using alternative synonymous codons. Due to broad positive correlation between codon bias and gene expression, as well as matching non-uniform abundance of tRNAs, the prevailing explanation for codon bias is natural selection for increased translation efficiency and accuracy (Plotkin et al., 2011). Codon adaptation index (Sharp and Li, 1987), CAI, is a popular quantitative measure of codon bias that compares codon frequencies in a query gene to frequencies in a reference set of highly expressed genes, making CAI a strong predictor of expression level. A measurement termed “effective number of codons”, abbreviated N_c, similarly measures codon bias by counting the number of codons used, but is not calculated in comparison to highly expressed genes, making it sensitive to other causes of bias, such as a reduction of 5’-end secondary structure that interferes

with translation initiation or a decrease in translation rate to allow for co-translational folding or modification.

Aspects of protein structure can also be detected from gene sequence, after translation to amino acids. Protein domains are spatially compact and distinct structural units of proteins that are often associated with specific molecular functions. Domains tend to be highly conserved (intriguingly, over half of domain families observed in Eukarya also appear in Bacteria and/or Archaea (Nasir, 2014)) and therefore it is precise, practical, and powerful to define domains by statistical models of their underlying sequences. For example, the Pfam database houses multiple-sequence alignments and hidden Markov models that represent protein regions identified from all proteome accessions in UniProt (Finn et al., 2016; Sonnhammer et al., 1998). A queried protein is matched to its domains by scores expressing how well each HMM model fits part of its sequence. In yeast, approximately 42% of genes have at least one domain, the most common being the protein kinase domain, which appears in 114 genes. Other examples of domains are the ATP binding domain of the ABC transporter, which moves substrates across membranes, the SH3 domain, which is frequently found in the proteins of signaling pathways, and the DEAD/DEAH box helicase domain, which unwinds RNA strands for various processes.

A complementary component of protein structure is intrinsic disorder, which describes regions of proteins that have no native structure, instead forming what is termed a random coil. Disorder can be predicted computationally with a classifier trained on proteins with known structure. DISOPRED, for example, predicts a protein's disordered regions from its amino acid sequence and a PSIBLAST position-specific scoring matrix, which contains information about variations of residues that are found at each position in evolutionarily related sequences (Jones and Ward, 2003). The authors that designed DISOPRED found that disordered protein regions occur as long segments in one third of eukaryotic proteins (Ward et al., 2004b). Despite the lack of precise structure-function associations, disordered regions can also be indicative of functional properties. Protein disorder is broadly associated with high rates of evolution; however, the disordered state of regions in *S. cerevisiae* proteins can also be highly conserved in other yeasts species, indicating functional importance. When the disordered state is conserved but the amino acid sequence varies between species, the disorder is termed flexible, and proteins with this type of disordered region have specific functional

associations, such as involvement in signaling, low expression, and having a single interface that binds different proteins at different times, among others (Bellay et al., 2011b).

Sequence comparison of a gene and its orthologs in close relatives can indicate effects of selection pressure. Evolutionary rate of a gene can be estimated from the ratio between nonsynonymous and synonymous mutations, denoted dN/dS, where the mutation counts are determined through comparing the query gene with an ortholog in a relatively closely related species (Goldman and Yang, 1994; Kimura, 1977). Alternatively, mutations can be counted. The Saccharomyces Genome Resequencing Project has sequenced 19 *S. cerevisiae* strains that have been isolated from a variety of environments (Liti et al., 2009). From these data, single nucleotide polymorphisms, SNPs, can be searched for in every gene. The incorporation of consequential mutations in a gene from any of the strains could mean an absence of stabilizing selection, and the confidence of this conclusion would increase with the number of SNPs and SNP-containing strains. Prediction of whether a SNP causes a functional defect in a gene, a significant refinement over solely distinguishing non-synonymous and synonymous mutations, is done by the SIFT algorithm, among others, which uses multiple-sequence alignments of homologs to infer each amino acid position's tolerance to specific substitutions, under the assumption that highly conserved amino acids are likely to have deleterious effects if changed by a SNP (Ng and Henikoff, 2003).

Analysis of the presence and absence of gene's homologs yields a number of interesting gene characterizations. Paralogs are genes within a single genome (e.g. that of *S. cerevisiae*) that evolved from one ancestral gene and have retained enough sequence similarity to be identified. These genes arise through duplication events in which genome segments are copied due to mistakes during DNA replication and repair or chromosomal segregation during cell division. The whole-genome duplication event is a particularly striking event in the history of *Sensu stricto* yeasts and *S. cerevisiae* duplicated genes can be described as whole-genome duplication (WGD), if they resulted from this event, or small-scale duplication (SSD) duplicates otherwise. Some differences between these classes were mentioned in section 1.2.1. An important result of small-scale duplication events is gene families, which arise from multiple (SSD) duplication events affecting one gene or set of paralogs and sometimes show fast rates of evolution.

The quantitative gene characteristic that reflects the duplication history is copy number, which is one plus the number of paralogs a gene has.

The interpretation of gene descriptors derived from orthologs, genes of separate species that evolved from a single ancestor gene, is relative to gene function in other species. Therefore, although orthologs are identified by sequence, we discuss these in the next section.

Phylogenetic characteristics

Describing a gene in terms of its orthologs gives a broad view of its evolutionary history by offering insight to the importance of it conserved function. The InParanoid project has performed ortholog identification for an extensive collection of 99 eukaryotic proteomes (Ostlund et al., 2010). The InParanoid algorithm (Remm et al., 2001) uses BLAST to compare all protein pairs within and between two species, providing sequence-based distance scores to a clustering algorithm. The clustering applies a number of rules to define clusters that distinguish between duplication events that occurred before or after species divergence. If two species are closely related, then synteny, the conservation of gene order on a chromosome, can be used to bolster support for sequence-based orthology predictions, as is done in the SYNERGY algorithm (Wapinski et al., 2007a). SYNERGY, however, is a substantially different approach from InParanoid: by making use of a pre-defined species phylogeny, it constructs a gene history tree that shows where, in relation to extant and ancestral species, duplication and deletion events occur. The advantage of this approach is that orthogroups (sets of orthologs) can be traced throughout all analyzed species. The designers of SYNERGY applied the algorithm to 23 Ascomycete species and noted that orthogroups varied strikingly in the tendency of their member genes to be duplicated or deleted, which they defined quantitatively as a measurement called volatility (Wapinski et al., 2007b). Characteristics of volatile orthogroups included signaling and stress response functions, variable expression in mutant genotypes and in different species, and expression regulation by the SAGA complex and the TATA box. In contrast, non-volatile orthogroups tended to be involved in essential growth processes, localize inside organelles, make up core components of protein complexes, and rarely show changes in expression. We use volatility as a gene descriptor, assigning each gene the volatility of its containing orthogroup.

Regardless of the method used to discover orthologs, two additional gene characteristics of evolutionary conservation can be defined from them. Given a phylogeny, the age of an *S. cerevisiae* gene can be expressed as the most distantly related species to have an orthologous gene. We defined a measure of gene age based on a tree that is a combination of the Wapinski et al. (2007b) and Ostlund et al. (2010) trees. We sequentially labeled, from most to least recent, all the last common ancestors between yeast and the extant species studied by the InParanoid and SYNERGY groups. Then gene age of an *S. cerevisiae* gene is defined as the highest (least recent) last common ancestor that the gene shares with any ortholog. Gene age captures only a slice of a gene's evolutionary history—essentially the depth of a gene's origin in an evolutionary tree. A second interesting measurement is the breadth of the presence of a gene's orthologs across many species. We term this "conservation" and calculate it as the number of species that contain an ortholog of a given gene.

Genome-wide experimental data-derived characteristics

Some of the most valuable characterizations of genes are those that inform their cellular functions through observation. One of the earliest, and subsequently most common, high-throughput genomic technologies is the microarray, which simultaneously measures the presence of specific RNA and DNA sequences in samples extracted from cells (reviewed in Zhang, 2006). A microarray is a solid plate onto which tens of thousands of oligonucleotides with known sequences, called probes, that are representative of genome content are tethered in grid pattern. To measure gene expression, complementary DNA (cDNA) may be produced from experimentally isolated mRNA, fluorescently labeled, and hybridized to complementary probes on the microarray. Scanning the microarray for fluorescence strength yields a quantitative measure of how much labeled cDNA hybridized to each probe. This measurement of mRNA abundance for tens of thousands of sequences is a snapshot of genes that have recently been transcribed. Because gene expression is regulated according to necessity, gene functions can be studied by measuring expression genome-wide in many different conditions or in a time series. For example, genes with expression that is upregulated during a certain phase of the cell cycle (Spellman et al., 1998) or in response to a DNA damage condition (Gasch et al., 2001) can be considered to have functions related to these conditions. Expression variance across many conditions can summarize gene

behavior: high variance may indicate that a gene functions in multiple cellular responses to conditions (Gasch et al., 2000) and suggests the gene is not part of a core cellular function requiring constant expression levels. A powerful method of analyzing expression data is calculating pairwise correlations between all gene expression profiles across experimental samples. Genes whose expression level patterns are very similar tend to be functionally related, and these relationships can be determined for all pairs of genes without making assumptions about how a cell may be responding to particular conditions (Stuart et al., 2003). By interpreting high correlations as connections in a network (Huttenhower et al., 2006), we can count the number of co-regulated partners a gene has, a gene characteristic called co-expression degree.

Other experimental datasets measure phenotypes associated with genes. At a molecular level, protein abundance (Newman et al., 2006) can indicate the functional importance of a gene, which is similar to expression level but takes into account post-transcriptional regulation. At an organism level, the phenotype of cell growth is a proxy for fitness, which can be measured for different genotypes. In yeast, there are collections of strains harboring individual deletions or mutant alleles for nearly all genes in the genome. Thus, single-mutant fitness measurements of the strains can be associated with individual genes (Costanzo et al., 2016; Giaever et al., 2002). Low single-mutant fitness for a given gene is an indication that the gene is important for cell growth or health, while high single-mutant fitness indicates its function has little importance in standard conditions or that the cell is able to adequately compensate for its loss. Fitness is not limited to standard conditions: it can also be measured while exposing cells to different chemicals. Lastly, Ohya et al. (2005) designed a high-throughput system to measure morphological phenotypes of internal cellular structures in yeast strains. The authors fluorescently stained cell walls, actin cytoskeletons, and nuclear DNA of the yeast deletion strains and used automatic image processing to measure over 250 parameters describing the shape, size, and positions of these structures. Levy and Siegal (2008) used this dataset to assess the phenotypic variation that occurred within the single-mutant populations and calculated a measure of phenotypic capacitance, which expressed the extent to which loss of a given gene led to loss of phenotypic robustness. They found that genes with high phenotypic capacitance (genes that promote uniform phenotypes with a population) are likely to have many genetic

interactions, protein interactions, and be involved in cellular processes that can have broad effects, such as transcriptional regulation and maintenance of DNA stability.

Curated data-based characteristics

As evidenced by the large number of data sets that contribute to sequence-based, experimental, and phylogenetic gene characteristics, many yeast genes have been well characterized. These genome-wide studies have complemented the many small-scale investigations of yeast biology that have been conducted for many decades. Given the large number and variety of all experimental characterizations of genes, manual work to assimilate these data into accessible and standard gene descriptions is very valuable. Specifically, the curators at the Saccharomyces Genome Database (Cherry et al., 2012) track all experimental evidence associated with individual genes, such as mutant and conditional phenotypes. They also annotate the genes with terms from the Gene Ontology (GO) (Ashburner et al., 2000). The GO represents a systematic hierarchical organization of much of the current knowledge about genes in all species by defining terms that describe molecular functions, biological processes, and cellular compartments. Approximately 87% of yeast ORFs that show evidence of a protein product have been annotated so far. A frequent use of GO annotations is to interpret genes by testing for statistically over represented terms in a group of genes. As for gene characteristics, the number of assigned GO terms or phenotypes can indicate gene multifunctionality, though GO is known to be somewhat biased and, of course, is limited to biology that has been investigated (Gaudet et al., 2017).

1.4 Dissertation focus

The yeast *Saccharomyces cerevisiae* is one of most studied and most easily genetically manipulated model species. Historically, it has been integral to the discovery of many fundamental eukaryotic cellular processes and both small-scale and high-throughput experiments continue to increase our knowledge of biology. However, progress in understanding human biology, including the identification of disease mechanisms and treatments, requires species-specific knowledge. Biological networks show only moderate conservation between species in terms of both nodes (genes or proteins) and interactions (physical or functional) among conserved nodes. While core modules tend to show conservation, other aspects of cellular networks, such as

between-module connections, duplicates, and expression regulation, appear to evolve quickly—and are thought to facilitate the development of species-specific traits. Consequently, there is need for methods to identify basic principles of biological networks that are universal among eukaryotes and to transfer knowledge from a model organism to the context of another species. In Chapter 2, we demonstrate that it is possible to build a model encapsulating patterns between gene characteristics and structure in genetic interaction networks. We show that the model works well to predict genetic interaction degree in two distantly-related species, *S. cerevisiae* and *S. pombe*. Importantly, the model we trained was not dependent on homology, indicating that this type of analysis can guide research efforts related to species-specific biology.

The recently completed yeast genetic interaction network, in combination with curated annotations of genes, allows the systematic investigation of gene functional behavior. In Chapter 3, we discuss a pipeline that creates functional profiles for genes by identifying all functional modules each gene participates in and summarizing them across 20 high-level biological processes. Gold-standard gene annotation schemes typically only reflect a single primary function for each gene, or, in the case of GO terms, are likely significantly affected by investigation bias. Thus, our systematic module-derived functional profiles are truer representations of gene functions. Chapter 4 presents a particularly exciting use of these functional profiles: genome-wide measurement of pleiotropy—the widespread phenomenon in which one gene impinges on multiple functions. After identifying pleiotropic genes, we find the properties that distinguish them from genes with very low pleiotropy and discuss why these properties may be associated with pleiotropy. Although multiple network analyses have shown that some high-degree genes (or proteins) act within modules and others interact with or in many modules, there are no systematic descriptions of this topological behavior that also consider the biological functions of the modules. Additionally, although there are many definitions of pleiotropy (reviewed in section 4.2.1), no one has previously used genetic interactions to detect gene pleiotropy.

A common theme in these two main lines of research is the combination of a diverse set of gene characteristics and GI network topological behavior. The motivation for this is to find fundamental principles of genomics that are universally informative and applicable. As discussed in this introduction, there are many ways in which gene characteristics are associated with function and evolution: gene sequences can reflect

the strength of evolutionary pressure or the existence of functional units, a duplicate gene can gain or lose functions, membership in a complex restricts ability to interact with new proteins, orthologs in many species indicates an important function, and many others. These relationships are most likely well conserved. In model organism research, a predominant motivation is the expectation that results will be relevant to many species due to homology, the evolutionary conservation of sequences and structures. There is no doubt that this expectation has been realized many times or that conservation can cover exceedingly long time periods. However, it has recently become clear that biological networks evolve considerably, incorporating new nodes and altering connections between nodes. As we demonstrate in Chapter 2, the relationship between gene characteristics and GI degree is conserved, and thus there may be many such ways to understand network-based behaviors of genes through their conserved relationships with gene characteristics. We anticipate that our results describing pleiotropic genes also fall into this category.

Some of the work presented in this document benefitted from contributions from collaborators of the author. These contributions are specified in introductory sections of individual chapters.

Chapter 2: Conserved rules govern genetic interaction degree across species

2.1 Chapter overview

Although many genetic interaction screens performed for *S. cerevisiae* (Costanzo et al., 2010) have yielded a genome-wide genetic interaction network, comprehensive genetic interaction networks have not been determined for other species. We therefore sought to model aspects of GI networks in order to enable the transfer of knowledge between species. This chapter presents the successful application of a machine-learning strategy to model GI degree. We apply the model to make predictions for *S. pombe* genes and conduct an analysis of rewiring between the species.

The text of this chapter has previously been published as an article in the journal Genome Biology (Koch et al., 2012). The author of this dissertation had a leading role in planning this work and writing the associated publication; all analysis was done by this author, except the aspects specifically noted in this paragraph, with contributions from collaborators. In addition to the author, Jeremy Bellay, Chad L. Myers, Michael Costanzo, Charles Boone, and Brenda J. Andrews conceived and planned the analysis. Jeremy Bellay made contributions to gathering gene properties and designing predictive models. Gordon Chua, Kate Chatfield-Reed, and Michael Costanzo performed the *S. pombe* GI screens and fitness measurements. Raamesh Deshpande constructed the co-expression network for *S. pombe*. Michael Costanzo, Gennaro D'Urso, Charles Boone, and Chad L. Myers contributed to writing the manuscript.

2.2 Background

Most genes are nonessential for eukaryotic life under standard laboratory conditions, which may reflect that organisms are highly buffered from genetic and environmental perturbations (Hartman et al., 2001). However, rare combinations of singly benign genetic variation can lead to synergistic effects, such as synthetic lethality, where mutations in two genes, neither of which is lethal independently, combine to generate an inviable double-mutant phenotype (Dixon et al., 2009). Because the natural variations that distinguish two individuals, such as single nucleotide polymorphisms, occur relatively frequently (Gibbs et al., 2003), and complex genetic interactions may underlie most individual phenotypes (Hartman et al., 2001), understanding the general

principles and rules that govern genetic networks may be critical for solving the genotype-to-phenotype problem and implementing personal medicine (Dowell et al., 2010).

Recently, we tested ~5.4 million *Saccharomyces cerevisiae* gene pairs for genetic interactions, mapping an extensive network of more than 100,000 interactions by synthetic genetic array (SGA) analysis (Costanzo et al., 2010). The study mapped both negative genetic interactions, the situation in which a double mutant exhibits a more extreme phenotype than the expected combined effects of the single mutants, as well as positive genetic interactions, the situation in which a double mutant exhibits a less pronounced phenotype than expected (Baryshnikova et al., 2010b). This study revealed the distribution of genetic interactions with respect to gene function and highlighted a central role for chromatin-related, transcription, and secretory functions as well as several fundamental physiological and evolutionary gene properties that are significantly correlated with genetic interaction degree in the *S. cerevisiae* genetic network (Costanzo et al., 2010). For example, we showed that the genetic interaction degree of a gene is highly correlated with single mutant fitness, such that genes with a substantial fitness defect show a large number of genetic interactions.

While genetic interactions have been the most extensively studied in the yeast *S. cerevisiae*, there is intense interest in developing and applying large-scale screening technologies in other species. For example, large studies have already been completed in *Escherichia coli* (Butland et al., 2008; Typas et al., 2008), *Schizosaccharomyces pombe* (Dixon et al., 2008; Roguev et al., 2008), *Caenorhabditis elegans* (Byrne et al., 2007; Lehner et al., 2006), *Drosophila melanogaster* (Agaisse et al., 2005; Boutros et al., 2004), and human cell lines (Barbie et al., 2009; Luo et al., 2009; Scholl et al., 2009). Although definitive comparative analysis of these networks across species would be premature given the sparsity of known interactions in the species other than *S. cerevisiae*, there have been preliminary comparative studies. In particular, the yeast *S. pombe* provides an attractive setting for this analysis due to the availability of a genome-wide deletion mutant collection (Kim et al., 2010a) and scalable technology for automated genetic analysis (Dixon et al., 2009). Furthermore, *S. cerevisiae* and *S. pombe* are estimated to have diverged approximately 500 million years ago and display markedly different physiological properties but share 75% of their gene content (Rhind et al., 2011; Sipiczki, 2000). The two comparative studies to date estimated approximately

30% conservation of individual negative genetic interactions, but also found substantial differences between the two species (Dixon et al., 2008; Roguev et al., 2008). These studies demonstrate the power and necessity of comparative analysis of genetic interaction networks, but have conducted only limited sampling of genetic interactions in *S. pombe*. The properties of these networks that are conserved across species and the rules governing their evolution remain largely open questions, making further characterization of the evolution of genetic interaction networks important.

2.3 Modeling interaction degree in the *S. cerevisiae* genetic interaction network

Highly connected genes in the *S. cerevisiae* genetic interaction network are often associated with slow-growing single mutants, protein products with disordered structure, gene pleiotropy as indicated by multiple Gene Ontology (GO) annotations, high connectivity in the physical interaction network, slower rates of evolution, and low expression variation (Figure 2.1A; Appendix 2, A2.2) (Costanzo et al., 2010), as well as a number of other sequence- and experimental-based gene features (Table 2.1). We reasoned that these correlations could serve as the basis for predictive modeling of interaction degree, enabling the prediction of interaction degrees for genes that have not yet been screened.

To this end, we applied a regression tree approach to model combinations of 16 gene characteristics (Appendix 2, A2.2) that are predictive of negative genetic interaction degree (Figure 2.1B). Regression trees are built by repeatedly splitting sets of training genes, according to the values of gene characteristics, until genes are sorted into small sets that each contain genes with similar genetic interaction degrees. The hierarchy of gene sets produced is visualized as a binary tree and the final sets of genes are each associated with linear regression models that assign predictions to query genes (Figure 2.1B). Bootstrapped subsets of the training data were used to build an ensemble of regression trees; this use of multiple models, bootstrap aggregation, is a typical method for building a robust predictive model (Breiman, 1996) (section 2.8.1).

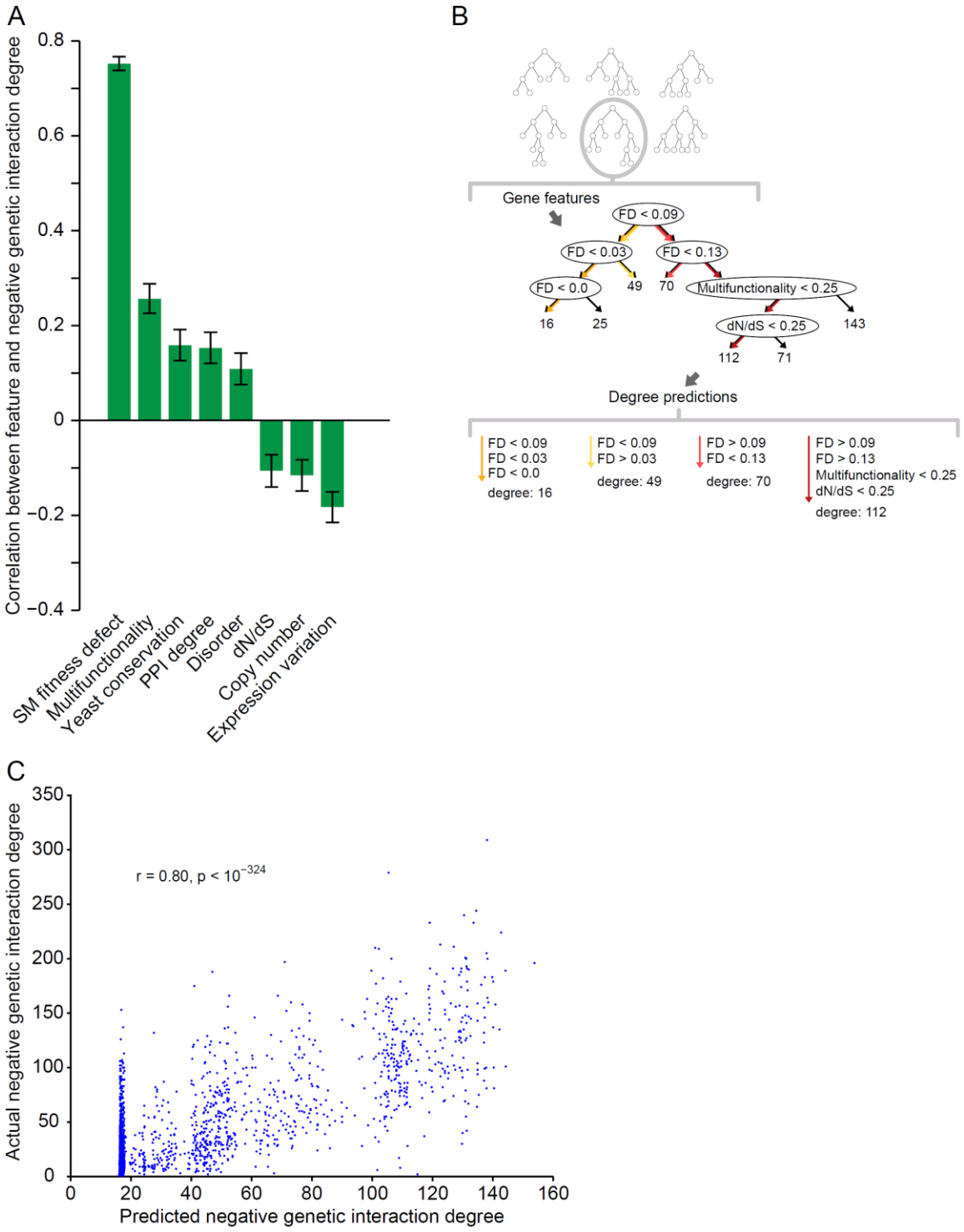


Figure 2.1. Physiological and evolutionary gene characteristics are predictive of genetic interaction degree. (A) Gene features are significantly correlated with negative genetic interaction degree. We measured Pearson correlation coefficient between gene

feature values and negative genetic interaction degree for 3456 nonessential *S. cerevisiae* genes. Error bars show 95% confidence intervals. A complete set of features and their correlations is given in Table 2.1; see section A2.2 for descriptions of gene characteristics. SM, single mutant. **(B)** Overview of the regression tree model for genetic interaction degree. An ensemble of 100 decision trees was built from bootstrap samples of genes. Combinations of values of characteristics are represented as paths from the root to the leaves of a tree. Internal nodes each split data (sets of genes) according to values for a single characteristic; leaf nodes are associated with predicted genetic interaction degrees. FD, single mutant fitness defect. **(C)** Scatter plot of negative genetic interaction degree and degrees predicted by the bagged decision tree model on held-out genes shows the significant relationship between predicted and actual degrees (Pearson's $r = 0.80$, $p < 10^{-324}$).

Table 2.1. Pearson correlations between features and negative genetic interaction degree in *S. pombe* (*pom*) and *S. cerevisiae* (*cer*) are observed to be significant in many cases.

		Pearson's r	p-value	95% CI
SM fitness defect	*pom*	0.48	9.90E-31	[0.41, 0.54]
	cer	0.75	0.00E+00	[0.77, 0.74]
Multifunctionality	*pom*	0.3	1.01E-12	[0.22, 0.37]
	cer	0.26	4.52E-53	[0.29, 0.23]
Conservation	*pom*	0.07	1.06E-01	[-0.01, 0.15]
	cer	0.16	7.11E-21	[0.19, 0.13]
Broad conservation	*pom*	0	9.30E-01	[-0.09, 0.08]
	cer	0.16	9.90E-19	[0.19, 0.12]
PPI degree	*pom*	0.2	5.84E-03	[0.06, 0.33]
	cer	0.15	1.49E-19	[0.19, 0.12]
Expression level	*pom*	-0.05	2.42E-01	[-0.13, 0.03]
	cer	0.11	7.40E-10	[0.14, 0.08]
Disorder	*pom*	0.13	3.05E-03	[0.04, 0.21]
	cer	0.11	1.91E-10	[0.14, 0.08]
CAI	*pom*	-0.02	6.49E-01	[-0.1, 0.06]
	cer	0.09	1.91E-07	[0.12, 0.06]
Protein length	*pom*	0	9.18E-01	[-0.08, 0.09]
	cer	0.05	3.57E-03	[0.08, 0.02]
Co-expression degree	*pom*	0	9.38E-01	[-0.08, 0.09]
	cer	0.05	3.75E-03	[0.08, 0.02]
Num. of domains	*pom*	-0.01	7.37E-01	[-0.1, 0.07]
	cer	0.01	4.54E-01	[0.05, -0.02]
Num. of unique domains	*pom*	-0.02	6.86E-01	[-0.1, 0.07]
	cer	0.01	6.98E-01	[0.04, -0.03]
Nc	*pom*	-0.01	7.56E-01	[-0.1, 0.07]
	cer	-0.08	2.95E-06	[-0.05, -0.11]
dN/dS	*pom*	-0.05	2.48E-01	[-0.14, 0.04]
	cer	-0.11	1.58E-09	[-0.07, -0.14]
Copy number	*pom*	-0.08	5.26E-02	[-0.17, 0]
	cer	-0.12	1.01E-11	[-0.08, -0.15]
Expression variation	*pom*	-0.15	4.11E-04	[-0.23, -0.07]
	cer	-0.18	3.87E-27	[-0.15, -0.21]

To validate our approach, we used our model to predict negative genetic interaction degree for all genes in the *S. cerevisiae* genetic interaction network (Figure 2.1C; section 2.8.1). A high correlation ($r = 0.80$, $p < 10^{-324}$) was observed between predicted and actual genetic interaction degrees of genes not used in training the models, indicating that our model accurately reflects topological features of the *S. cerevisiae* genetic interaction network (Figure 2.1C). A strength of this type of model, in addition to providing degree predictions for previously unseen genes, is that the learned tree structures highlight rules consisting of combinations of gene characteristics that explain variation in degree (Figure 2.1B).

2.4 Predicting genetic interaction degree in a distantly related species

If the rules governing genetic network topology are conserved, then a model based on *S. cerevisiae* gene features should be predictive of genetic interaction degree in other organisms. To test this, we examined the same gene features of *S. pombe* genes that we found to be predictive of *S. cerevisiae* interaction degree, including a quantitative measurement of single mutant fitness defect across the genome (section A2.2; section 2.8.2). Surprisingly, comparative analysis of the various characteristics between pairs of orthologs revealed that a number of non-sequence-based features are only modestly conserved between the two yeast species (Berglund et al., 2008) (Figure 2.2A; section A2.2). For example, we found a significant but relatively weak correlation in single mutant fitness defect (Pearson's $r = 0.20$, $p < 10^{-8}$) between 1,100 one-to-one orthologous gene pairs for which we could derive fitness measurements in both yeasts. The lack of strong conservation of deletion mutant fitness is somewhat surprising given that approximately 80% of *S. pombe* orthologs of *S. cerevisiae* essential genes have conserved essentiality (Kim et al., 2010b). Thus, while *S. cerevisiae* and *S. pombe* share a common set of genes that are indispensable for viability, our findings suggest that the severity of fitness defects imposed by the deletion of orthologous nonessential genes for growth under standard laboratory conditions is not well conserved. Other gene properties, including protein-protein interaction degree, dN/dS, and multifunctionality, exhibit a similar lack of conservation (Figure 2.2A).

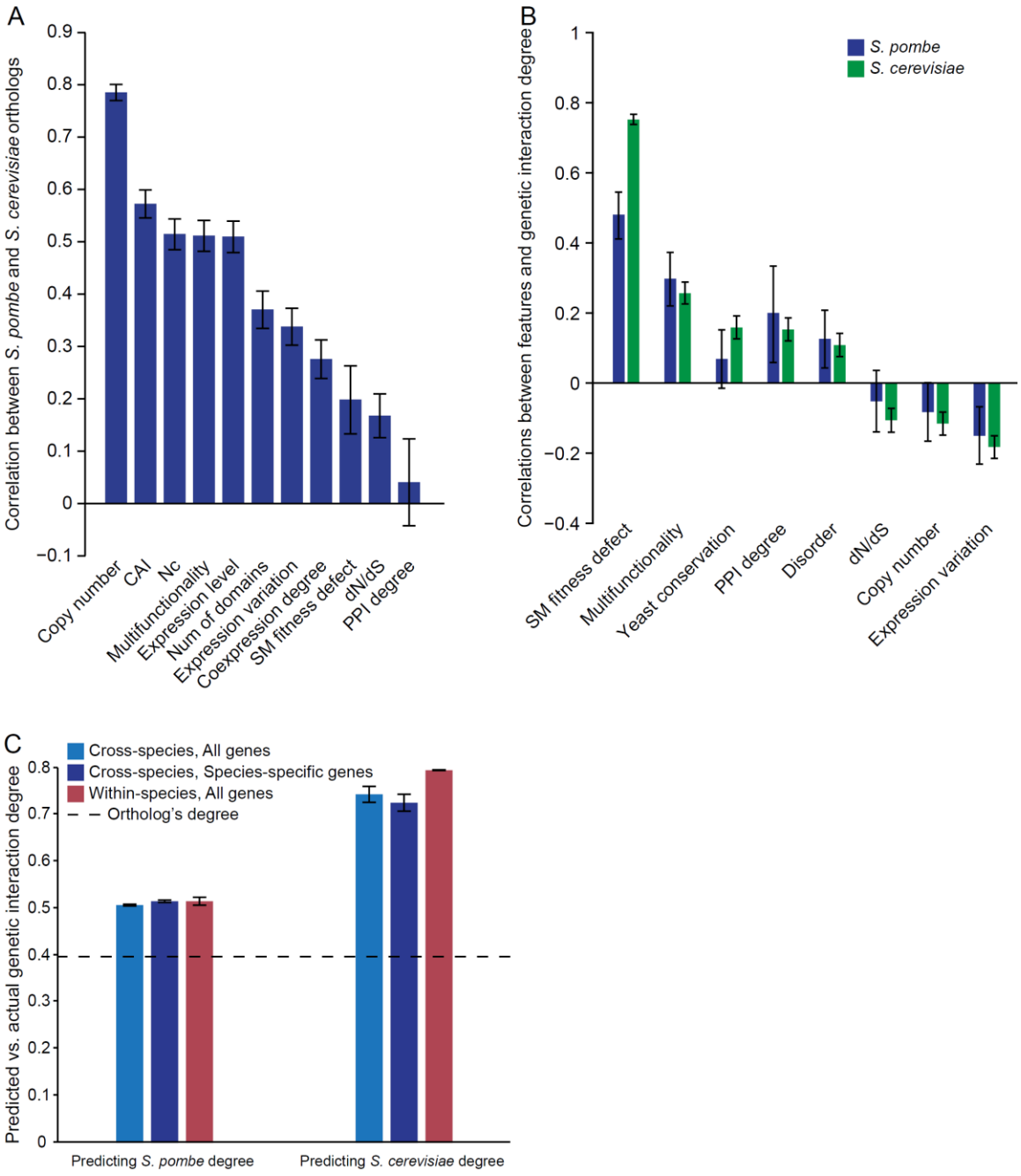


Figure 2.2. Cross-species analysis of the predictive model for genetic interactions. **(A)** Pearson correlations between one-to-one *S. cerevisiae* and *S. pombe* orthologs for their values of gene characteristics. Note that a number of characteristics are sequence-based and are thus not independent of the sequence-based ortholog identification; features that appear to have trivial correlations are not included here. Error bars show 95% confidence intervals. **(B)** Pearson correlations between features and degree in *S.*

pombe are observed to be significant in many cases and similar to those in *S. cerevisiae*. A complete set of features and their correlations is given in Table 2.1; see section A2.2 for descriptions of characteristics. Error bars show 95% confidence intervals. **(C)** Predictive abilities of bagged regression tree models were evaluated by measuring Pearson correlations between predicted and actual degrees. The left set of bars shows the performance of predictions made for ~550 *S. pombe* genes and the right set of bars shows the performance of predictions made for all nonessential deletion mutants in *S. cerevisiae*. For each scenario, models were trained both on data from the same species (red bar) as well as data from the other species (blue bars). The light blue bars correspond to predicting degrees of all genes in the test species, while the dark blue bars correspond to predicting degrees for the subset of genes lacking orthologs in the training species. Error bars show standard deviations of bootstrapped predictions. For a baseline, the dashed line shows the correlation between observed degrees of one-to-one orthologous genes (a simple prediction method that can be applied to only orthologs).

Despite the low conservation of single mutant fitness and the varying correlations between individual gene properties for orthologs, we found that relationships between *S. pombe* gene characteristics and genetic interaction degree were strikingly similar to those observed in *S. cerevisiae* (Figure 2.2B, Table 2.1). Consistent with *S. cerevisiae* trends (Figure 2.1A, Table 2.1), fitness defect was the strongest predictor of *S. pombe* genetic interaction degree. That is, *S. pombe* strains with severe fitness defects often exhibit high numbers of genetic interactions. The observed trends suggested that in addition to correlations with individual gene features, higher-level combinations of features that are predictive of connectivity in the *S. cerevisiae* genetic interaction network (Costanzo et al., 2010) (Figure 2.1A) may also be informative of *S. pombe* genetic interaction degree.

To test this hypothesis, we built a predictive model relating the combination of available gene characteristics to genetic interaction degree in *S. cerevisiae* and then applied the resulting model to predict genetic interaction degree in *S. pombe* (section 2.8.1). Interestingly, we observed significant correlation ($r = 0.51$, $p < 10^{-36}$) between interaction degree predicted by our model and the number of interactions previously

determined (Roguev et al., 2008) for 548 *S. pombe* genes (Figure 2.2C, left side, light blue bar).

Our ability to predict interaction degree from a small set of gene-specific properties is evidence that rules governing genetic interaction network topology are conserved across a large evolutionary distance (Figure 2.2C). Importantly, there is no significant decrease in correlation between predicted and actual interaction degree when predictions were restricted to genes unique to *S. pombe* (Figure 2.2C, left side, dark blue bar), indicating that the model performs equally well when applied to genes lacking orthologs in the species used to learn relationships in the model.

As a baseline comparison for our cross-species predictive model, we built a model from *S. pombe* gene characteristics and genetic interaction degrees instead of from *S. cerevisiae* data. Within-species predictions for *S. pombe* interaction degrees are not significantly more accurate than predictions made by the cross-species model (Figure 2.2C, left side, compare red and light blue bars). We also note that although a simplistic predictor that maps the degree of a *S. cerevisiae* gene directly to its *S. pombe* ortholog provides reasonable performance (Pearson correlation approximately 0.4), this strategy is out-performed by our cross-species model and is limited to conserved genes. Strikingly, the models trained on *S. pombe* interactions and features were also able to predict interaction degree in the *S. cerevisiae* network with high accuracy (Figure 2.2C, right side, compare red and light blue bars). In general, interaction degree predictions for *S. pombe* genes were weaker than *S. cerevisiae* interaction degree predictions, which may reflect the limited functional diversity of available *S. pombe* genetic interaction studies (Dixon et al., 2008; Roguev et al., 2008). Nonetheless, the ability to predict interaction degree using characteristics measured in either yeast species is evidence that relationships between genetic interactions and fundamental physiological and evolutionary properties are generally conserved.

The strong correlation between single mutant fitness defect and negative genetic interaction degree has the unsurprising consequence that the models are considerably influenced by this feature. To explore the reliance of our model on fitness defect, we constructed two types of bootstrapped regression tree models that were trained on all characteristics except fitness defect. The first of these additional models is trained to predict negative genetic interaction degrees and is able to successfully make both within- and cross-species predictions (Figure A2.1, section A2.1). The second model

was trained to predict the residual negative genetic interaction degree that remained after subtracting degree predictions made from a regression tree model that was trained on the single feature single mutant fitness defect. The prediction of these residuals by the remaining features was also significant (Figure A2.2, section A2.1). We therefore consider the inclusion of many other features to be a worthwhile part of our model, since they capture aspects of genetic interaction degree that fitness defect alone does not.

2.5 Validating predictions with *S. pombe* whole-genome GI screens

As an independent validation of our model, we conducted genome-wide *S. pombe* genetic interaction screens. Eight query gene-deletion mutants spanning diverse cellular functions were crossed into an array of 2,907 nonessential *S. pombe* deletion mutants (Dixon et al., 2009; Kim et al., 2010a), making approximately 23,000 double mutant strains (Figure 2.3A; section 2.8.2).

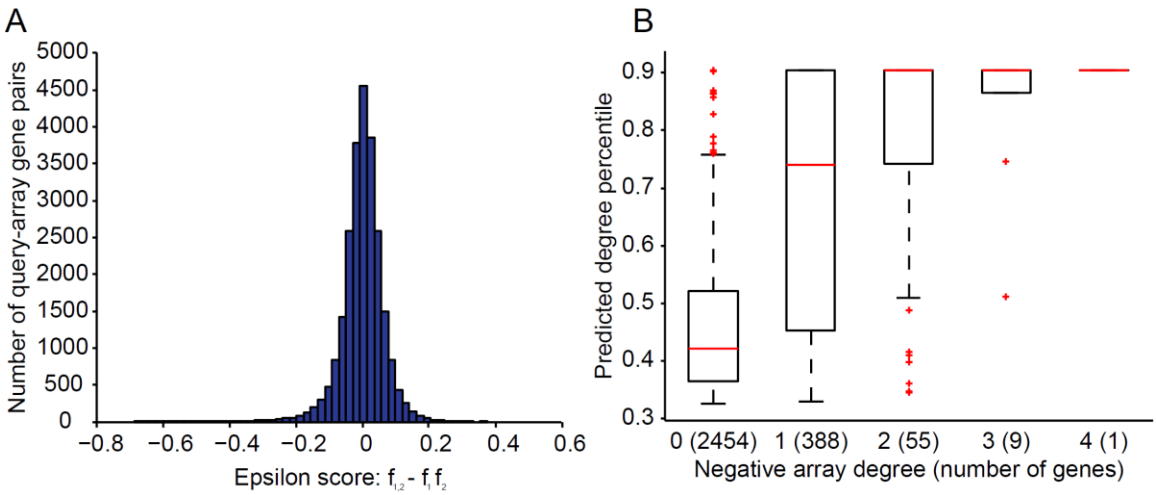


Figure 2.3. Observed genetic interactions between *S. pombe* genes support degree predictions. (A) Model predictions were validated on a second, whole-genome set of interaction screens in *S. pombe* that are independent of the training data. Eight query deletion mutants were crossed with the entire *S. pombe* nonessential deletion collection. In total, genetic interaction (epsilon) scores were measured for approximately 23,000 gene pairs. Epsilon scores are tightly centered at 0, thus interactions called for scores of +/-0.08 or more extreme are rare. **(B)** The collection of nonessential *S. pombe* genes (n = 2907) were grouped by the number of interactions each has with the eight

query genes for which full-genome screens were performed. Numbers in parentheses give the number of genes for which this degree was observed. For each degree, the box plot shows the distribution of predicted degrees, which are expressed as percentiles. There is a strong positive correlation (Pearson's $r = 0.40$, $p < 10^{-111}$) between predicted and actual degree.

Consistent with our results for a published dataset (Roguev et al., 2008) (Figure 2.2C), we observed a significant correlation ($r = 0.40$, $p < 10^{-111}$) between the predicted number of interactions and the total number of experimentally derived interactions observed for a given array mutant in this genome-wide deletion set. Grouping genes with the same observed degree, we found that the distributions of our predictions were reflective of actual degrees (Figure 2.3B). For example, the median degree percentile predicted for genes with a degree of one was approximately 0.72, while the median prediction for genes with zero interactions was approximately 0.42. Importantly, the significance of the correlation was robust to the choice of interaction cutoff and persisted for a higher-confidence, sparser network (section 2.8.2).

2.6 Identifying network rewiring suggested by cross-species predictions

Although many individual genes are conserved, yeast genetic interaction networks may have undergone substantial rewiring, as only approximately 30% of the interactions are conserved (Dixon et al., 2008). Similarly, a low conservation of genetic interactions has also been observed between *S. cerevisiae* and *C. elegans* (Tischler et al., 2008). To examine the extent of network rewiring, we first inferred interaction degree for the entire *S. pombe* genome using our cross-species model. Because the predictions did not depend on sequence orthologs (Figure 2.2A, C), they can be used to compare the topologies of the *S. cerevisiae* and *S. pombe* networks even though only a small fraction of the *S. pombe* network has been screened.

We found several instances where the predicted interaction degree for a given *S. pombe* gene was quite different from the observed degree of its *S. cerevisiae* ortholog, suggesting that the gene acquired or lost interactions differentially as the species diverged. To identify larger functional modules that were targets of this rewiring, we grouped functionally related genes according to a catalog of 65 annotated protein

complexes (Baryshnikova et al., 2010b) and 545 GO biological process annotations (Ashburner et al., 2000) (section 2.8.3), and compared the median interaction degree determined for orthologous protein complexes and functional groups (Figure 2.4A; Figure A2.3, section A2.1). Many groups of functionally related genes and several complexes were statistically indistinguishable in terms of network connectivity, indicating that these modules act either as network hubs in both species or non-hubs in both species.

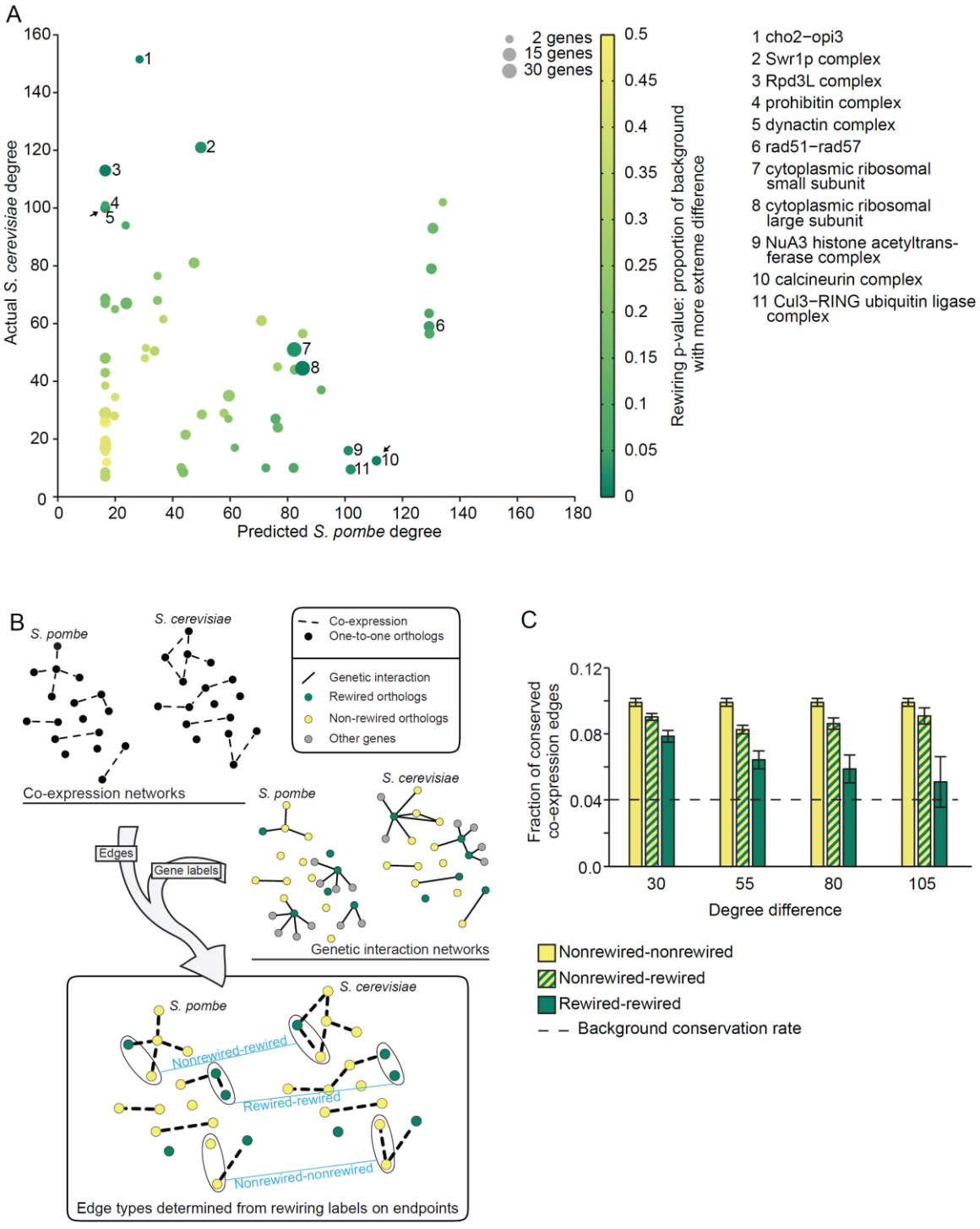


Figure 2.4. Global analysis of rewiring based on whole-genome predictions in *S. pombe*. **(A)** Points in the scatter plot each represent groups of between two and 22 genes whose protein products are in the same protein complex (section 2.8.3). Darker

color represents complexes that are predicted to have significant rewiring. Generally, genes in complexes that fall on the diagonal are predicted to have conserved degrees, while those that fall off-diagonal show evidence for large degree differences between the two species. Significantly rewired complexes (at a threshold of 0.05) are labeled by their names. **(B)** To validate our predicted rewired genes, we constructed separate networks of co-expression relationships among genes for each yeast species, then labeled genes according to our rewiring designation. Only one-to-one orthologs that are nonessential in both species were included in the networks. Edges in the co-expression network were classified by whether involved genes were both rewired, only one was rewired, or neither was rewired. We then calculated fractions of conserved co-expression relationships between species within each of these classes. **(C)** There is a clear relationship between these classes of edges and their conservation across the two yeast species. For rewiring at four levels of magnitude, we counted the number of conserved edges (among all edges in the union of the two networks). A conserved edge appears in the networks of both species and a non-conserved edge appears in exactly one. The magnitude of rewiring increases along the x-axis for the rewired class (differences of >30, >55, >80, >105 interactions), but the non-rewired class is defined as the set of ortholog pairs with less than a 30-edge difference in degree. Edges in the two rewired classes consistently showed significantly lower levels of conservation than edges in the non-rewired class ($p < 0.01$, Fisher's exact test). Error bars show the binomial proportion 95% confidence interval. The dashed line is the expected rate of conservation if edges are randomized in one of the co-expression networks. There are 12472 edges among 509 genes in the conserved-conserved network. Numbers of edges and genes at rewiring thresholds, in bold, are as follows, where the conserved-rewired case is given as the first pair and the rewired-rewired case is given second: **30:** (14532, 832), (4684, 323) **55:** (8730, 695), (1659, 186) **80:** (5358, 620), (644, 111) **105:** (2822, 565), (176, 56).

However, we also identified many examples of possible rewiring, in which a significant difference in network connectivity, observed in *S. cerevisiae* and inferred in *S. pombe*, was found for orthologous modules (Figure 2.4A; Figure A2.3; section 2.8.3). These predicted rewired groups represent complexes or biological processes that may have evolutionarily diverged in terms of their importance in the genetic interaction network, acting as hubs in one species but not in the other. In particular, we found that

11 of 65 (17%) protein complexes and 44 of 545 (8%) GO biological processes may have undergone significant rewiring (Figure 2.4A; Figure A2.3) at a level of significance expected to identify only 3 and 27 (5%) rewired modules, respectively. For example, components of the dynactin complex are hub genes in the *S. cerevisiae* genetic interaction network (complex average of 85th percentile; Figure 2.4A) whereas the orthologous genes were predicted to exhibit average connectivity in the *S. pombe* genetic interaction network (complex average of approximately 50th percentile; Figure 2.4A). Dynactin, a multi-subunit protein complex known for interacting with dynein and enabling long-range movement along microtubules (reviewed in Schroer, 2004), has been implicated in a *S. cerevisiae* cell cycle checkpoint pathway that arrests cell cycle progression in response to perturbations in cell wall synthesis (Suzuki et al., 2004). A similar checkpoint has not been reported in *S. pombe*, suggesting that the difference in the number of genetic interactions observed across species may reflect a dynactin-specific role in monitoring *S. cerevisiae* cell wall integrity.

In addition to *S. cerevisiae*-specific genetic interaction hubs, we also identified gene groups predicted to be hubs in the *S. pombe* network but not observed as such in the *S. cerevisiae* genetic network. One such case is the calcineurin-associated protein complex (Figure 2.4A). A difference in network connectivity might reflect a unique role for calcineurin in the regulation of bi-polar growth activation in *S. pombe* (Kume et al., 2011). Unlike an *S. cerevisiae* cell, which grows predominantly via an actin-dependent budding mechanism, an *S. pombe* cell grows in a highly polarized bi-polar manner from its two ends. Following cell division, cell growth is initiated from the old end first, and later, after completion of S phase, from the newer end that forms at the site of cell septation (referred to as new end take off, or NETO). Calcineurin has been shown to play an important role in the delay of NETO by directly dephosphorylating critical targets involved in microtubule dynamics at the site of cell growth. This mechanism is dependent on activation of Cds1 kinase, best known for its role in the intra-S phase DNA replication checkpoint (Boddy, 1998). A connection between the intra-S phase checkpoint and inhibition of bipolar growth activation is so far unique to *S. pombe* and distinct from the checkpoint controls operating in *S. cerevisiae*. Additionally, calcineurin is dispensable for growth in *S. cerevisiae* (Sugiura et al., 2001); in *S. pombe*, its deletion leads to defects in cell growth, cytokinesis, cell polarity, mating, and spindle pole body

positioning, which are widespread effects consistent with its hub-like activity (Yoshida et al., 1994).

While our method of identifying rewired modules reports several statistically significant differences, we note two caveats in interpreting these results. First, since degrees of genes within functional modules may be systematically poorly predicted, our procedure may incorrectly identify modules as significantly rewired in cases where our test statistic would also have indicated that the within-species difference between predicted and observed degree was significant. Therefore as a control, a version of this rewiring experiment that compares observed and predicted *S. cerevisiae* degrees will enable identification of cases that do not reflect true cross-species rewiring (Figure A2.4A, B). Second, due to variations in the experimental protocol for measuring genetic interactions, there are differences in the media on which fitness defects were measured in *S. cerevisiae* and *S. pombe*, which may also contribute to apparent rewiring (Baryshnikova et al., 2010a).

Functional properties of genes can be captured by many types of biological networks, so we turned to an independent dataset for confirmation of our rewiring predictions. To enable a comparative analysis of gene expression profiles across the two yeasts, we constructed a species-specific *S. pombe* co-expression network using a previously published approach (Huttenhower et al., 2006) and large collections of publicly available expression data (section 2.8.4), and obtained a previously published *S. cerevisiae* network (Myers et al., 2005). Each species's network contained 832 genes that are one-to-one orthologs between the two yeasts and connected genes are those pairs that have high co-expression values surpassing a threshold of the 95th percentile. At our selected density of 0.05, there are approximately 17,000 edges in each network. In general, we found evidence of conservation between the *S. cerevisiae* and *S. pombe* networks: co-expression edges between two genes occurred in both networks for 9.2% of the gene pairs that were co-expressed in at least one network. This is about twice the background conservation rate of approximately 4.3%, as determined through comparison to a randomized network produced by a degree-preserving procedure.

To explore the connection between genes predicted to be rewired in the genetic interaction networks and differences between the co-expression networks, rewiring predictions were overlaid on the co-expression networks. Specifically, all nonessential one-to-one orthologs were classified as either rewired or non-rewired based on our

prediction of genetic interaction degree (Figure 2.4B). Using this rewiring labeling, we measured the conservation rate of three types of co-expression edges: co-expression edges connecting two non-rewired genes, connecting two rewired genes, and connecting rewired to non-rewired genes.

We found that co-expression edges involving predicted rewired genes are consistently less-conserved than edges with exclusively non-rewired endpoints (Figure 2.4C), a trend that is robust over different co-expression thresholds used for network sparsification (Figure A2.5). For example, when genes whose degrees differ by 55 interactions or more are considered rewired, 6.9% of the co-expression relationships connecting rewired genes are conserved (107 of 1,659), in contrast to the significantly higher 10.1% of co-expression relationships that are conserved between non-rewired genes (1,238 of 12,472, Fisher's exact test $p < 10^{-6}$). This trend grows stronger when considering genes that were predicted to have even larger differences between *S. pombe* and *S. cerevisiae*. This analysis independently confirms predictions of highly rewired genes between the two species and suggests that changes at the level of gene expression regulation are at least one mechanistic factor that contributes to these differences.

2.7 Conclusions

Although individual interactions and gene-specific properties may not be strongly conserved between species, our findings suggest that these properties influence genetic interaction networks in a similar manner. For example, while the genes important for normal growth may vary, the relationship between a gene's fitness contribution and the genetic interactions it exhibits appears to be conserved. Indeed, models trained on both *S. cerevisiae*- and *S. pombe*-derived gene properties were significantly predictive of cross-species genetic interaction degree (Figure 2.2C), suggesting that the general principles governing genetic interaction network structure are retained through evolution. Thus, a complete genetic interaction network for an organism such as *S. cerevisiae* should serve as a reference network to guide studies to uncover genetic interactions in more complex systems. Predicting specific pairwise interactions across species is of course the next (more difficult) challenge, but models that can accurately predict the variation in number of interactions across the genome provide a foundation for cross-

species interaction analysis. Our results also demonstrate that integrative comparisons leveraging multiple functional genomic datasets across species may be one approach to build confidence in differential network analysis. As more data become available, both the extent and nature of network conservation should reveal how functional conservation and divergence can be recognized and utilized in distantly related species.

2.8 Methods

2.8.1 Models and evaluation

Our models are bagged regression trees that use the 16 features described in section A2.2. Breiman (Breiman, 1996) suggests that using an ensemble of only 25 classifiers can result in nearly all improvement gains that bagging can produce over a single classifier; however, we used 100 trees because the computation required in training is relatively low and we were interested in analyzing the tree structures. Individual trees were trained by MATLAB's classregtree function, which minimizes node impurity according to mean squared error. For each tree, a bootstrap sample was used to select, with replacement, a set of training genes the same size as the set of total genes (therefore each tree is trained on approximately 63.2% of all genes) and held out genes. The final prediction for a single gene of the species used to train the model (that is, the within-species prediction) is the median of all predictions from trees for which the gene was not in the training set. The final prediction for a gene of the species not used to train the model (that is, the cross-species prediction) is the median of all predictions from all trees.

To assess the performance of the model, we calculated the Pearson correlation coefficient between predicted and actual degrees of genes with known degrees. To estimate stability of performance, we repeated the model construction and evaluation 25 times and reported predictive ability as the mean Pearson correlation coefficient and its standard deviation across all 25 repetitions for within- and cross-species cases (Figure 2.2C).

2.8.2 *S. pombe* genetic interaction screens

Eight whole-genome *S. pombe* genetic interaction screens were completed using the method described in (Dixon et al., 2008). The query strains were deletion mutants for each of the following genes: SPCC1682.08c, SPBC21D10.12, SPBC13E7.09, SPAC4G8.13c, SPAC3A11.13, SPAC27D7.13c, SPAC22F3.09c, and SPAC16A10.07c. The resulting double mutant colonies were processed as described in (Baryshnikova et al., 2010b). Negative interactions were derived from the scores by applying an interaction cutoff of <= -0.08 and P-value cutoff of < 0.05. Degree measurements were then derived for all nonessential genes by counting the number of significant interactions across the set of eight queries. Significant correlation with the predicted degrees was also observed when a stricter cutoff was applied (interaction score <= -0.12, P-value < 0.05 yielded a correlation r = 0.41, P-value < 10^{-117}).

2.8.3 Rewiring groups and significance assessment

To make comparisons between degrees of orthologs in the genetic interaction networks of the two yeast species, we considered genetic interaction degree to be predicted percentile for all *S. pombe* genes, while percentiles of actual degrees were used for *S. cerevisiae*.

To search for groups of functionally related genes that have been rewired since the divergence of *S. pombe* and *S. cerevisiae*, we defined gene groups in two ways. The first simply grouped genes whose protein products form a complex in a set of complexes defined in (Baryshnikova et al., 2010b). The number of proteins per complex ranges from 2 to 81, with the vast majority having six or fewer proteins.

The second method for making sets of functionally related genes grouped genes that share a biological process GO term annotation (Ashburner et al., 2000). We considered GO terms that are annotated to greater than 3 and fewer than 50 genes in either of the two species. Additionally, a group of *S. cerevisiae* genes was required to have a minimum number of two genes with known genetic interaction degrees; a group of *S. pombe* genes was required to have a minimum of two genes with known fitness defect. Since GO terms tend to be highly redundant, we filtered gene groups so that no pair of groups overlapped by more than 50% of either group's genes.

To determine orthologous pairs of groups that have significantly different average degrees, we calculated the difference between the median degrees of genes in each species's group, and then compared the differences to a distribution of differences produced from randomly grouped genes. We generated this background by creating groups of randomly selected genes in one species, then identifying orthologous groups in the other species composed of the selected genes' orthologs. A query gene-group pair was compared to a background containing only random gene-group pairs whose group sizes were identical to the query groups. For example, a protein complex of five individual *S. cerevisiae* proteins may contain four genes that have *S. pombe* orthologs; this query gene-group pair would be compared with a background of groups with five random *S. cerevisiae* genes matched with a group of four of their *S. pombe* orthologs.

2.8.4 Comparative analysis of co-expression networks

To independently validate genetic interaction degree differences across species, we performed a comparative analysis of co-expression networks of the *S. cerevisiae* and *S. pombe* genes. The *S. cerevisiae* network was previously published (Huttenhower et al., 2006) and is based on integration of a large collection of expression datasets. To construct the *S. pombe* network, data from nine expression studies were collected from the GEO database (Barrett et al., 2011). Genes with missing values for more than 30% of the samples were removed, and the remaining missing values in each dataset were imputed using KNNImpute (Troyanskaya et al., 2001). Datasets reflecting probe intensities (rather than relative ratios) were log-transformed. After processing, the nine *S. pombe* expression datasets were integrated as described in (Huttenhower et al., 2006; Huttenhower et al., 2008). The naïve Bayes approach for dataset integration requires a gold standard set of positives, for which we used direct gene co-annotation to any term in the GO that contained between 2 and 100 genes. *S. pombe* gene annotations were downloaded from the GO website (Ashburner et al., 2000; The Gene Ontology Consortium, 2012) in May 2011. All analysis and integration of expression data were completed using the Sleipnir library (Huttenhower et al., 2008).

We applied a 95^{th} percentile cutoff to edges in both the *S. cerevisiae* and *S. pombe* co-expression networks, such that only the highest scoring 5% of edges were retained.

To estimate the overlap between the *S. cerevisiae* and *S. pombe* networks in the absence of biological conservation, we randomized the edges of the *S. cerevisiae* network and considered the background conservation to be the overlap between this randomized network and the *S. pombe* network. The randomizing procedure repeatedly chose two random edges that do not share an endpoint and exchanged an endpoint of one edge with an endpoint of the other edge, thus maintaining the degrees of genes in the network. The number of endpoint swaps performed was 20 times the number of edges in the network, which is a sufficient number of swaps to remove the original relationships between genes.

Chapter 3: Functional annotation of genes with network modules

3.1 Chapter overview

Biological networks, including GI, PPI, and co-expression networks, are frequently used to assign functions to genes. This often involves identifying network clusters that represent functional modules. Previous work has been successful in identifying functional modules, which demonstrates quality of data sets and the validity of using cluster membership to annotate genes. However, many of the most popular clustering methods treat module detection as a single problem with a global solution that describes a data set by breaking it into largely distinct components. This ignores the very common phenomenon of pleiotropy, in which a gene is involved in many functions.

In this chapter, we describe a gene-centric method to create functional profiles of genes based on their genetic interactions. Because modules have been shown to not only be a dominant feature of GI network topology but also directly correspond to functional processes of the cell, we opt to first extract modules from the GI network, and then characterize genes by their containing modules. We use biclusters discovered through frequent item set mining because they provide highly significant modular context to nearly all interactions in the GI network

The main components of our strategy to define a functional profile include a systematic method for selecting biclusters that represent the functions of each strain in the yeast negative genetic interaction network (Costanzo et al., 2016) and annotation of biclusters with high-level functional processes to summarize each gene's participation in different aspects of cell biology. The most salient functions in the gene functional profiles closely matched gold-standard annotations of the genes, demonstrating the accuracy with which they capture functions. But because these functional profiles are much more complete than previous annotation sets, they can be used for assessing gene pleiotropy, an application that is discussed in Chapter 4.

Raamesh Deshpande and Jeremy Bellay provided helpful code used in the application of XMOD. All other aspects of analysis were performed by the author of this dissertation. The work presented in this chapter builds on some ideas suggested in Bellay et al. (2011a).

3.2 Background

3.2.1 Popular methods of identifying clusters in biological networks

Community detection in networks is a highly studied area and there are many algorithms designed for general network cluster detection. Cluster analysis applied to PPI networks generally aims to identify dense subnetworks under the expectation that most will represent protein complexes or tight functional modules. To this end, many clustering algorithms calculate local densities of neighborhoods (Brohée and van Helden, 2006; Shih and Parthasarathy, 2012). MCODE (Bader and Hogue, 2003), for example, identifies areas of high density by seeding all clusters with the nodes that have the highest clustering coefficient when considering only neighbors with degrees high enough to meet a pre-determined threshold. Markov clustering (MCL)(Van Dongen, 2001) is a particularly popular method that forms clusters through simulating random walks. This algorithm alternately calculates random walk distributions from every node, updating the network by formation of new connections between nodes, and strengthens links within well-connected groups of nodes. The outcome of this iterative process is a set of prominent nodes that are each connected to a set of nodes that have no other connections; each of these star-shaped components defines a cluster.

The accuracy of these methods is typically measured by comparison to gold-standard sets of complexes or to GO terms, which allows authors to design algorithms with a practical balance between precision and recall. This often results in low coverage of network nodes by the clusters: Shih and Parthasarathy (2012) reported node coverage by clusters from 15 different algorithms in three PPI networks and the large majority of clustering results covered under 70% of nodes, many covering less than 50%. Low node coverage indicates that these clusters are not representative of a large fraction of the genome.

Clustering of gene expression profiles is performed to identify sets of co-expressed genes. Some clustering methods treat each profile as a high-dimensional data item; other algorithms operate solely on profile similarities, which amounts to clustering the co-expression networks. Co-expression gene clusters are usually not measured against a gold standard set of modules, but are evaluated by traditional, more generic metrics that compare similarity of genes clustered together to the similarity of genes assigned to different clusters. Two popular, and illustrative, algorithms for co-

expression clustering are CLICK and c-means. CLICK (Sharan and Shamir, 2000) is a divisive algorithm that makes minimum-weight cuts on a weighted graph (derived from a profile similarity network), splitting connected components until each is expected to contain only nodes that are part of the same cluster. In between multiple rounds of divisive cluster formation, any non-clustered nodes are used to expand the clusters by comparing original profiles to profiles that have been calculated for the clusters. The fuzzy c-means (FCM) algorithm (Bezdek, 1981; Dunn, 1973), has also been applied and further developed for discovery of co-expression clusters (Gasch and Eisen, 2002; Maji and Paul, 2013). FCM is similar to the k-means algorithm, but allows items (e.g. profiles) to belong to multiple clusters with partial membership. The output is a so-called fuzzy partition, in which each item is associated with a membership weight for each cluster, such that all weights sum to one. Given the memberships of all items to all clusters, a cluster's centroid can be calculated as a weighted combination of all items. After initiation of a specified number of random profiles as the cluster centroids, FCM iteratively determines each item's cluster memberships and updates the centroids according to the new membership.

Many authors have dwelled on general topological properties that are common to diverse types of networks (Barabasi and Oltvai, 2004; Clauset et al., 2008), including genetic interaction networks (Tong et al., 2004), such as hierarchical modularity and power-law-like degree distributions. They suggest that one clustering scheme should work well for all these networks (Clauset et al., 2008; Girvan and Newman, 2002; Palla et al., 2005). However these topological statistics are superficial: modular structures in GI networks often follow the distinct patterns of the within- and between-pathway models. Thus it is unlikely that algorithms designed to find organic and sprawling modules in PPI networks or expansive co-expression modules will be able to home in on typical GI structures. Despite this, few targeted module-detection algorithms have been applied to large GI networks. By far, the most common practice in analyzing GI network clusters is applying hierarchical clustering to both sides of the network's matrix representation and manually browsing the clustered matrix. Hierarchical clustering, formulated as an agglomerative algorithm, starts by treating individual GI profiles as clusters and iteratively merges the two most similar clusters, as determined with a profile similarity metric and a method of comparing two clusters. For example, Pearson's correlation coefficient (PCC) is often used to compare gene profiles and two sets

(clusters) of profiles may be compared by calculating the maximum PCC of all pairwise profile comparisons between the two clusters. Clusters are merged until all profiles have been joined in a single top-level cluster. In the next section, we discuss how hierarchical clustering methods have been specifically applied to GI networks. However, afterwards, in section 3.2.3, we introduce a method that is able to effectively capture the network structures that are most common among genetic interactions.

3.2.2 Systematically annotating genes with clusters

The application of network clustering does not always have modules as an end-goal. Module discovery may be used to find functional information about specific genes. Two recent publications argue that the topologies of large-scale biological networks are so rich that systematically derived clusters should be used to assign data-driven annotations to genes. The tool NeXO (Dutkowski et al., 2013) builds an ontology based on a hierarchical clustering dendrogram, while SAFE (Baryshnikova, 2016) uses the spatial layout of a network to locate modules in overlaid data. While both methods make substantial use of GO terms, they are driven by network structure and the resulting annotations are un-biased, or less biased than GO, and able to suggest functions for all genes in the network, including completely uncharacterized genes. They also reiterate curated information, from specific modules to the hierarchical organizations of modules. For example, NeXO was applied to a combination of yeast biological networks and identified 60% of GO cellular compartment terms in addition to hundreds of modules that did not map to any GO term.

Although somewhat successful in their goal of automatic gene annotation, NeXO and SAFE fail to represent pleiotropy of genes, a property that has been long-recognized to be common and is well-represented in GO. NeXO is reliant on global similarities of network profiles and assigns each network node to only one set of close neighbors. Similarly, the SAFE publication highlights global profile correlations as an ideal data type to use. Further, SAFE constrains network nodes to appear in one location in a network layout that determines which nodes will be grouped during statistical analysis; this could cause a multifunctional node to be located in between multiple modules that all represent its true functions, but outside of the statistically enriched areas for all the modules. In these ways, both methods strongly emphasize a one-node-one-function

assumption. This assumption is also common in network clustering algorithms that identify dense subnetworks: the vast majority are not local because they nearly always attempt to optimize the entire set of clusters found, likely including some clusters at the expense of others (e.g. clusters emerge simultaneously as algorithms perform iterative updates affecting many memberships in FCM, MCL, etc., or try to optimize an objective function explicitly as in FCM). From another perspective, most algorithms only produce disjoint clusters (e.g. MCL, CLICK) or only allow overlapping clusters in a restricted manner (e.g. fuzzy c-means allows multiple memberships by decreasing membership weights, MCODE). The requirement of disjoint clusters precludes the ideal identification of a cluster containing a gene that has partial profile similarity with genes that are in separate clusters.

3.2.3 Biclusters

Biclustering is a type of module discovery in which groups of genes are assessed for local, as opposed to global, similarity (Hartigan, 1972). Given a matrix representation of a data set, a bicluster is composed of a subset of rows paired with a subset of columns. For example, biclustering methods have been actively developed for identification of co-expressed genes. In this context, a bicluster groups genes that have similar expression patterns over some, but not all, of the experiments or time points. Genes that are members of multiple regulatory groups may fall into two or more biclusters that each represents a different subset of data samples. Ideal solutions to the problem of discovering coherent biclusters in real-valued data, like expression data, are infeasible because they would require solutions to NP-hard problems, such as finding a maximum weighted subgraph of a bipartite graph or covering a bipartite graph with a minimum set of bicliques (Cheng and Church, 2000; Tanay et al., 2002). Additionally, different data sets and applications required discovery of different types of patterns, like genes with expression patterns that scale with each other or expression that is constant throughout the bicluster. These two challenges motivated the development of a large variety of algorithms. The first application of biclustering to expression data used a greedy approach that begins with the entire data matrix and removes rows and columns to produce a bicluster meeting a consistency threshold (Cheng and Church, 2000). Another used a combinatorial search for a heavy subgraph of the gene-condition

bipartite graph (Tanay et al., 2002). Still another approach used Gibbs sampling to model the expression patterns of biclusters and determine gene and condition membership (Sheng et al., 2003). Many algorithms initialize a random bicluster and perform a local search that considers additions and removals of rows and columns that may improve the consistency of the expression patterns in the bicluster (Bergmann et al., 2003; Ihmels et al., 2002; Reiss et al., 2006). As the number of algorithms expanded, researchers continued creating variations that were more efficient and more inclusive of expression patterns. While the algorithms designed for expression data could be applied to genetic interactions, they generally do not provide any guarantees about the completeness of the discovered biclusters, tend to emphasize large structures, and avoid overlap of clusters. For these reasons, most biclustering algorithms are not ideal for genetic interactions.

Bellay et al. (2011a) made the observation that the between- and within-pathway models of genetic interactions will be captured precisely by bicluster structures. In a genetic interaction network, a bicluster can be described as a set of genes (or more precisely, strains) that, as a group, show dense interactions with a second set of genes (strains). While the two subsets of genes are experimentally distinguishable (e.g. query and array strains in SGA), the interpretation of each side of a bicluster is the same—a functional module. The Bellay et al. (2011a) study discovered biclusters using frequent item set mining, which is guaranteed to find all possible biclusters with complete density in a binary network. The exhaustive nature of the method proved highly useful, as the authors were able to draw relatively definitive conclusions about the frequency of the GI pathway models that are composed of negative or positive interactions. Importantly, they found that biclusters covered over half of the negative interactions in the GI network. This indicates that exhaustive biclusters are an ideal data set to help characterize not just modules, but the multiple functions of individual genes, and therefore are an ideal basis for functional profiles.

3.2.4 Frequent item set mining

General description

Frequent item set mining is a method to discover repeated co-occurrences of items in a large collection of sets of items. Although there are many diverse applications

of frequent item set mining, this field of data mining was originally developed to analyze consumer purchases from a store, and consequently, some terminology reflects this type of data. The set of all items that appear in the database to be analyzed is the item base $B = \{i_1, \dots, i_n\}$, which may be, for example, all the products a store sells. Any subset of the item base is called an item set. The database $T = \{t_1, \dots, t_m\}$ is composed of many item sets, termed transactions. Consistent with the idea that a single customer with a specific set of needs bought items as one purchase, the items in a transaction are suspected to have some underlying connection. Further, items that appear together in many transactions may have an important connection that warrants labeling them as an interesting group. Given the transaction database T, the support of an item set, $s_T(I)$, is the number of transactions that contain the item set. Any item set with support meeting a user-defined minimum support threshold is called a frequent item set, and the discovery of such sets is the goal of frequent item set mining.

Although the data description and the following algorithm description are asymmetric, there is no required inherent difference between items and the mechanism that groups items into transactions, such as a customer. Any two dimensional, binary data set can be used, and this includes networks. In analyzing the genetic interaction network, items are genes (e.g. drawn from the network columns) and each transaction is the set of genes that interact with a single gene (e.g. drawn from the network rows).

Because there are 2^n item sets in B and it is neither efficient nor necessary to calculate support for all of them, frequent item set mining algorithms define the search space carefully. The support of any item set $J \subseteq B$ cannot exceed the support of any subset $I \subseteq J$, i.e. $s_T(J) \leq s_T(I)$. Therefore, if an upper bound is determined for the support of I, that bound will apply to J as well. Incorporating the minimum support into this idea, we have the Apriori property:

$$\forall I \subseteq J \subseteq B: s_T(I) < s_{min} \longrightarrow s_T(J) < s_{min},$$

which states that a superset of an infrequent item set cannot be frequent. This property forms the basis for an efficient algorithm by defining where the search space may be pruned.

Given the relationship between an item set and its subsets, a natural method of exploring the search space begins with the smallest item sets and works its way through larger sets, only calculating the support of a larger set if all subsets are known to be

frequent. The enumeration of the subsets of B therefore can be structured as a prefix tree, which requires items to be ordered and represents item sets as sequences that are uniquely defined by a path from the root to a leaf (Figure 3.1A). The root of the prefix tree represents the empty set and each edge signifies the addition of an item. The edges follow the constraint that any path through the tree must only add items in order. This defines a one-to-one mapping between item sets and sequences, preventing a set from being generated more than once.

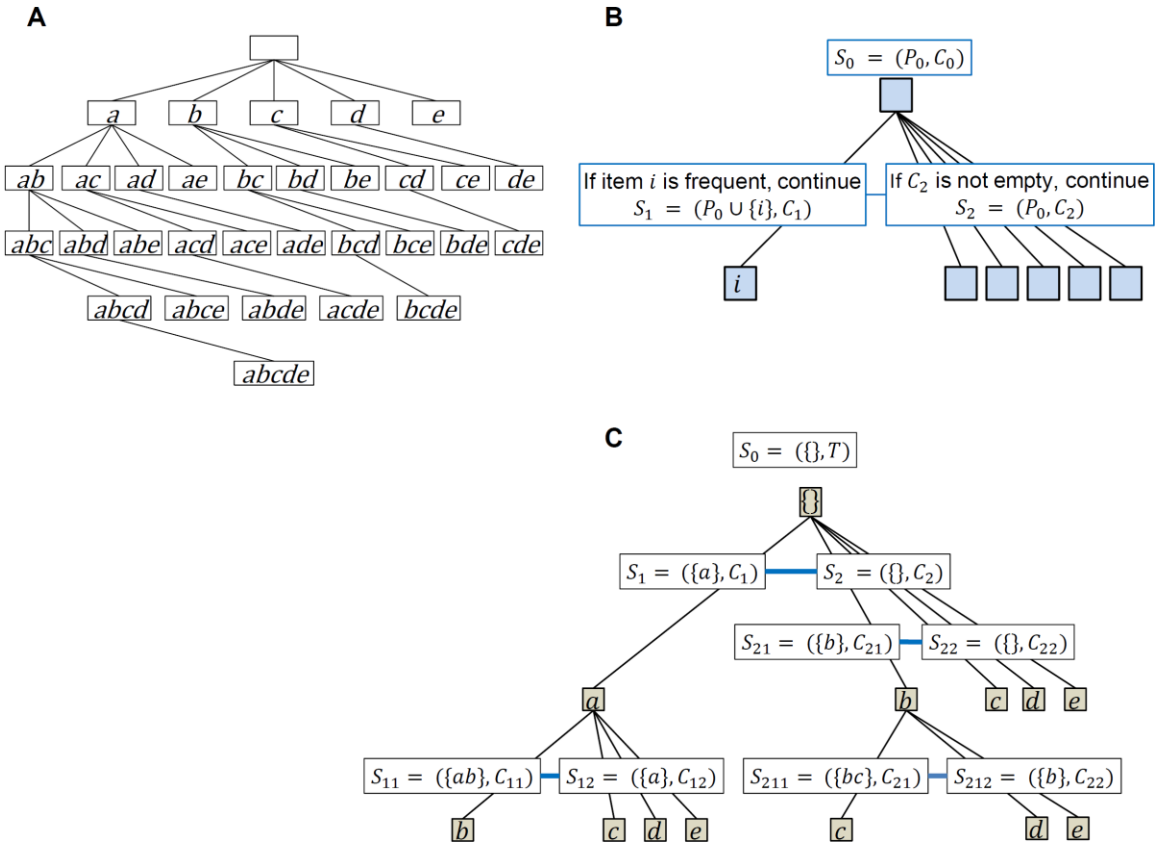


Figure 3.1. Diagrams of trees used for set enumeration. (A) The complete prefix tree for the item base {a,b,c,d,e}. **(B)** The recursive subproblems defined for any position in the prefix tree. **(C)** Some of the top-most recursive subproblems for the item base {a,b,c,d,e} displayed on the prefix tree. In B and C, the subproblem definitions overlap exactly the edges leading to nodes whose associated prefix will be explored by the specified subproblems.

The Eclat algorithm

The Eclat algorithm (Zaki et al., 1997) defines a recursive procedure that incrementally extends frequent item sets with a given set of available items, reporting new frequent item sets as they are discovered. A single recursive call sets up a divide-and-conquer strategy, creating two subproblems based on an initial item set and an item i that may be used to extend the initial item set. The first subproblem discovers all frequent item sets that include item i and the second subproblem discovers all frequent item sets that do not include item i. This strategy is a depth-first exploration of the prefix tree described above. To give a formal explanation of Eclat, we trace a path through the prefix tree while accounting for the operations that maintain an efficient database.

A subproblem S, which is solved at each node of the prefix tree, is expressed by its two associated inputs as $S = (P, C)$, where P is a prefix item set and C is a conditional transaction database. P is a frequent item set and will be added to every frequent item set subsequently found; it corresponds to a unique position in the prefix tree. The database C contains only the transactions that include P and items that have not been investigated. The transaction database is organized by associating each item with a list of all transactions that contain it. These lists, called transaction ID (TID) lists, enable efficient modifications to C that mirror the depth-first traversal of the prefix tree. The initial inputs for finding all frequent item sets in the transaction set T are $P = \{\}$, which corresponds to the root of the prefix tree, and $C = T$, in which all transactions trivially contain P. An example relating subproblem definitions to the prefix tree is in Figure 3.1C.

The evaluation of each subproblem, $S_0 = (P_0, C_0)$, creates two further subproblems, S_1 and S_2, based on an item i that may be used to extend the item set P_0. This is depicted in Figure 3.1B. It first selects and removes item $i \in B_0$, which is the set of items contained in C_0. Next, it determines if $P_0 \cup \{i\}$ is a frequent item set. Because all transactions in the conditional database contain P_0, the support of $P_0 \cup \{i\}$ is equal to the support of item i in C_0, which is easily determined from the size of i's TID list. If i is not frequent in C_0, the Apriori principle states that no extension of $P_0 \cup \{i\}$ will be frequent and all branches of the associated node in the prefix tree can be eliminated from the search space.

If i is frequent in C_0, then its supersets must be explored and the subproblem $S_1 = (P_1, C_1)$ is created to do so using the new prefix $P_1 = P_0 \cup \{i\}$. The new conditional

transaction database C_1will include only the data needed to evaluate supersets of P_1 and is efficiently derived from C_0: to limit the transactions to those that include P_1, the TID lists for all items are intersected with item i's TID list. The TID list for item i can now be removed because, as with all other items in the prefix P_1, item i is guaranteed to be in all transactions in C_1.

Evaluation of the second subproblem of S_0, $S_2 = (P_2, C_2)$, discovers all frequent item sets that do not include item i, but are supersets of P_0. The subproblem therefore encompasses all branches of P_0's prefix tree node other than the one associated with item i (Figure 3.1B and C). Reflecting the omission of item i, the prefix P_2 is set to P_0 and item i's TID list is removed from C_0 to create C_2. If C_2 is not empty, the subproblem S_2 is completed by a recursive call. If it is empty, then there are no further extensions of P_0 that are frequent and no recursive call is made.

Filtering frequent item set results

A perennial problem in frequent item set mining is that frequent item sets typically overlap considerably as a result of noise. The simplest case is when a frequent item set is a subset of another frequent item set and both have the same support, which implies that both are supported by the same set of transactions. There is no information contained in the smaller set that isn't also in the larger set, so the smaller can be safely discarded. More formally, only an item set that cannot be extended without decreasing its support needs to be kept; such a frequent item set is called closed. For every frequent item set, there is exactly one superset that is closed. Many applications that use an asymmetric interpretation of the data (i.e. frequent item sets are of more interest than groups of transactions) filter the frequent item sets for those that are as large as possible, called maximal item sets, and extension by any item would yield a non-frequent item set. To generate the set of all maximal item sets, any frequent item set for which a superset is also frequent can be discarded. One of these two methods of filtering frequent item sets is nearly always performed, yet for large data sets, they are usually not sufficient to produce few enough results for manual inspection or to produce item sets that are meaningfully distinct enough to summarize statistically. Thus further limitations to the frequent item sets are applied.

In the method we developed, described below, a heuristic is used to determine biclusters that are most likely to represent functional modules. By discarding some overlapping biclusters, we are left with a set that are likely to represent distinct functional modules.

3.3 Procedure for identifying GI biclusters

3.3.1 Bicluster discovery using frequent item set mining

Our first task in creating a functional profile for every yeast gene was identifying functional modules composed of negative genetic interactions. We opted to search for biclusters with frequent item set mining based on two benefits: first, there is no limit on the number of biclusters a gene can have membership in and, second, frequent item set mining is exhaustive, finding all possible dense bipartite structures in a network. Within an SGA-derived GI network, a bicluster takes the form of one set of query strains and one set of array strains, with interactions occurring between all query-array pairs; this structure is a complete bipartite subgraph in the network.

To apply frequent item set mining to the negative interactions in the most recent yeast genetic interaction network (Costanzo et al., 2016), we used the XMOD procedure, which was developed by Bellay et al. (2011a) to determine the statistical significance of biclusters. In this method, each bicluster is assigned a p-value, calculated by a comparison to biclusters mined from ten randomized versions of the network that preserve the degree distribution of the original network. Specifically, all biclusters are assigned a score that represents the likelihood of all their contained interactions occurring if genes interacted randomly, conditioned on the genes' interaction degrees. Then, the scores of the random biclusters are used as a null distribution to assign p-values to the real biclusters, which are expected to have lower scores due to non-random gene associations. Biclusters with p-values higher than a chosen significance level are discarded.

A number of steps were taken to prepare the GI network for bicluster discovery. The network from Costanzo et al. (2016) is actually composed of two distinct data sets: the TSA (temperature sensitive array) and DMA (deletion mutant array) networks. Each was handled separately and identically. Because frequent item set mining requires

binary input, we took the preliminary step of binarizing the networks by defining interacting strains as pairs with significant SGA genetic interaction scores less than or equal to -0.08, according to an established intermediate cutoff. We additionally added self interactions to the network, so that a strain could occur on both sides of a bicluster; this would be expected in the case of a set of genes all interacting with each other, and self interactions are trivial functional relationships.

Next, we accounted for the fact that some yeast genes are associated with multiple strains. The set of queries in each GI network includes mutant strains with DaMP and temperature-sensitive alleles of essential genes, often with multiple alleles of a single gene. Due to the biased multiplicity of many essential genes within the set of query strains, bicluster datasets generated from the complete data set may be uninteresting or difficult to interpret because bicluster significance would be driven by the highly correlated profiles of alleles of the same gene. To overcome this problem, we produced replicates of the binary networks, with each replicate containing one randomly selected allele for each gene. Using many replicate networks yields good representation of different alleles and allows different combinations of alleles to be chosen. In all, 15 replicates of each of the DMA and TSA GI network were created.

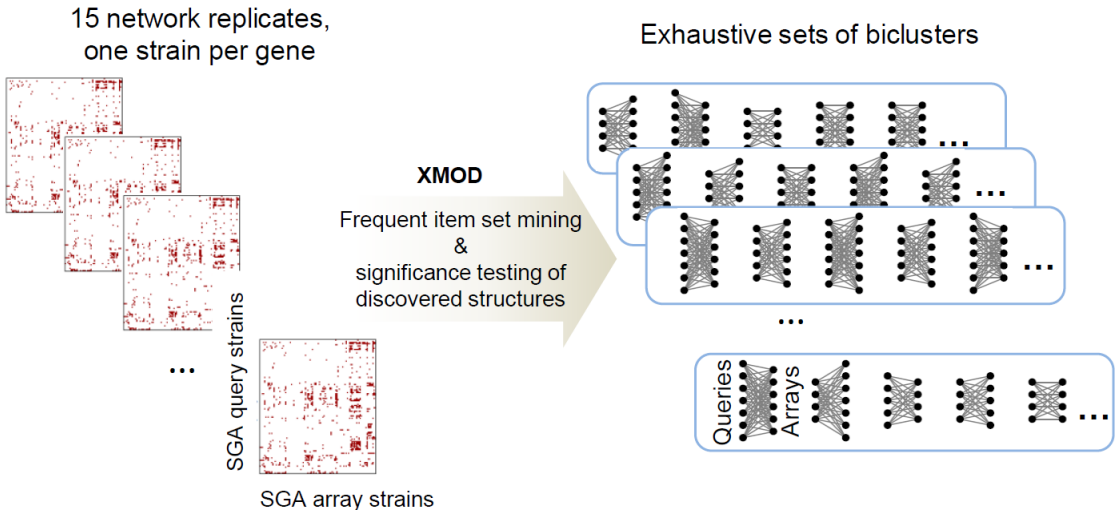


Figure 3.2. Application of XMOD to one SGA genetic interaction network. Each replicate contains one randomly chosen strain for each gene and is individually mined for biclusters. Each bicluster contains a set of query strains and a set of array strains.

Figure 3.2 depicts bicluster discovery for one SGA network: network replicates were input to separate runs of frequent item set mining and XMOD. All frequent item set mining described here was performed using an implementation of the Eclat algorithm (Zaki et al., 1997) by Christian Borgelt, which is available at http://www.borgelt.net/eclat.html, using the "-tc" option to report only closed item sets. A single test run on the DMA network yielded over 37 million biclusters with a size of at least four query strains and four array strains. Based on the observation that biclusters with sizes 4x4, 4x5, and 5x4 make up ~38% of discovered biclusters, yet only ~2.7% of these are significant at a p-value threshold of 10^{-4} (Figure 3.3), we used the Eclat option to remove from the results all biclusters of these three mentioned sizes and those with either dimension smaller than four in order to reduce memory usage in XMOD; the Eclat option string to accomplish this is "-s-4 –m-4 -F-6-5-4". After eliminating these small biclusters, the DMA network replicates contained an average of ~24.5 million biclusters and the TSA network replicates contained an average of ~20 million biclusters.

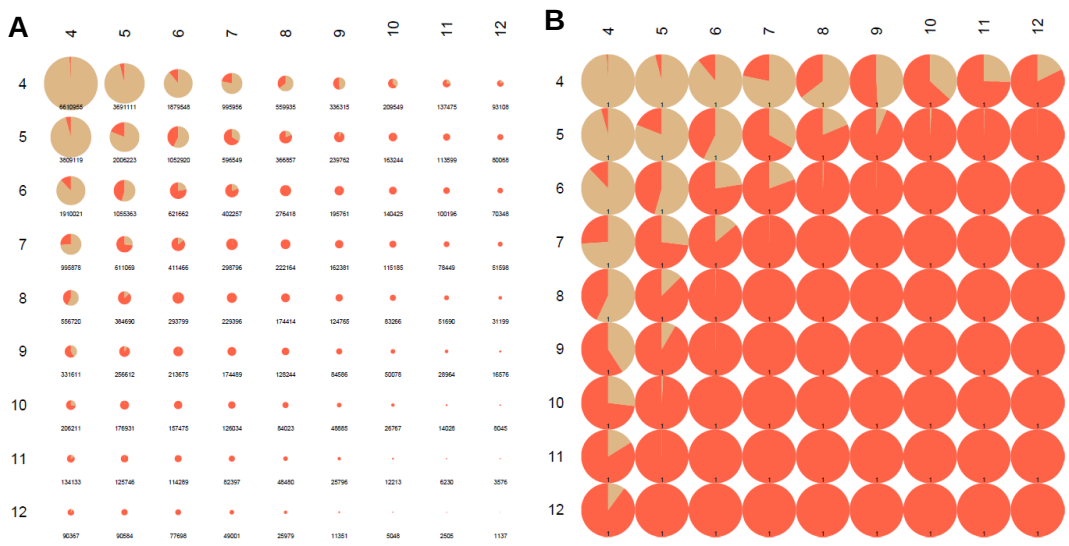


Figure 3.3. Percent of discovered biclusters that are significant. Axes describe the dimensions of the biclusters, with the number of query strains on the x-axis and the number of array strains on the y-axis. The red section of each pie chart shows the fraction of biclusters with p-values < 10^{-4}. The size of the pie charts in **(A)** show the number of discovered biclusters for each size; **(B)** shows only the fraction significant.

These grids of pie charts are truncated; there were many biclusters with dimensions larger than 12.

As described in Bellay et al. (2011a) and briefly above, XMOD determines empirical p-values of biclusters through comparison to biclusters discovered in ten randomized networks. Combining biclusters from ten random networks supplies a better sampling of biclusters of larger dimensions than one random network could, however, there was an overabundance of random-network biclusters with small dimensions (e.g. 4x6, 6x4, and 5x5). So for better speed and memory use, we randomly discarded biclusters to keep a maximum of two million for each size.

All biclusters with p-values higher than 10^{-4} were discarded, leaving ~14 million (57%) DMA and ~10 million (50%) TSA biclusters per network replicate. The vast majority of biclusters containing 30 or more interactions (e.g. 6x6, 7x5, and larger) were significant (Figure 3.3), since very few large biclusters were found in the randomized networks.

3.3.2 Selection of biclusters for a non-redundant set

Bicluster discovery through frequent item set mining produces modules that overlap, i.e. a bicluster usually shares some of its interactions with other biclusters. While this certainly reflects reuse of genes in different cellular functions, it is also caused by our inability to discover larger modules that are fractured by false negatives (biological or technical). Because our goal is to simply identify gene membership in different functional modules, it is not necessary that modules be recovered in their entirety. However, it is important to remove the redundancy of biclusters that reflect the same functional module in order to prevent over-counting the functional memberships of genes. We do so by making pairwise comparisons of biclusters and discarding one bicluster whenever a pair has too much overlap. Our application of this bicluster-removal process specifically aligns with our goal of creating a functional profile for every gene: we do not remove any biclusters globally, we make removals separately within each single-mutant strain's set of containing biclusters. In this way, removal of a bicluster does not remove a gene's membership in the associated functional module because an overlapping bicluster remains to represent the module.

To remove redundancy from the sets of biclusters associated with each single-mutant strain, we used a method described in Bellay et al. (2011a). The procedure is greedy and proceeds as follows. First, order all biclusters from best to worst. Then, select the biclusters in order and upon the selection of a bicluster, remove overlapping biclusters from any future consideration. We defined “overlapping” as the smaller bicluster having 10% or more genetic interactions in common with the larger.

To define the best-to-worst ordering, we determined preferences for different bicluster sizes, and built a size-lookup table to pick between differently sized biclusters. For our use, the quality of a bicluster can be measured by how well it reiterates a set of genes annotated by a GO term. We selected Jaccard similarity between the bicluster gene set and an enriched GO term as a simple statistic to measure this. Since calculating GO term enrichments on all bicluster gene sets would take too long, we used a sample of biclusters to rank bicluster sizes (expressed in two dimensions) according to results from Jaccard similarity analysis. First, for each bicluster size, we collected all biclusters up to a maximum of 10,000 and removed redundancy from this set by consecutively selecting biclusters in random order and removing any other bicluster from future selection if more than 10% of its interactions overlapped with the selected one. Next, statistically enriched GO terms were determined for every bicluster (using genes from one side) and the maximum Jaccard similarity obtained from each bicluster was recorded, yielding a distribution of maximum Jaccard similarities for each bicluster size. Finally, as a summary statistic representing likelihood of reflecting a known module, we kept the median Jaccard score for each bicluster size, organized as a lookup table to consult. This analysis was done separately for the DMA- and TSA-derived biclusters.

A size preference table is specific to the bicluster dataset (either TSA or DMA) and the side of the bicluster (query-strain or array-strain) that is intended to be used in further analyses. Therefore, a total of four tables were created (Table A3.1).

For both the TSA and DMA negative genetic interaction network, every strain has 15 sets of biclusters that it appeared in, one set from each of the network replicates. All of these individually had redundancy removed twice: first for the purpose of annotating the query-strain sides of biclusters, and second for the purpose of annotating the array-strain sides, using the appropriate bicluster size preference table in each case.

3.4 Bicluster-derived functional profiles for genes

3.4.1 Annotation of biclusters

After identifying non-redundant sets of biclusters for strains in the GI network, the remaining steps in creating functional profiles summarize the biclusters in terms of their functions and collect network replicates. The final functional profiles are calculated for strains, since those are represented in the genetic interaction network. However, these are equally considered gene profiles since each strain is associated with a single gene's functions.

Biclusters can be functionally annotated using the annotations of genes that are represented by their constituent strains. As alluded to earlier, biclusters may be annotated based on either of their two strain sets, the query strains or the array strains. The set chosen determines the interpretation of the bicluster's functional annotation in relation to the member strain of interest. If the strain of interest is within the set used to determine the bicluster annotation, then the annotation should reflect a function the strain participates in. If not, the annotation is based on the strain's interactors, and represents a functional relationship that may be direct or indirect. We analyze query-strain profiles and use bicluster annotations derived from the query-strain gene sets the remainder of this chapter and in Chapter 4.

We annotated the query-strain side of each bicluster with biological process terms that have been manually annotated (MA, Table A3.2) or systematically annotated (SAFE, Table A3.3) to yeast genes. Every bicluster was annotated by any term for which its gene set had significant enrichment or to which at least 40% of queries were annotated. In most cases, that majority of each strain's set of biclusters received annotations (Figure 3.4), meaning this layer of abstraction from specific biclusters to high-level annotations is likely to faithfully represent the breadth of a gene's module memberships. In the TSA network, ~94% of strains have at least 50% coverage of their biclusters with MA annotations and ~42% have at least 80% coverage; for the DMA network these numbers are lower, at ~76% and ~18%, respectively.

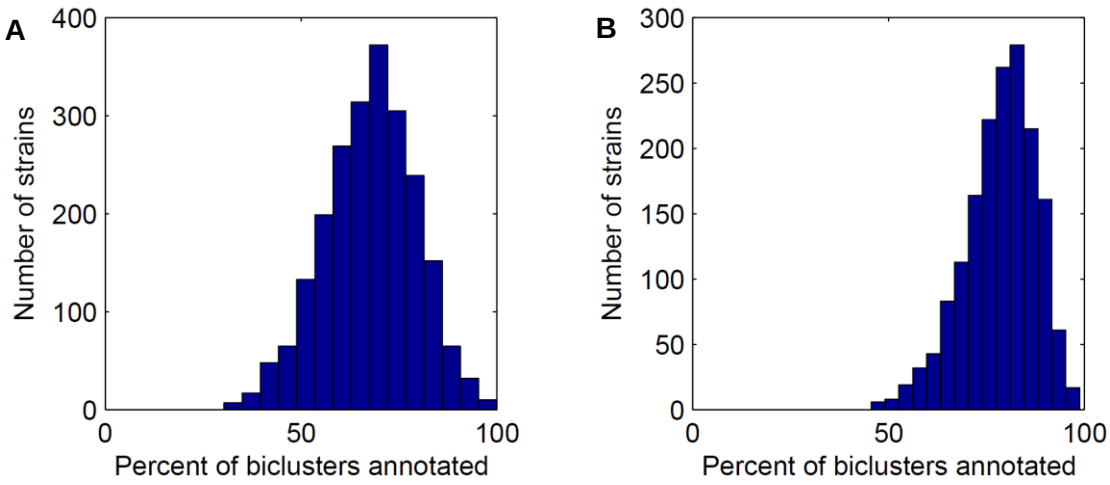


Figure 3.4. Distribution of annotation coverage in sets of biclusters associated with individual query strains of the DMA (A) and TSA (B) GI networks. MA terms were used to annotate query-strain sides of biclusters based on gene annotations.

The numbers of annotations to each term were counted and normalized within each of the 15 sets of biclusters associated with each query or array strain, creating functional profiles. These replicate profiles were averaged, yielding one bicluster-based functional profile per strain.

3.4.2 Validation of bicluster functional profiles

To compare each gene's bicluster-derived annotation profile to its gold standard annotations, either MA or SAFE, we used a simple one-dimensional version of the clustering algorithm DBScan (Ester et al., 1996) to find the most striking highly annotated process or processes for each functional profile. Before clustering with DBScan, we normalized each profile by dividing all elements by their maximum value. Our implementation of DBScan visits values from highest to lowest and labels a value as an outlier if it has no neighbors (the "minPts" parameter is 1) at a distance of less than 0.2 (the "Eps" parameter), otherwise it defines a cluster and expands the cluster following the standard algorithm. We defined profile-predicted annotations as (1) all outliers that are higher than the first cluster, if there are any, or (2) the highest cluster, if there are no outliers. For ~90% of profiles, DBScan identified only one or two annotations.

We calculated precision and recall statistics that assess the similarity of these predictions to the MA and SAFE gold-standard annotation schemes (Table 3.1). Since genes can have more than one gold-standard annotation, we calculated precision and recall separately for genes with one and two annotations. For genes with two annotations, we defined a true positive two ways: as at least one prediction matching a gold-standard annotation and as two predictions matching both gold-standard annotations. Precision and recall values were generally very high, indicating that the functional profiles accurately capture known annotations of genes.

Additionally, we demonstrated the usefulness of annotated biclusters over annotated GI partners for making accurate process predictions. We performed the DBScan and precision-recall analysis using the MA gold-standard annotations of a gene's negative genetic interaction partners (at both the intermediate, -0.08, and strict, -0.2, SGA score interaction thresholds) to build a functional profile (Table 3.2). Precision and recall are always substantially lower for predictions made by genetic interactions as compared to the corresponding statistics for bicluster functional profile predictions. The superiority of our functional profiles is likely due to the fact that modules combine individual interactions that represent only one gene function and ignore false positive interactions.

Table 3.1. Precision and recall summary for top biological processes predicted from bicluster-derived functional profiles. DMA, TSA: GI networks; MA, SAFE: gold standard annotation schemes; TP: true positives; green/red color scale indicates low to high values and matches Table 3.2.

		Bicluster-dervied functional profiles			
		DMA, MA	TSA, MA	DMA, SAFE	TSA, SAFE
Recall (%)					
All genes	1 annotation	79.7	81.5	99.1	98.6
All genes	2 annotations, >=1 TP	88.1	87.7	100.0	97.9
All genes	2 annotations, 2 TP	37.7	52.2	71.2	61.7
Precision (%)					
All genes	1 annotation	57.2	56.8	75.9	71.0
All genes	2 annotations, >=1 TP	74.9	77.2	94.7	93.8
All genes	2 annotations, 2 TP	44.9	57.6	78.7	72.5

Table 3.2. Precision and recall summary for top biological processes predicted from negative genetic interaction profiles. DMA, TSA: GI networks; TP: true positives; green/red color scale indicates low to high values and matches Table 3.1. The MA annotation scheme was used.

		Negative genetic interaction profiles			
		DMA, -0.08	DMA, -0.20	TSA, -0.08	TSA, -0.20
Recall (%)					
All genes	1 annotation	48.1	52.0	51.2	60.3
All genes	2 annotations, >=1 TP	49.3	53.6	56.6	63.5
All genes	2 annotations, 2 TP	3.6	8.0	16.1	24.6
Precision (%)					
All genes	1 annotation	31.2	29.1	29.7	31.9
All genes	2 annotations, >=1 TP	31.7	31.6	40.7	44.5
All genes	2 annotations, 2 TP	4.3	8.2	18.0	24.8

3.5 Conclusions

We have designed a pipeline that takes advantage of the modular structure of the yeast GI network to summarize individual gene participation across high-level biological processes. The comprehensiveness of the derived functional profiles represents an improvement over many alternatives. In particular, our use of frequent item set mining followed by redundancy removal performed within strain-specific sets of biclusters avoided the limitations of clustering approaches that use global calculations that may give unequal preference to the strongest network structures at the expense of others. Additionally, the recently published SGA genetic interaction network is unprecedented in its completeness of genome coverage with single mutant strains, meaning these networks may contain local structures of genes that were not included in other genetic interaction screens. Although our method is straightforward at a high level and is similar to the previous application of XMOD (Bellay et al., 2011a), nuances of the GI data set and our prioritization of individual strains required a careful, detailed implementation as well as a computationally intense execution.

While this chapter demonstrates a strong comparison between bicluster-derived functional profiles and gold standard annotations, it does so purely as validation of the fact that the profiles contain the simplistic view of the annotations. A more complete exploration of the novelty of our functional profiles is presented next.

Chapter 4: Pleiotropy derived from yeast genetic interaction modules

4.1 Chapter overview

Pleiotropy, the phenomenon of a single genetic locus with multiple phenotypic effects, has vast implications on the genotype-phenotype relationship, as well as the robustness and adaptability of cellular networks. It has previously been measured according to many definitions, which typically count phenotypes associated with genes. Although modularity is frequently—and rightly—discussed as a key organization principle in biological networks, pleiotropy has not been measured in the same network context. Therefore, an important component of gene functional behavior is still unexplored.

In this chapter, we systematically measure pleiotropy within the context of modularity by using the module-based functional profiles described in Chapter 3. Our method calculates the entropy of functional profiles to measure the spread of each gene's set of containing modules among high-level biological processes. We measure the pleiotropy of ~3200 essential and nonessential genes, which are all the genes that participated in enough biclusters to have a reliable functional profile.

We compare gene pleiotropy to our panel of gene characteristics to search for fundamental principles of how multi-module gene behavior relates to different types of genes. Pleiotropy is significantly associated with a number of gene characteristics, including some unexpected functional and evolutionary properties, like high expression variance and high copy number, which have interesting implications.

The author of this dissertation had a leading role in conceiving and planning this work; all analysis was done by this author, with contributions from collaborators. In addition to the author, Chad L. Myers, Michael Costanzo, and Charles Boone conceived the analysis. Chad L. Myers and Michael Costanzo provided suggestions regarding the analysis and its written presentation. Raamesh Deshpande and Jeremy Bellay provided code used for methods described in Chapter 3 and referenced in the current chapter.

4.2 Introduction

4.2.1 Organization of functions in biological systems

Modularity of cellular functions has become a central tenet of systems biology, supported by evidence from diverse types of genomic data. Segal et al. (2003) designed a method that, from yeast gene expression data, inferred functionally coherent sets of genes that were regulated as a group according to different conditions. Gavin et al. (2006) described protein complexes in terms of core components and attached modules, using various data as evidence that grouped proteins act as single functional units. Costanzo et al. (2010) noted that the yeast genetic interaction (GI) network is well suited to define clustering of genes at various levels, from broad high-level biological processes down to specific pathways. With an eye to evolution, Hart et al. (2007) found that most *S. cerevisiae* protein complexes are composed of genes that are either all essential or all nonessential. Further, Ryan et al. (2013) noted that complexes conserved in *S. pombe* had the same property, but notably, proteins in some complexes switched essentiality as a group, indicating that this organization persists in the context of evolutionary changes. Finally, Roguev et al. (2008) compared genetic interactions between the same yeast species and found evidence that while GIs are highly conserved within modules, a lower conservation of GI between modules allows “rewiring” to occur as the species diverge.

Despite the seemingly tidy nature of modules and their properties, considerable complexity characterizes modular organization due to substantial reuse and diverse effects of cellular components. Pleiotropy, when considered at the molecular level of genes, is the case in which perturbation of one gene influences multiple functions (Paaby and Rockman, 2013; Stearns, 2010). For example, specific subcomplexes of nucleopores play important roles in gene silencing and DNA damage repair in addition to controlling nuclear import and export (Strambio-De-Castillia et al., 2010). As another example, multiple proteins responsible for mRNA decay in the cytoplasm, such as XRN1p, have a complementary chromatin-binding function that promotes genome-wide transcription initiation and elongation, mechanistically maintaining steady state mRNA levels (Haimovich et al., 2013). Famously, mammalian apoptosis pathways are triggered by components of the electron transport chain, such as cytochrome C (Ow et al., 2008). Other genes have a single molecular function but are fundamentally upstream of diverse cellular pathways, such as the HSP90 family of chaperones, which aid the folding of

functionally diverse proteins (Taipale et al., 2010), and class V myosins, which use the actin cytoskeleton to localize mRNA and various organelles with help from cargo-specific receptor proteins (Hammer and Sellers, 2011). Because of the diverse physical interactors of the protein products, varied phenotypic effects appear when these genes are mutated.

In exploring the general notion of pleiotropy, researchers have used distinct definitions and datasets, showing that pleiotropy exists as many types of one-to-many genotype-to-phenotype relationships (Paaby and Rockman, 2013). All levels of biological organization have been considered: pleiotropy can connect DNA mutations or genes to phenotypic traits of molecular networks, cellular structures, organisms, populations, etc. Further, a phenotype may be described in the context of an environment, such as a genetic background, population, chemical, or nutrient availability. In humans, pleiotropy was recently explored by Pickrell et al. (2016), who used GWAS results to identify 341 loci in humans that are associated with multiple traits, including diseases. In yeast, phenotypic effects that stem from one gene have previously been measured by reverse genetics methods that screened the yeast deletion collection for phenotypes, such as measuring over 250 morphological phenotypes (Ohya et al., 2005) or measuring sensitivities to different stresses (Dudley et al., 2005; Ericson et al., 2006; Hillenmeyer et al., 2008, respectively assessing 21, 6, and 180 evironments). These studies variously estimate that between 5% and 30% of yeast genes could be considered pleiotropic according to counted numbers of traits or environmental sensitivities. Although different environmental challenges can require different functional roles, these studies do not link conditions to specific functions, leaving the possibility that genes sensitive to many environments may belong in a generalized stress response category.

Given the extensive sets of gene annotations assembled by The Gene Ontology (Ashburner et al., 2000) for human and model organisms, counting annotations is a natural way to identify pleiotropic genes and has been employed in a number of studies. Khan et al. (2014) used semantic similarity of GO terms that clustered into non-overlapping functions to identify moonlighting proteins, a strict but particularly interesting type of pleiotropy in which functions are physically separable but not as a result of physical partitioning in the protein. The authors found that moonlighting proteins often (48% of the time) contain disordered regions. Pritykin et al. (2015) carefully considered the structure of the GO tree in order to identify genes with distinct functions. The

multifunctional genes were tested for associations with a number of gene properties, revealing that multifunctional genes are more likely to be large and multi-domain, essential, broadly expressed, central in PPI networks, have many regulators, and contain disordered regions.

4.2.2 A genome-wide and modular basis for pleiotropy

Biological networks can naturally represent modularity and pleiotropy, providing detailed molecular-level context for gene functions. In comparison to characterizations of individual genes, networks are more comprehensive representations of cell function because they reflect cellular processes as systems, which can be seen as, for example, associations between phenotypes and entire pathways (Kim and Przytycka, 2012; Vidal et al., 2011; Yu et al., 2016). Further, gene functions are not limited to associations with phenotypes: genes can affect network properties, such robustness and flexibility (Burga et al., 2011; Levy and Siegal, 2008; Park and Lehner, 2013; Rutherford and Lindquist, 1998). An estimation of pleiotropy as effects measured within a molecular-level network may therefore be crucial in order to capture a gene's importance to multiple components of a complex cellular system.

In protein interaction networks, a popular network-based characterization of hub proteins is classification as an intra-modular ("party") node, which mainly functions as part of a module and has correlated expression with its neighbors, or an inter-modular ("date") node, which coordinates between modules or has multiple functions (Agarwal et al., 2010; Han et al., 2004; Pritykin and Singh, 2013). A strength of physical protein interactions is that they are mechanistically interpretable; however this type of relationship is limited by physical locality. In contrast, genetic interactions identify a variety of functional relationships, including partial redundancy within the same module, pathway buffering, and dependency/similarity within a spatially or temporally directed pathway. We believe that genetic interactions provide a novel and valuable view of pleiotropy because they (i) are known to appear both within and between pathways, (ii) capture functional relationships between genes that operate in different high-level processes, (iii) can recover functions that are buffered by other genes, and (iv) can reflect any biological process, not just those in a restricted set of measured phenotypes. This last point is a solution to the problem of trait selection that many estimates of

pleiotropy are bound by. The GI network is therefore an informative place to assess the molecular pleiotropy of genes.

In this work using genetic interactions, we consider pleiotropy to be one gene affecting multiple sectors of cellular function such that there is a phenotypic consequence of fitness defect. With this definition, we are able to characterize the properties and behavior of genes that impinge on diverse functional modules and affect multiple traits at the molecular level.

The gene functional profiles described in Chapter 3 form the basis of the pleiotropy measure we describe in this study. In the creation of these profiles, frequent item set mining was used to exhaustively discover modules of genetic interactions, termed biclusters, which covered the majority of the yeast genetic interaction network (Costanzo et al., 2010). A bicluster is composed of two sets of genes and each gene in one set interacts with every gene in the other set—put another way, it is a complete bipartite subgraph of the GI network. Biclusters represent genes with similar behavior, because all genes on one side of a bicluster share a set of interaction partners; in this way, they are similar to clustered gene profiles, a popular framework for identifying functionally related genes. However, biclusters are built from subsets of a gene's interaction partners, meaning they can identify multiple functions per gene and thus represent reuse in addition to modularity. Bellay et al. (2011a) found that the bicluster coverage of interactions in a hub gene's profile may relate to pleiotropy, since this correlated with the number of GO terms annotated to the gene as well at the number of drug sensitivities (Hillenmeyer et al., 2008). In the following analysis, we describe a novel definition of pleiotropy derived from GI biclusters in a new, nearly complete yeast GI network (Costanzo et al., 2016). We measure pleiotropy using an entropy measure computed on the set of each gene's genetic interaction biclusters to describe the functional spread of its effects on phenotype. We evaluate characteristics of the high- and low-pleiotropy genes identified by our approach and report several physical, functional, and evolutionary properties that differ between the two pleiotropy classes.

4.3 Measuring pleiotropy from participation in GI modules

Genetic interactions are able to capture relationships between genes involved in different processes, and biclusters, which are groups of genes densely connected by

genetic interactions, are able to characterize the functional context of these relationships. We define a measure of pleiotropy that expresses a gene's functional distribution within these bicluster modules (Figure 4.1). Our first step was to apply XMOD to an input GI network to obtain its set of biclusters. For each gene, we collected all biclusters that contain it and removed clusters that were redundant (see Chapter 3). A bicluster consists of two sets of genes, densely connected by a set of genetic interactions bridging them. In the context of calculating pleiotropy for a specific gene *G*, we refer to the set containing *G* as the "associate side" and the set of genes interacting with *G* as the "adjacent side." We use simple criteria to annotate biclusters with high level biological processes: if the associate-side genes are statistically enriched for or are at least 40% composed of genes annotated by a term, then the bicluster is labeled with that term. Having identified and annotated a gene's biclusters, we then count the process annotations as described in Chapter 3, resulting in a functional profile of the gene's modules (Figure 4.1, 4.2A). The final pleiotropy score is the entropy of the process annotations counted in the profile (Figure 4.1, 4.2A and section A4.1). Entropy is a non-negative value that is 0 in the case that all annotations are the same and reaches a maximum when all possible annotations occur an equal number of times. The number of terms used for annotations, not the number of annotations a gene's biclusters receive, determines the maximum value entropy can reach. We used a set of 20 manually annotated (MA) biological processes (Table A3.2) and the entropy scores could range from zero to approximately 4.3 (Figure 4.2B).

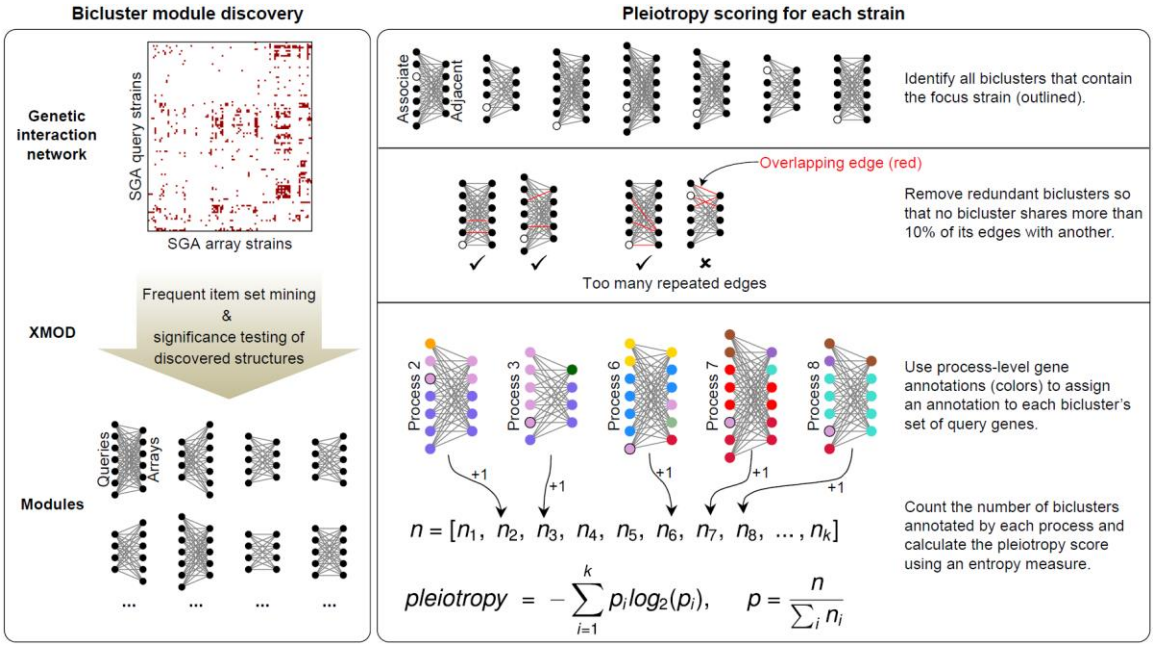


Figure 4.1. Measuring pleiotropy from GI modules. Bicluster modules are obtained by applying XMOD to the genetic interaction network **(Left box)**. The input SGA-derived network is binary and contains negative genetic interactions between query and array single mutant strains, reflecting SGA experimental setup. Interactions are added between query and array strains representing the same gene to allow modules containing these. Discovered complete bipartite modules have one set of query strains and one set of array strains. The pleiotropy of a focal strain, depicted as an outlined circle, is calculated from the functional distribution of its containing bicluster modules **(Right box)**. Bicluster annotations are determined by the associate side of the module, the set of genes that contains the focal gene and is drawn as the left side of each bicluster (the right strain set is the adjacent side). Colors represent gene annotations. The vector *n* contains counts of annotation occurrences.

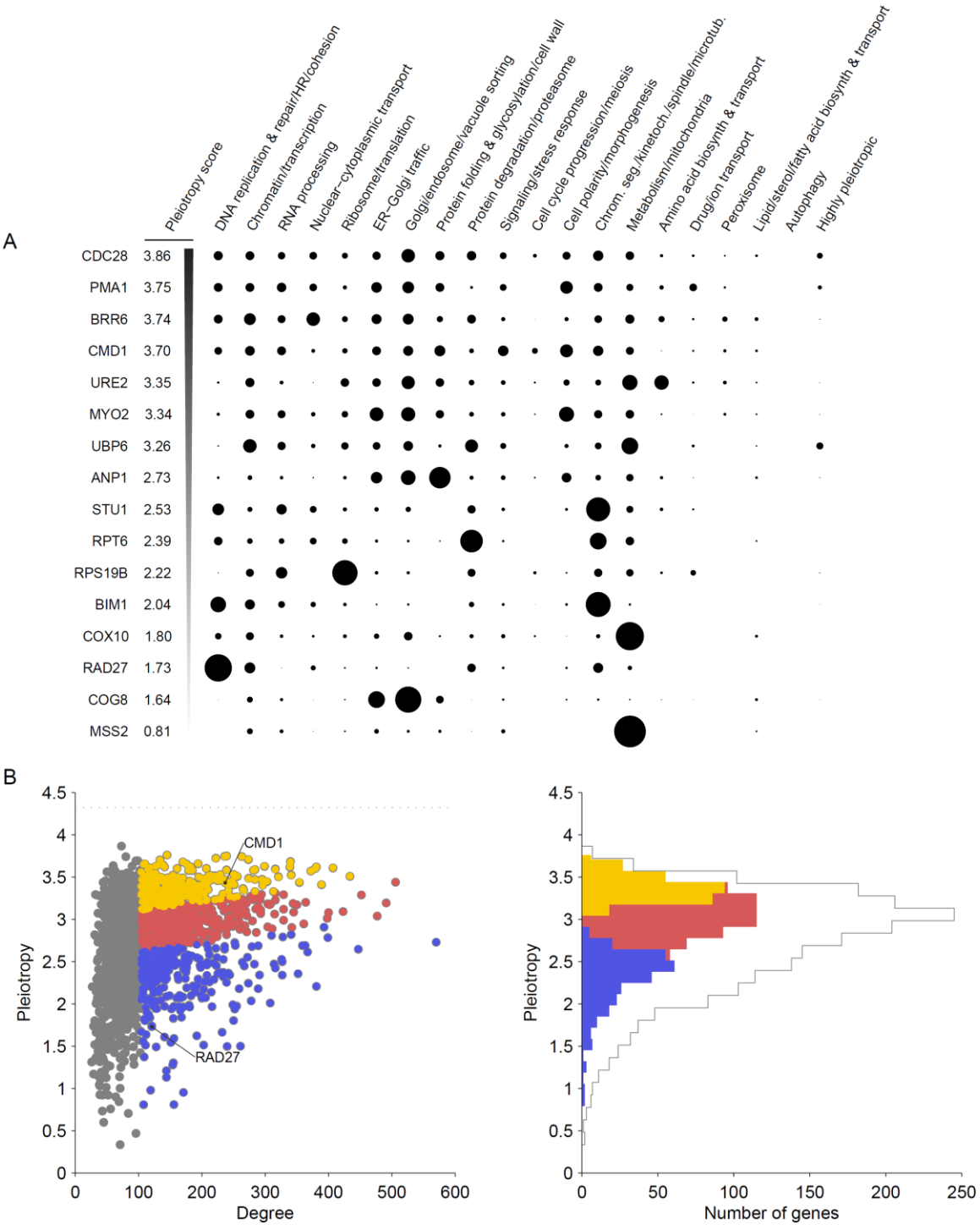


Figure 4.2. Pleiotropy scores. (A) Functional profiles of example genes with a range of pleiotropy scores are sorted with the most pleiotropic at the top. The distribution of annotation occurrences from each gene's containing modules was normalized (equal to vector *p* in Figure 4.1) and displayed such that circle area represents the fraction of

module annotations in each bioprocess. Data for each are from a single query strain. **(B)** Gene pleiotropy scores are correlated with genetic interaction degree, but still show further variation. High, low, and medium pleiotropy groups are only assigned to genes with degree of at least the 60^{th} percentile (vertical boundary between gray and colored markers, left plot) and are determined from residuals of the regression of pleiotropy scores against degree (cause for sloped divisions between the high-medium and medium-low boundaries, left plot). Histogram bars in the right-hand plot are stacked and count the genes assigned to pleiotropy groups. Both plots show DMA-derived pleiotropy scores.

In implementing this pleiotropy measure, we used negative genetic interactions of the latest, near-complete yeast GI network (Costanzo et al., 2016). This network comprises two distinct datasets reflecting the experimental organization used in its construction. The separation of the two GI networks persists throughout our work here: we derived pleiotropy scores from each. The first GI network, called the TSA (temperature sensitive array) network, contains 2112 query genes screened for interactions with 560 essential array genes and 173 nonessential array genes. The second, called the DMA (deletion mutant array) network, contains 4004 query genes screened for interactions with 3827 nonessential array genes. The query genes of both networks include nonessential genes, experimentally represented by gene deletions, and essential genes, which were represented by temperature sensitive and DAmP alleles. Accordingly, the biclusters from both networks can have a mixture of essential and nonessential genes on one side, the query side. We focused primarily on measuring pleiotropy for query genes (more precisely, strains), so in this case the associate-side module enrichment step of our pleiotropy method analyzed mixed-essentiality groups of genes. We also implemented our pleiotropy measure with a different data orientation, computing pleiotropy scores for array instead of query genes, and with a second annotation scheme, the experimentally derived SAFE annotations (Table A3.3) (Baryshnikova, 2016; Costanzo et al., 2016) instead of the manual set. We use the term “scoring configuration” to refer to a data setup used in generating pleiotropy scores, which specifies the annotation method and type of strains analyzed; in total, there are six configurations, which are described in section 4.9.1.

Pleiotropy scores systematically identified a broad range of gene functional behavior within the GI network. Figure 4.2A uses example genes to illustrate the relationship between pleiotropy scores and bicluster-derived functional profiles, which count biological process annotations: some genes participate in many processes, while others have functions concentrated in a few processes. In total, we were able to construct bicluster-derived functional profiles and measure pleiotropy for 3236 yeast genes.

When using the genetic interaction network, a straightforward pleiotropy metric could be the number of interactions observed for a given gene. A gene's genetic interaction degree is very informative about the magnitude of a mutation's effect. For example, degree is strongly correlated with fitness defect (Pearson's $r = 0.78$, $p < 10^{-300}$; (Costanzo et al., 2016, nonessential strains)), and is also correlated with the number of GO terms ($r = 0.23$, $p < 10^{-42}$) and the number of curated phenotypes ($r = 0.65$, $p < 10^{-300}$), two gene features that can indicate multiple functions. The pleiotropy score we developed, however, is more specific than GI degree—it is designed to distinguish different functions of a gene, first by organizing GIs into modules, and second, by assessing annotation profiles with fractions instead of counts in the entropy calculation. This decoupling of degree and pleiotropy is evident by the variation depicted in Figure 4.2B, which illustrates how a high degree alone is not sufficient for a gene to have high pleiotropy. Nevertheless, the Spearman correlation of 0.45 ($p < 10^{-53}$) between entropy and degree suggests that attempts to characterize pleiotropic genes may simply recover trends already associated with degree. To focus specifically on the functional breadth of genes, independently from their interaction degree, we controlled for GI degree when defining pleiotropy classes. Specifically, we first take pleiotropy as the residual of the regression of entropy against degree. We then limit genes to those that have a negative GI degree at or above the 50^{th} percentile. Finally, we classify these high-degree genes as high, medium, or low pleiotropy by binning scores into the highest 30%, middle 40%, and lowest 30% (Figure 4.2B). These three classes are used for all statistical analyses discussed in the following sections. As previously mentioned, we used the TS-derived and DMA-derived GI networks separately in measuring pleiotropy, therefore we have a set of three pleiotropy classes for each network. We specify source GI data in the text when discussing specific results.

4.4 Examples of high and low pleiotropy: Calmodulin and RAD27

As an example, we highlight the high-scoring pleiotropic gene CMD1 (Figure 4.3A, Figure 4.2A), which encodes the binding protein calmodulin and is conserved in all eukaryotes. It is well known to regulate many processes, a functional ubiquity that likely is enabled mechanistically by the capacity to bind calcium ions in four different sites in most species as well as bind various target proteins, many of which trigger function-specific conformations of calmodulin. Evidence of binding site functional specificity comes from Ohya and Botstein (1994), who found four groups of mutations that resulted in distinct phenotypes. Using its namesake ability to detect Ca+ ions, CMD1 activates calcineurin and two protein kinases when Ca2+ concentration is high, which control a number of downstream processes (Cyert, 2001). Within the GI network, CMD1 appears in dozens of biclusters. Nine of them are shown in Figure 4.3A to illustrate both how GI-derived pleiotropy is apparent from structured modules and the functional coherency that characterizes these modules. One of Calmodulin's known localizations is the bud tip and neck due to its physical interaction with Myo2p, a myosin protein that is required for polarized growth (Stevens and Davis, 1998). This relationship is reflected in the bicluster labeled "Cell polarity/morphogenesis" (Figure A4.1), which contains cell wall integrity genes SLT2, BCK1, and SWI4, bud neck and wall localized proteins SKT5, CHS3 (recruited by SKT5 and MYO2), and ROM2, ARP2/3 activator PAN1, and polarity-establishing BEM1 (Duncan et al., 2001; Levin, 2005; Madden and Snyder, 1998). Another established localization behavior of calmodulin is association with the spindle pole body throughout the cell cycle. During mitosis, it is involved in attachment of microtubules to the SPB and is required for correct spindle function (Sundberg et al., 1996). This explains its membership in the bicluster labeled "Chrom. seg/kinetoch./etc" (Figure A4.2) along with spindle organizers CIK1 and STU2 (a SPB-interactor), as well as a number of kinetochore genes, AME1, OKP1, and NSL1, and kinetochore recruitment gene CTF13 (De Wulf et al., 2003; Kosco et al., 2001; Page et al., 1994). The shared negative interactors of these genes, adjacent in the bicluster, are genes from the spindle assembly checkpoint (SAC), which can buffer a dysfunctional spindle by prolonging prometaphase. Lastly, Cmd1p is thought to regulate the final stages of vacuolar fusion (Peters and Mayer, 1998). The bicluster labeled "Golgi/endosome/vacuole" (Figure A4.3) reflects this role, containing two components of the cytoplasm-to-vacuole targeting pathway complex TRAPPIII and GYP1, which

respectively activate and deactivate the vesicle docking regulator YPT1, as well as SEC17, which is required before vacuole membrane fusion events, and three members of the COG complex (Du and Novick, 2001; Lynch-Day et al., 2010; Ungermann et al., 1998). Though GI modules do not explain specific functions of a gene, the example of CMD1 shows how genetic interactions can recover evidence for functions established in literature.

In contrast to the highly varied functions of CMD1, we also make an example of RAD27 (Figure 4.3B). This gene has a focused functional influence on cellular processes, and therefore low pleiotropy, with nearly all of its containing biclusters representing DNA replication and repair functions. Despite the clear theme of RAD27's modules, we still see individual pathways clustering together. For example, the associate side of one bicluster (Figure A4.4) contains genes in complexes that initiate and drive the replication fork during DNA replication (Medagli et al., 2016). The genes PSF1 and SLD5, as half of the GINs complex, and SLD3 help to assemble the pre-initiation complex, which includes MCM2, ORC2, and CDC45, at replication origin sites. Many of these genes go on to form the CMG complex, the helicase that unwinds duplex DNA and progresses in the core of the replication fork. This set of genes negatively interacts with genes involved mitotic checkpoints for DNA damage and DNA replication, MRC1, RAD9, RAD24, DDC1, RAD17, and CSM3, which appear in the bicluster's adjacent side. Another of RAD27's biclusters contains histone-related genes in both sides (e.g. SWC3, SWR1, ARP6, VPS71, HTZ1, YTA7, and EAF6), and others contain a number of genes related to RAD27's known functions, Okazaki fragment processing and double-strand break repair (Figure A4.5) (e.g. POL31, RNH203, RNH201, XRS2, MRE11, and RAD50).

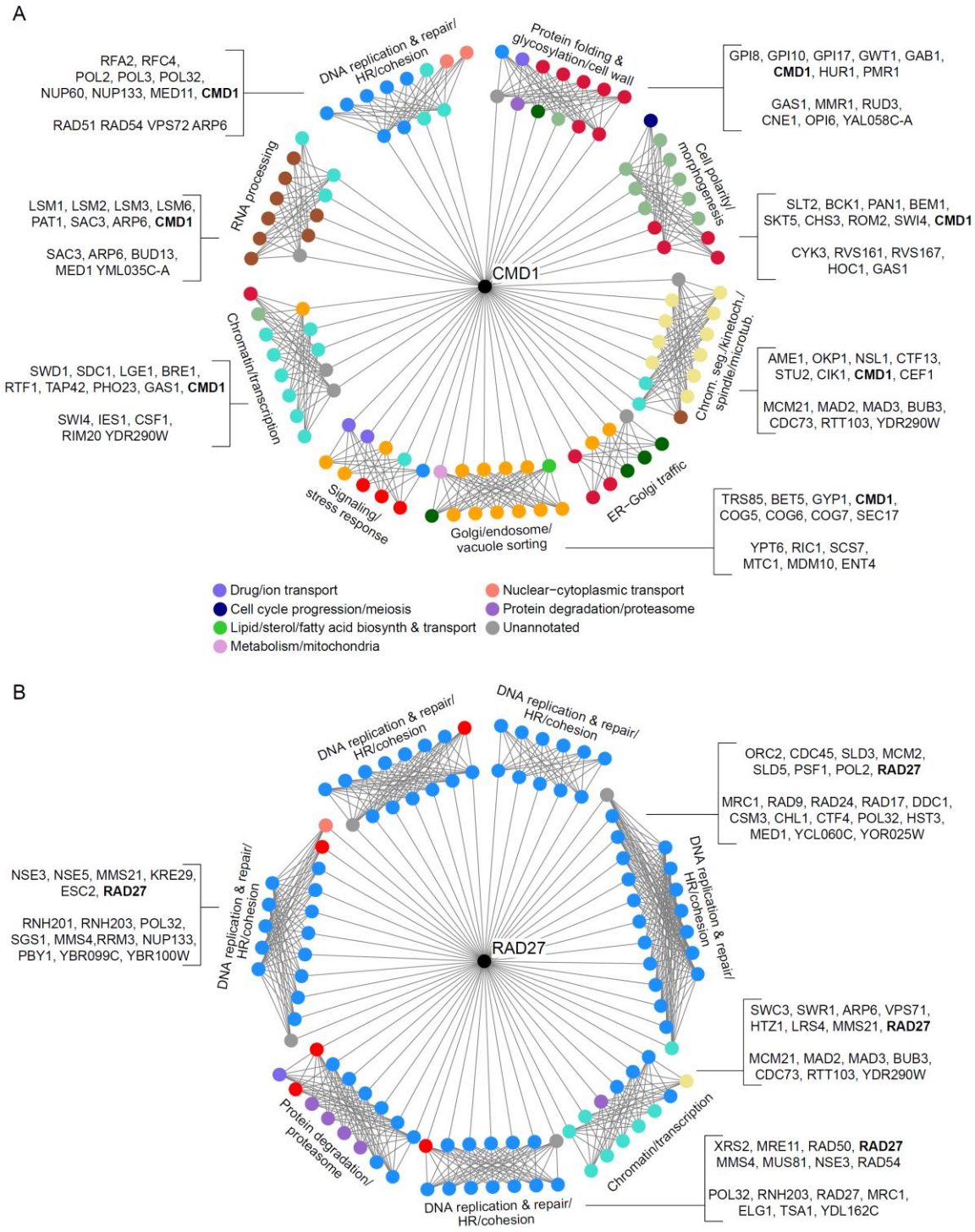


Figure 4.3. Selected biclusters of the pleiotropic gene CMD1 (A) and the non-pleiotropic gene RAD27. (B). Nodes represent genes and edges represent negative genetic interactions extracted from the DMA-derived GI network. Only genetic interactions that define each bicluster are displayed, although there are often

interactions between genes on the same side of a bicluster. The biclusters' adjacent-side sets of genes are connected to the focal genes CMD1 or RAD27. The genes arranged on the outside of the each diagram, with the addition of the focal gene, are the associate sides. Gene names list the members of some biclusters; the first group of names in each bracketed pair lists the associate-side genes and the second group lists the adjacent-side genes. Text labels are bicluster annotations determined from the associate-side genes. Colors of nodes indicate the functional annotations of genes, which can be inferred from the bicluster annotations (*e.g.* sea green represents "Cell polarity/morphogenesis"). Any colors that cannot be interpreted with bicluster labels are listed in the legend. Any genes that have multiple process annotations are colored preferentially to match the annotation given to the bicluster. Both panels use the same color scheme.

4.5 Many primary functions are represented in high-pleiotropy genes

Many of the genes that displayed high pleiotropy have known associations with particular pathways. The chaperone HSP90, whose pleiotropy score is in the highest 30%, is a classic example of how participation in a central maintenance pathway allows the gene to suppress phenotypic variation in many aspects of cellular biology (Rutherford and Lindquist, 1998). We found that this is not unique; genes involved in many other cellular functions also exhibited high pleiotropy. The following are brief examples of some of the many functional annotations already associated with genes in our high pleiotropy class: cell cycle regulation (CDC28, CKS1, cyclin CLN3, whole genome duplicates SWI5 and ACE2, RAM pathway component TAO3); the ubiquitin system (UBI4, UBP1, DOA1, CDC53, RAD6, RSP5, TOM1, UBP6, UBR2, HRT1, UFD1, UBP14); stress response and protein folding (chaperones HSP82, CDC37, and CNS1, HSP82 regulator HSP1); membership in the CCR4-NOT complex, a global transcription regulator (CDC36, CDC39, NOT3, CAF120); ribosome biogenesis (MAK11, NEW1, DBP7); nuclear-envelope membrane functions (BRL1, BRR6, and APQ12); and vacuole functions (VPS62, VAC7, VAC14, VPS66, ZRT3, IML1).

4.6 GO term enrichment within pleiotropy classes

In order to discover any particular cellular processes or components that are significantly biased in their composition of pleiotropic genes, we performed hypergeometric tests for enrichment of GO terms in our high and low pleiotropy classes. We found that high pleiotropy genes were not enriched for any GO terms. Although this result is somewhat surprising, it is consistent with the observation that pleiotropic genes work in diverse primary functions. Low-pleiotropy gene classes from both GI networks were enriched for a number of terms. Low-pleiotropy genes derived from the DMA network were enriched for Golgi vesicle-mediated transport, as well as more general transport and localization terms, and mitochondrial respiration. For example, 43 of all 55 background genes annotated by the GO component “mitochondrial inner membrane” have low pleiotropy (enrichment, $p < 10^{-11}$). The low-pleiotropy class derived from the TSA network was enriched for vesicle transport also, and DNA replication and proteolysis terms. For example, 15 of the 16 genes in the cytosolic proteasome complex have low pleiotropy (enrichment, $p < 10^{-4}$).

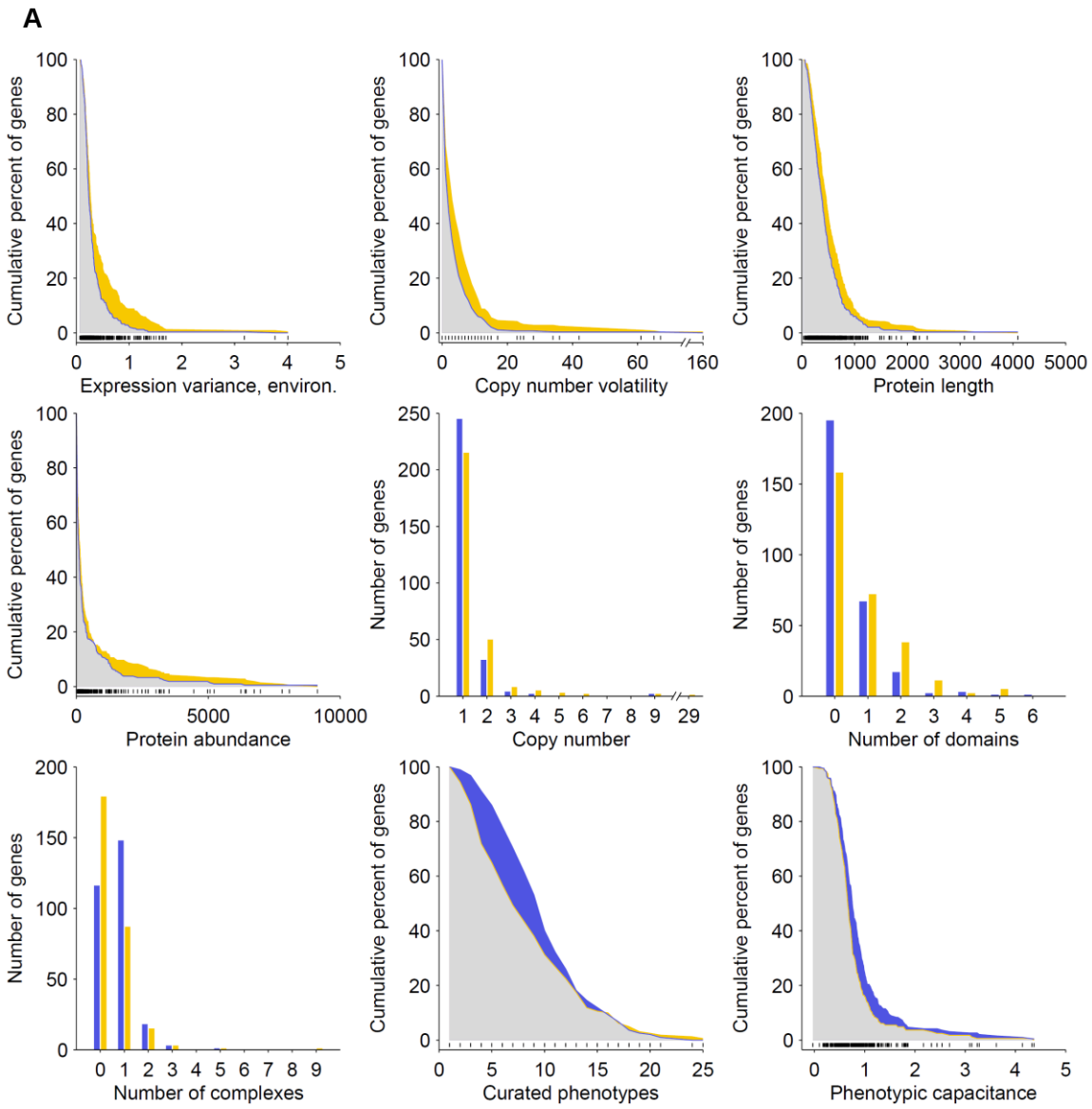
4.7 Differences between high- and low-pleiotropy genes

Next, we searched for evolutionary, structural, and functional properties of high- and low-pleiotropy genes by testing for associations with 36 gene characteristics. Many of the gene characteristics were described in Chapter 1, section 1.3.2; the exact set used and their associated methods are listed in Appendix 4, section A4.2. Briefly, the characteristics include quantitative summaries of individual gene behavior in various phylogenetic analyses, sequence-based calculations, and genome-wide experimental data sets.

A number of gene properties differed significantly between the two pleiotropy classes in Wilcoxon rank-sum tests (Figure 4.4, Table 4.1). High-pleiotropy genes were positively associated with expression variation, high gene copy-number-based features, high protein abundance, and many domains, while the low-pleiotropy genes tended to participate in protein complexes and, surprisingly, had more curated phenotypes. Specific statistics presented in the sections below are based on pleiotropy scores of query strains, our default analysis set, derived from both the TSA and DMA GI networks; we also explored other scoring configurations (section 4.9.1). In most cases, pleiotropy

associations with gene characteristics were consistent across different scoring configurations, with few exceptions. For example, SMF, essentiality, and Expression variance, genetic-B were significantly associated with both high- and low-pleiotropy genes, depending on the scoring configuration used (Table 4.1).

Our reporting of results in the following sections is conservative: we tested 22 variations of our method (section 4.9.2 and Table 4.2) and report results that are robust across most test variants for multiple scoring configurations (Table 4.1; non-robust results, Table A4.2). For example, SGA interrogates essential genes with mutant strains that are temperature-sensitive point mutations or DAmP (low expression) alleles. Because it is easy to imagine a point mutation that affects only the subset of a gene's functions that is dependent on a single part of the protein, one variation of our rank-sum tests excludes TS strains, leaving just DAmP alleles to represent the behavior of essential genes.


A
Cumulative percent of genes
Expression variance, environ.
Copy number volatility
Protein length
Protein abundance
Number of genes
Copy number
Number of domains
Number of complexes
Curated phenotypes
Phenotypic capacitance

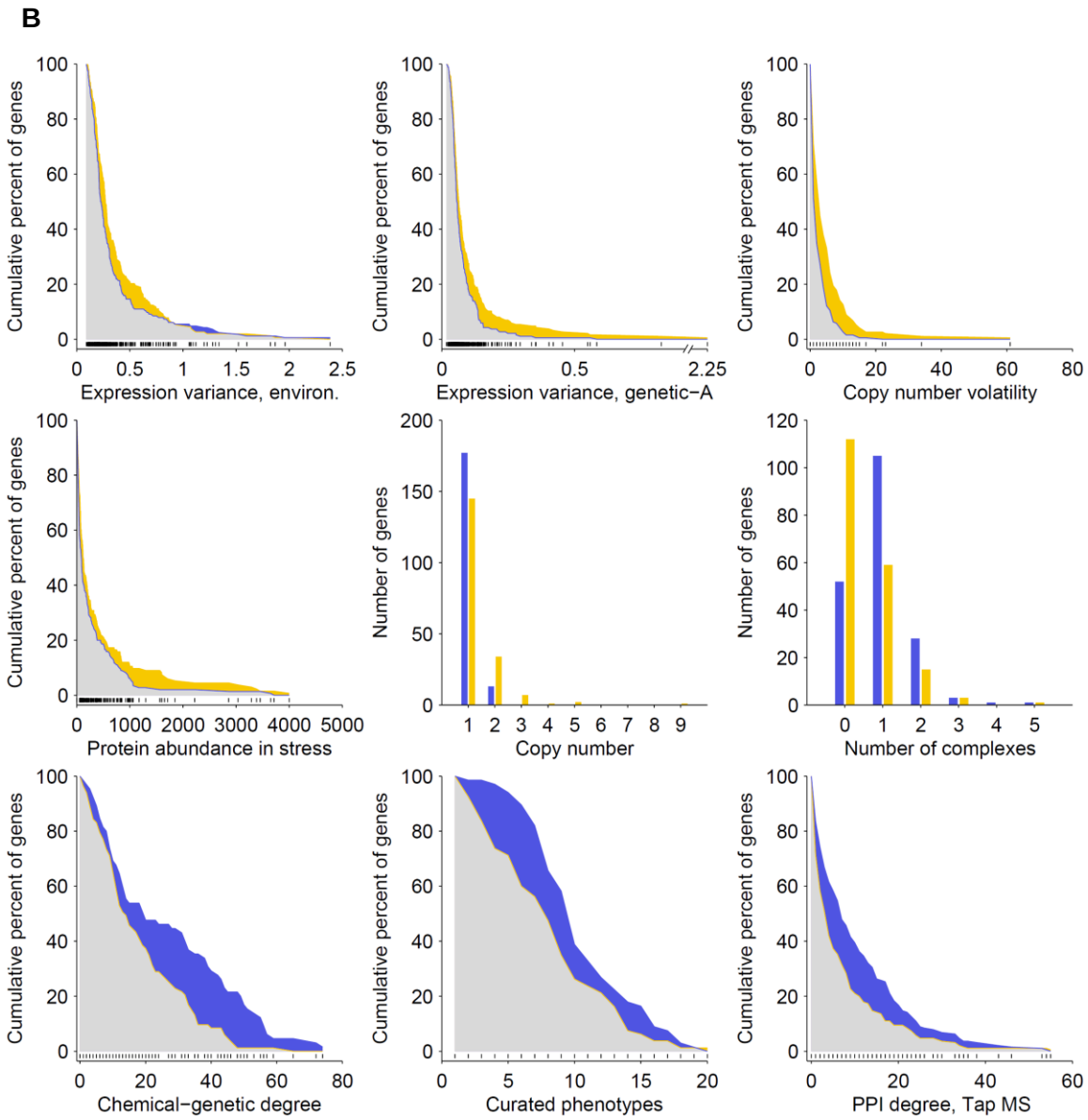


Figure 4.4. Gene properties significantly associated with high (yellow) or low (blue) pleiotropy derived from the DMA (A) and TSA (B) GI networks. Cumulative plots are displayed for properties that take on many values. For a pleiotropy group and property, the plotted line shows the percent of the genes that have a property value greater than or equal to any point on the x-axis. Percentage calculations only take into account genes that have measured values for the property (all characteristics shown had good data coverage, see section 4.9). The area between the blue and yellow lines is filled with color indicating which pleiotropy group has a higher percent of genes with high

property values. Black hash marks plotted above the x-axis mark all values found in the genes' property values. Bar plots are displayed for properties that take on few values. There is a total of 268 genes in both the high and low DMA-derived groups **(A)** and a total of 163 genes in both TSA-derived groups **(B)**. P-values from rank-sum tests for each property from left to right are (A, top row) 7×10^{-3}, 8.4×10^{-4}, 7.4×10^{-4}; (A, second row) 3.4×10^{-2}, 4.3×10^{-4}, 1.1×10^{-4}; (A, third row) 1.1×10^{-5}, 2.7×10^{-4}, 1.6×10^{-3} (B, top row) 5.4×10^{-3}, 6.4×10^{-3}, 2.4×10^{-6}; (B, second row) 3.4×10^{-2}; 8.7×10^{-6}, 7.5×10^{-7}; (B, third row) 5.8×10^{-3}, 2.1×10^{-3}, 5.9×10^{-5}.

Table 4.1. Summary of gene characteristics associated with high- and low-pleiotropy genes. Tests were performed for pleiotropy scores derived from different pleiotropy scoring configurations (columns); TSA array configurations are relegated to Table A4.1 due to sparsity of significant results. Values shown are the number of rank-sum tests that yielded a significant p-value, out of a total of 22 variations performed for each query-strain scoring configuration and 12 variations performed for each array-strain scoring configuration (see section 4.9 and Table 4.2). Blank cells indicate zero tests with significant results. Values in parentheses indicate significant results that contradict the result column by associating the gene property with the opposite pleiotropy class. Asterisks indicate features that were associated strongly enough with both pleiotropy classes that the property is listed in two rows. The significance of p-values from rank-sum tests was determined using the FDR-control procedure described in Benjamini et al. (2006), counting tests for 37 gene properties as a family.

Gene/protein properties	Pleiotropy class with positive association	DMA						TSA		
		Query			Array			Query		
		Associate		Adjacent	Associate		Adjacent	Associate		Adjacent
		MA	SAFE	MA	MA	SAFE	MA	MA	SAFE	MA
Copy number volatility	High	22	18	10	12	10		22	22	18
Protein abundance in stress	High		18	22	8	12		15	22	22
Copy number	High	21	18			8	2	22	22	22
WGD duplicate	High	21	20			8		21	22	19
Expression var., environ.	High	10	17	6	1	10	5	17	21	22
Protein abundance	High	3	20	22	12	12	2	1	11	22
Expression var., genetic-A	High	2	15			12	2	19	22	18
CAI	High	1		20		11	2	3	22	18
Transcription level	High			18	5	12	2	1	12	20
Expression level	High	1		20	5	7	2	1	5	21
Protein length	High	22	21	20	2					
Number of domains	High	21	22	13					3	
Number of unique domains	High	21	22	12					3	
Single mutant fitness defect*	High			12		7	8			1
Originated in Saccharomyces	High			4				22	21	11
Essential*	High	12	2	20						
Expression var., genetic-B*	High								21	1
Complex member	Low	22	22		12	8		22	22	22
Number of complexes	Low	22	22		12	8		22	22	21
Curated phenotypes	Low	22	22	11	6	5	6	22	22	11
Multifunctionality	Low		4	1	11	4		21	22	20
Phenotypic capacitance	Low	11	22			5		5	22	13
PPI degree, Tap MS	Low			13	3	1		22	19	13
Single mutant fitness defect*	Low	22	22		6			10	4	
Chemical-genetic degree	Low					10		9	22	17
Effective number of codons	Low			8	8	12	2		7	21
Essential*	Low							14	12	9
Yeast conservation	Low		2	15	5			5	22	
dN/dS	Low		1	9	1	10	3			3
Broad conservation	Low			8				15	10	
Gene age	Low			5				18	5	
Expression var., genetic-B*	Low			20						

4.7.1 Expression variance and protein abundance are higher among high-pleiotropy genes

Two different measurements of gene expression level variance were robustly associated with high-pleiotropy genes. Environmental expression variance is determined by subjecting yeast cells to many environments and measuring gene expression levels, then calculating variance for each gene (Gasch et al., 2000). A Wilcoxon rank-sum test showed that genes in the high pleiotropy class had higher environmental expression variance than genes in the low-pleiotropy class using both the DMA ($p < 7 \times 10^{-3}$) and TSA ($p < 6 \times 10^{-3}$) pleiotropy scores (Expression variance, environ., Figure 4.4A,B). Among the 50 genes with the highest environmental expression variance, 22 have high DMA-derived pleiotropy compared to only 5 with low pleiotropy (Figure 4.5A; 23 have a

medium pleiotropy). We found that, regardless of pleiotropy level, most genes with high variance reached their extreme expression levels during heat shock and cold shock conditions and during stationary phase. Regulatory response to environmental stresses consists of induced expression of some genes and suppression of others, a program that is similar in all stress environments, not condition-specific (Gasch et al., 2000). We find that there is no bias in the high pleiotropy genes towards having increased or decreased expression during stress conditions (Figure 4.5B).

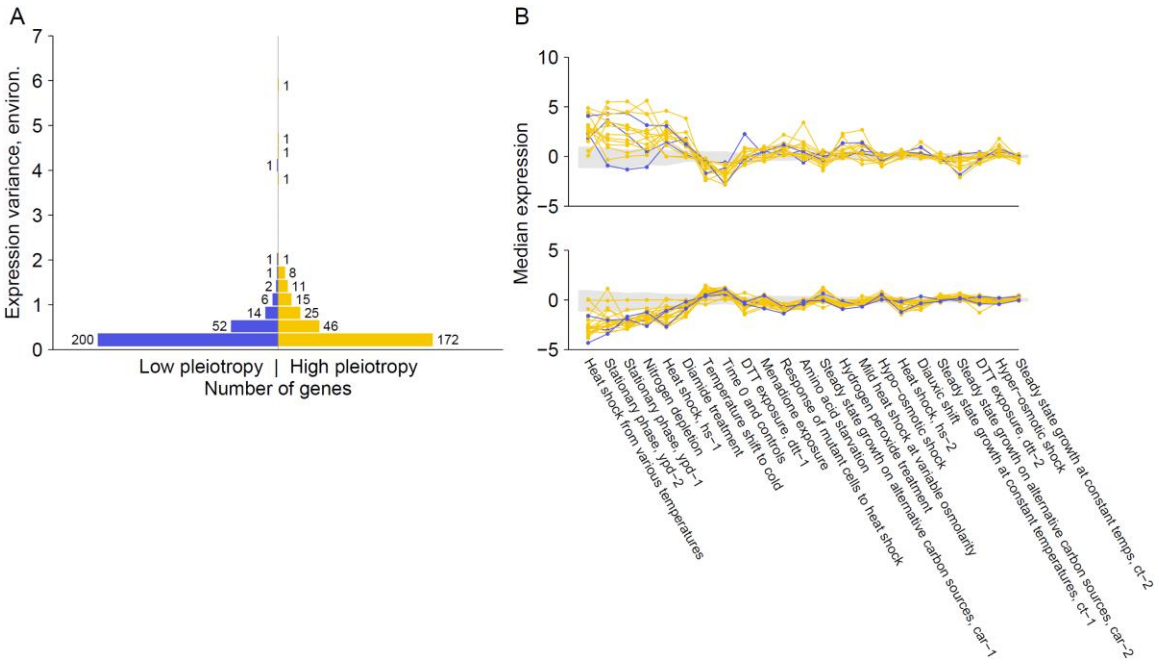


Figure 4.5. High-pleiotropy genes have higher environmental expression variance than low-pleiotropy genes. (A) Histograms show the distribution of expression variance, shown on the y-axis, for all high- and low-entropy genes. The number of genes in each pleiotropy class that fall into each bin is given next to each bar. **(B)** Expression of all high- and low-pleiotropy genes that are among the 60 genes (regardless of pleiotropy) with the highest expression variance was plotted for each environment. Values plotted are medians of multiple time points measured for individual environments. Separate axes are for visual clarity only: genes that tend to increase in expression during stress are plotted on the top, while genes with the opposite trend are on the bottom. Yellow indicates high-pleiotropy and blue indicates low pleiotropy.

Genetic expression variance, calculated from the genome-wide expression profiles of many crosses between the diverged *S. cerevisiae* strains BY and RM (Brem and Kruglyak, 2005)(Expression variance, genetic-A, Figure 4.4B), was also associated with high pleiotropy genes identified from both GI networks (DMA-derived pleiotropy, SAFE annotations: $p < 7 \times 10^{-3}$; TSA-derived pleiotropy: $p < 7 \times 10^{-3}$), although this result depended on the configuration used for the DMA-derived pleiotropy classes (Table 4.1). Among the 50 genes with the highest variance, 22 had high pleiotropy and 7 had low pleiotropy (21 have medium pleiotropy). Similarly, a second measure of genetic expression variance (Skelly et al., 2013)(Expression variance, genetic-B, Table 4.1), which is measured from the gene expression of many diverged and geographically varied *S. cerevisiae* strains, was associated with high pleiotropy genes for one scoring configuration. (For another scoring configuration, in which pleiotropy of DMA queries was measured using adjacent-side bicluster enrichments, this expression variance feature was associated with low-pleiotropy genes.)

The two expression variance measures strongly associated with high-pleiotropy genes, environmental and genetic-A, had a Pearson's correlation of 0.21 ($p < 2.6 \times 10^{-13}$), suggesting that highly variable genes defined by the two measures overlap. However, environmental variance remained robustly associated with pleiotropy after controlling for the genetic-based feature (TSA-derived pleiotropy, rank-sum $p < 0.017$). Genetic expression variance was not correlated after controlling for the environmental feature, suggesting that environment-induced expression variation is more strongly linked with high pleiotropy.

Protein abundance levels (Newman et al., 2006) offer further observation of cellular usage of a gene, since translation and protein degradation are regulated. Protein abundance, including protein abundance under stress conditions, tended to be higher in high-pleiotropy genes than in low-pleiotropy genes (nonstress, SAFE: $p < 7 \times 10^{-3}$; stress, SAFE: $p < 5 \times 10^{-3}$; Figure 4.4).

4.7.2 Copy number is higher in high-pleiotropy genes

High-pleiotropy genes tended to have higher copy number, which is the number of genes that arose through duplication of a single ancestral gene of a given gene, compared to low-pleiotropy genes (DMA: $p < 4.3 \times 10^{-4}$; TSA: $p < 8.7 \times 10^{-6}$; Figure 4.4A,

B). High-pleiotropy genes with a copy number greater than two are in protein families that function in environmental responses as transmembrane proteins or components of signaling pathways, consistent with previous characterization of genes that have frequently duplicated (Wapinski et al., 2007b). The most extreme copy number is that of high-pleiotropy gene RGT2, which, with its paralogs, is in a family of transmembrane sugar-transport channels, including some that trigger response to intracellular sugar concentrations. A more well-known example is the hub IRA2, which is a negative regulator of RAS2 and has two paralogs.

Duplicate gene pairs that arose from an ancient *S. cerevisiae* whole-genome duplication (WGD) event are distinguished from all other duplicate pairs, which resulted from small-scale duplication (SSD) events (Guan et al., 2007; Hakes et al., 2007; Kellis et al., 2004; Wapinski et al., 2007b; Wong et al., 2002). We found that this difference is important with respect to pleiotropy. WGD genes were strongly associated with high pleiotropy genes, while SSD genes had only slight evidence of an association (Table 4.1, Table A4.2). This difference is difficult to explain, since there is no consensus on how evolutionary models apply differently to these scenarios. However, it is possible the genome state after a whole-genome duplication helps genes diversify by providing broader redundancy, like entirely duplicated complexes and pathways.

The DMA-derived group of high pleiotropy genes contained 38 WGD genes, significantly more than the 18 classified as low pleiotropy ($p < 4.8 \times 10^{-3}$; Figure 4.6). TSA-derived groups, from an overall smaller dataset, showed the same trend with 18 and 4 WGD duplicates in the high and low groups, respectively.

WGD gene pairs have been shown to sometimes have unequal allocation of importance, though they typically retain similar function (Kellis et al., 2004; VanderSluis et al., 2010), and pleiotropy roles reflect both these scenarios. We investigated the behavior of the duplicate partners of the DMA-derived high pleiotropy genes (most partners of genes in the TSA-derived pleiotropy groups have not been screened in SGA). First, considering only the handful of WGD pairs whose members both meet our degree criteria and therefore both have assigned pleiotropy classes, we find similarity in pleiotropy (Figure 4.6). Two WGD pairs were composed of two high-pleiotropy genes each (the pair ACE2 and SWI5, and the pair RPL40A and RPL40B) and, similarly, one pair had two low-pleiotropy members. However, no paired genes that both had high degree contained a low- and high-pleiotropy gene—a hint that duplicate partners of high

pleiotropy genes tend to have higher pleiotropy than partners of low-pleiotropy genes (Wilcoxon rank-sum $p < 0.016$). Beyond these cases, the duplicate partners of high-pleiotropy WGD genes ranged broadly in both pleiotropy and GI degree (Figure 6), therefore representing both similarity and difference in the pleiotropy level of paired duplicate genes (Figure 4.6).

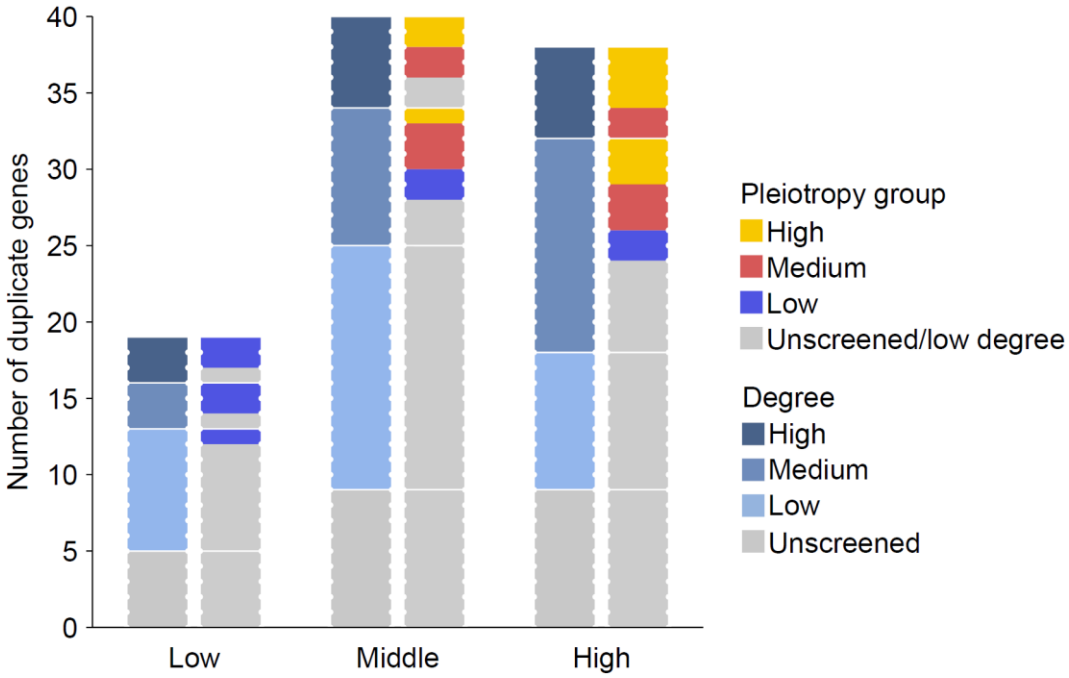


Figure 4.6. High-entropy genes are more likely to be whole-genome duplicates than are low-entropy genes. Bar heights show the number of whole-genome duplicate genes that are in the classes of high-, medium-, and low-pleiotropy, as labeled on the x-axis. GI degree and pleiotropy data shown as bar coloring describes the WGD *partners* of the high-degree classified genes. Bars on the left side show partner GI degree, where genes are considered "high" if their degree is at least the 50 percentile "hub" threshold used to define pleiotropy classes (near 100); "medium" if degree is at least 50 but lower than high cutoff; and "low" if degree is lower than 50. Bars on the right side show partner pleiotropy scores, which use the same thresholds as the standard pleiotropy classes defined for high-degree genes. Stacked sections of pleiotropy scores correspond to the matching sections of degree, as shown by horizontal white lines. For example, of all the WGD partners of classified low-pleiotropy genes, there are six with low GI degree (light blue, left bar) and, of these, one participated in enough biclusters that it could be given

and pleiotropy score, which was low (one unit of dark blue, right bar). As another example, there are six high-pleiotropy genes whose WGD partners have high degree and, of these, four also have high pleiotropy. High-degree genes will be counted as both a classified gene and a partner of a classified gene.

A third copy-number based feature, copy number volatility (Wapinski et al., 2007b), was also higher in high pleiotropy genes ($p < 8.4 \times 10^{-4}$, $p < 2.4 \times 10^{-6}$, Figure 4.4A, B). This property measures the number of times a gene is lost or duplicated within extant or ancestral yeast species. We note that although copy number and copy number volatility are correlated with the binary WGD duplicate feature, these features each remain significantly associated with high pleiotropy genes after controlling for WGD duplication ($p < 0.032$ and < 0.011, respectively).

4.7.3 Domains are more common in high-pleiotropy genes

The proteins of high pleiotropy genes tended to have more domains than those of low pleiotropy genes (DMA $p < 1.1 \times 10^{-4}$; Figure 4.4A), a relationship supported by a recent GO-based measure of multifunctionality (Pritykin et al., 2015). Speculating that the association between the number of domains and pleiotropy is driven by functions of individual domains, we tested for enrichment of specific domains and combinations of domains, but did not find significant results for either medium- or high-pleiotropy genes.

4.7.4 Characteristics of low-pleiotropy genes

Genes that have low pleiotropy are characterized as highly prominent genes that are well-studied, conserved, and important. We found that low-pleiotropy genes are involved in more complexes than high pleiotropy genes (DMA $p < 1.1 \times 10^{-5}$, TSA $p < 7.5 \times 10^{-7}$, Figure 4.4A, B), a characteristic derived from a literature-curated protein complex standard (Costanzo et al., 2016). For TSA-derived pleiotropy, this result is also supported by a tendency to have a higher number of protein interactions in Tap-MS experiments ($p < 6 \times 10^{-5}$) (Table 4.1). Participation of low-pleiotropy genes in protein complexes likely has the result of constraining the evolution of these genes (Lovell and Robertson, 2010).

A second characteristic of low-pleiotropy genes is that, compared to the high-pleiotropy genes, they have higher phenotypic capacitance, which is a measure of average morphological variance upon deletion of a nonessential gene and therefore also an indication of ability to buffer variability in phenotypes (Levy and Siegal, 2008). The authors who investigated phenotypic capacitors described a subset of capacitors that function in protein interaction clusters containing multiple capacitors and have a number of specific GO enrichments. This suggests that some capacitors promote phenotypic robustness by working in specific pathways. The abundance of these capacitors in our low-pleiotropy shows that our process of measuring a gene's functional behavior through GI modules has distinguished between genes with specific roles in varied pathways (high pleiotropy) and genes whose deletion effects, but not necessarily wild-type behavior, has a variety of phenotypes.

Low pleiotropy genes have a higher number of annotations in the form of curated phenotypes (DMA $p < 2.8 \times 10^{-4}$, TSA $p < 2.1 \times 10^{-3}$) and multifunctionality (TSA $p < 8.6 \times 10^{-4}$), which is derived from GO biological process annotations. These results are difficult to explain in relation to low pleiotropy, but might be caused by investigation bias in our understanding of yeast. Indeed, for the TSA-derived pleiotropy groups, we find tendencies of the low-pleiotropy genes to be old and conserved and to have strong mutant phenotypes (Table 4.1, "Yeast conservation", "Broad conservation", "Age", "Single mutant fitness defect"), which describe genes that may be frequently studied. The low-pleiotropy genes from the DMA-derived scoring configuration "query, adjacent" may also trend in the conserved direction, but not robustly, with only two rank-sum test variations displaying significance. In order to test the possibility that curated phenotypes and high multifunctionality specifically highlight well-studied genes, we compared these two gene characteristics to the number of publications associated with individual genes, which we obtained from PubMed ("Links from Gene") and SGD ("primary references"). Curated phenotypes had significant correlations of 0.49 and 0.43 with PubMed and SGD literature counts, respectively; multifunctionality had significant correlations of 0.67 and 0.65. We therefore calculated new annotation-count characteristics by regressing curated phenotypes and multifunctionality against each of the literature counts and recording the residuals of the two characteristics. Next, we repeated the rank-sum tests to compare the literature-controlled annotation characteristics between the high- and low-pleiotropy classes. Despite accounting for possible literature bias, the association

between DMA-derived low pleiotropy and curated phenotypes remained. However, using the annotation characteristics that controlled for PubMed literature counts completely removed the associations between TSA-derived low pleiotropy and both curated phenotypes and multifunctionality (rank-sum $p > 0.56$ and 0.18). This means that we cannot exclude literature bias as a major driver of the original surprising results. Importantly, there is a reasonable expectation that literature counts strongly reflect truly interesting genes, so these results are far from conclusive.

Given the fact that a control for literature bias did not alter the association between DMA-derived low pleiotropy and curated phenotypes, and that we expect curated annotations to accurately reflect biology even with some bias, our low pleiotropy results remain puzzling. It is possible that our measure of pleiotropy is only evident at the molecular level or may depend on the functional depth that genetic interactions reveal by removing redundancy in pairs of genes that have hidden effects. This suggests that our novel, unbiased measure of pleiotropy captures an as-yet unappreciated amount of functional influence that flourishes in many functionally buffered and newly evolved genes that are difficult to characterize experimentally.

4.8 Discussion

Genetic interactions provide a valuable view of pleiotropy by revealing gene functions at a molecular level. Clusters of within-pathway interactions highlight modules of genes related to specific cellular processes, like pathways or protein complexes, while between-pathway interactions occur when two pathways buffer each other. With this sensitivity to such a variety of gene-gene relationships, genetic interactions are well-suited for identifying diverse functions. Importantly, genetic interactions are calculated solely from phenotypic measurements, namely growth rates, in our case. Therefore, all genetic interactions represent functions that are evolutionarily relevant. Despite the fact that only one phenotype is measured, the functions represented by genetic interactions span most aspects of cellular biology (Costanzo et al., 2010; Costanzo et al., 2016).

Another strength of genetic interactions is their ability to reveal functions that may be undetectable in single mutants. A negative genetic interaction between two genes indicates a shared function that either one can perform individually, i.e. a buffering relationship. By assessing a gene's pleiotropy within the GI network, we leverage the

context of many (individual) background mutations, effectively removing layers of buffering and exposing the gene's formerly hidden phenotypes. Some pleiotropy studies suggest that most genes affect few traits (as reviewed by (Paaby and Rockman, 2013; Wagner and Zhang, 2011)), but none of the considered datasets measure gene roles that are normally buffered in single mutants, leaving both theoretical and empirical discussions (Wang et al., 2010) to possibly underestimate pleiotropy. Still, the importance of recognizing buffered functions depends on the extent to which individuals in natural populations harbor genetic variations that have genetic interactions.

A key element or our pleiotropy measure is the organization of the GI network into biclusters, which has multiple benefits. First, we have higher confidence in structures of genetic interactions than in individual interactions because dense clusters are very unlikely to occur by chance. Second, the functional level of a module removes redundancy by treating a set of genes as a unit. Because our method uses the associate-side of biclusters to determine annotations, genes that share a function are treated as a single unit. These functional units are summarized by an entropy measurement, the final pleiotropy score, which describes the shape of the distribution of modules among functions and differentiates broad from focused functional influence of a gene.

Through characterization of genes classified as having high and low pleiotropy, we found that evolution-related properties distinguished the groups. High-pleiotropy genes were more likely to be duplicated and to change in copy number throughout 30 yeast species. Contrasts in functional behaviors of the pleiotropy classes showed that high-pleiotropy genes have greater variability in expression, while low-pleiotropy genes are likely to be part of protein complexes. These interesting characterizations may shed light on the evolutionary processes through which genes may acquire multiple functions.

We propose that functional freedom is an important property enabling pleiotropy. Gene duplication and divergence is considered to be the primary source of raw material through which adaptions appear. The fact that WGD duplicate genes tend to have high pleiotropy suggests that this process of new adaptations allows the accumulation of diverse functions in single genes, as opposed to only yielding two specialized (i.e. low-pleiotropy) genes. Partial functional buffering, relatively common between duplicates (Musso et al., 2008; VanderSluis et al., 2010), likely plays a role in this process. One model consistent with these ideas is subfunctionalization, in which the functions of the

original gene are partitioned between the paired duplicates (Force et al., 1999). Following this process, the two genes complement each other such that each gene has functional regions maintained by selective constraint and degenerate regions that tolerate mutations and possibly acquire new functions. Even duplicates that are asymmetric in GI degree have been shown to maintain buffering relationships (VanderSluis et al., 2010). A second mechanism by which duplicates may diverge and become pleiotropic is suggested by the tendency of high-pleiotropy genes to have high variance in expression. Changes in regulatory patterns occurring soon after duplication may provide a route for acquiring environment-specific roles (Conant and Wolfe, 2006; Mattenberger et al., 2017). Functions unneeded in particular conditions may be altered to respond to new challenges, thus diversifying the gene's functions. Acquisition of new functions through variable expression is not limited to duplicates, but is proposed as a general mechanism promoting environmental and phenotypic adaptations (Stern et al., 2007; Tirosh et al., 2006). Finally, the significantly low number of pleiotropic genes that have membership in protein complexes suggests an avoidance of evolutionary constraint of sequence changes and consequent barrier to gaining novel functions.

While the characterization of pleiotropic genes as being sheltered from functional constraints provided by duplicates buffering each other and as lacking physical interactions in protein complexes offers insight into the kind of genes that are able to acquire new functions, it remains to be shown how pleiotropic genes have risen to such prominence that they are genetic interaction hubs. Indeed, the functional freedom suggested by our characterization of pleiotropic genes is a contrast to Fisher's classic geometric model of pleiotropy, which predicts that pleiotropic genes will be evolutionarily constrained and has been advanced by the "cost of complexity" model (Orr, 2000; Welch et al., 2003). However these characterizations can coexist at different time periods in a gene's life cycle: pleiotropy may originate over a relatively short period of time following de novo birth of a gene, a gene duplication, or a regulatory change buffered by a non-duplicate alternative pathway, and subsequent loss of buffering. Intriguingly, the TSA-derived high-pleiotropy genes contained a significantly higher number of *Saccharomyces*-specific genes as compared to the low-pleiotropy genes, which is evidence that participation in many processes can occur near the beginning of the lifecycle of a gene. Additionally, there may be mechanisms in place that stabilize the effects of evolving genes. Post-transcriptional regulation may be strong enough to

counteract expression-level patterns, therefore stabilizing protein levels when needed (Artieri and Fraser, 2014), and explaining the association between high-pleiotropy genes and high protein abundance. Similarly, genetic hubs have been shown to typically have steady expression levels, likely as a consequence of their importance (Park and Lehner, 2013), but the hubs that have variable expression levels are enriched for duplicates that may be able to buffer the effects of low expression (Park and Lehner, 2013). Overall, there appears to be a complicated relationship between pleiotropy, adaptation, and genomic robustness that has yet to be elucidated.

4.9 Methods for characterizing high- and low-pleiotropy genes

We used Wilcoxon rank-sum tests to compare the values of gene characteristics in our high- and low-pleiotropy gene classes, which were defined for genes with degree of at least the 50th percentile among all genes with pleiotropy scores. The gene characteristics are listed in section A4.3. All gene characteristics have data coverage of over 75% of genes (most have coverage over 95%), with the exception of three characteristics that were measured only for nonessential genes: phenotypic capacitance, curated phenotypes, and chemical genetic degree. These characteristics have coverage of nearly 60% of classified genes. We performed tests using pleiotropy scores obtained from the six different pleiotropy scoring configurations (for each of the TSA and DMA GI networks) and 22 testing variants, which are described below. The significance of p-values from rank-sum tests was determined using the FDR-control procedure described in Benjamini et al. (2006), treating the sets of 36 tests with identical set-ups, but different gene properties, as families.

4.9.1 Description of scoring configurations

Because there are multiple ways to derive gene functional profiles from SGA networks using the method presented in Chapter 3, we explored six reasonable versions, called pleiotropy scoring configurations. A scoring configuration comprises three specifications of the method (Figure 4.1 illustrates relevant details). First, functional profiles can be calculated for a network's query strains or array strains and this determines the set of strains (and genes) that will be assigned a pleiotropy score; we

use the label “query” or “array”. Second, only one of the two bicluster sides is used for determining bioprocess annotations and, in relation to the strains being analyzed (the first specification), it either includes the strains or consists of their interacting partners; we use the label “associate” or “adjacent”. Third, the set of annotations used may be the manual annotations, “MA”, or the systematic annotations “SAFE”. The six combinations of these three methods specifications that we used are “Query, associate, manual”, “Query, associate, SAFE”, “Query, adjacent, manual”, “Array, associate, manual”, “Array, associate, SAFE”, and “Array, adjacent, manual”.

4.9.2 Description of testing variants

For each scoring configuration, we performed multiple rank-sum tests, called “testing variants,” that explore different ways to define the high- and low-pleiotropy classes and control for possible biases that different types of mutant alleles may cause. The set of methods choices we considered that may affect any gene includes controlling for the GI degree of genes; removing strains that show weak signs of batch effects; and altering the percent of genes that are added to the high and low pleiotropy classes, which may be 20%, 30%, or 40%. Methods related to genes represented by TS and DAmP strains are the following: determining the gene’s pleiotropy score by taking the mean or maximum of the strains scores; discarding all strains of a mutation type (DAmP or TS); and applying a minimum GI degree threshold of 50 before averaging the degree of alleles to determine the high-degree genes that may be classified.

For each of these possibilities, we selected a default for use in reporting results and making figures. The default test variant is the following: degree controlling was used; in the case of multiple strains representing one gene, pleiotropy scores were averaged; only genes with the highest 50% of GI degree *after averaging strains* were kept (called "high-degree"); of high-degree genes, those with the highest and lowest 30% of pleiotropy scores were classified as having high and low pleiotropy, respectively.

Finally, to understand how robust our rank-sum results are, we selected a total of 22 testing variants (including the default) in which different methods are used. Only 12 of these are relevant for “array” scoring configurations because the sets of array strains in SGA networks are nearly all essential (TSA) or all nonessential (DMA). All combinations of these methods that we implemented are presented in Table 4.2.

Table 4.2. Test variants used for comparing gene properties in high- and low-pleiotropy genes. This table describes test variants in terms of modifications that were made to the default method described in the text. The method modification of not controlling for degree was paired with all other selected combinations of modifications to the default method, so we shortened the table by omitting it; each row represents two variants—one with and one without controlling for degree.

Short description	Modifications made to the default method
Default method	None
Max allele	In the case of multiple alleles representing one gene, the maximum pleiotropy score was used.
Remove batch-affected strains	Strains showing weak signs of batch effects were removed before selecting high-degree genes.
Min GI deg of 50, pleiotropy tails 20%	Strains with a degree less than 50 were removed before selecting high-degree genes; genes with the highest and lowest 20% of pleiotropy scores were classified as high and low pleiotropy.
No DAmP alleles	DAmP strains were removed before selecting high-degree genes.
Min GI deg of 50 for DAmPs	DAmP strains with GI degree lower than 50 were removed before selecting high-degree genes.
No TS alleles	TS strains were removed before selecting high-degree genes.
Min GI deg of 50 for TSs	TS strains with GI degree lower than 50 were removed before selecting high-degree genes.
Min GI deg of 50	All TS and DAmP strains with GI degree lower than 50 were removed before selecting high-degree genes.
Pleiotropy tails 20%	Genes with the highest and lowest 20% of pleiotropy scores were classified as high and low pleiotropy.
Pleiotropy tails 40%	Genes with the highest and lowest 40% of pleiotropy scores were classified as high and low pleiotropy.

The pleiotropy classes of high, medium, and low, are determined for each pairing of scoring configurations and testing variants because these methods choices affect which genes are considered. First, high-degree genes are identified as those with degree in the top 50% out of all genes that have been screened, have a pleiotropy score, and have not been removed by one of the test-variant modifications. For test variants that include degree control, we regress pleiotropy scores against degree and keep the pleiotropy residuals in place of pleiotropy scores. Then the pleiotropy scores (or score residuals) are divided into classes with the genes whose pleiotropy is in the top

30% of the high-degree genes labeled high pleiotropy, genes in the bottom 30% labeled as low pleiotropy, and the remaining 40% of genes labeled as medium pleiotropy.

Chapter 5: Conclusions and future work

5.1 Dissertation summary

Network structure is an important component of functional genomics and understanding it is required for prediction of phenotypes from genotypes. It is widely accepted that genes function through modules and that mechanisms of network structure promote robustness and adaptation, making biological complexity feasible. However, there are no characterizations of biological networks that successfully unify topological behavior with the functional roles of genes in determining phenotypes. The work described here investigates structure in the yeast genetic interaction network through the application of machine learning and data mining strategies. We use interpretable evolutionary, functional, and physical properties of genes to relate network structure to gene functional behavior.

We first showed that there is a high conservation of the relationship between gene characteristics and network structure by building an ensemble decision tree model that predicts negative GI degree. This model was trained using data from *S. cerevisiae*, but we successfully applied it to predict GI degree of *S. pombe* genes. We have therefore demonstrated that the structural properties of the *S. cerevisiae* GI network can be encapsulated in a model that is useful for other species. An important aspect of this result is that it suggests a practical method to guide the design of genetic interaction screens in other species, which have risen in interest since advances in gene-editing technology have made large-scale genetic interaction experiments feasible in more complex organisms (Doudna and Charpentier, 2014), including human. Since the yeast genome is extensively annotated and its GI network is completely mapped, models built in yeast may be able to capture patterns more complex than degree; our work suggests that these will be applicable to incomplete GI networks of other species.

Our next investigation of genetic interaction network structure led to gene functional characterization and a novel measurement of gene pleiotropy. The GI network is an ideal context for measuring pleiotropy because it is systematically measured and highly modular. We began with tackling the problem of module discovery by using frequent item set mining to extract all dense bipartite subgraphs of the GI network. Generally, previous studies explored networks using clustering methods that cannot reveal pleiotropy. The typical methods used, such as hierarchical clustering, are far too

limited to fully make use of large, comprehensive networks. After taking steps to remove redundancy of each gene's associated modules, we annotated modules with biological processes and used them to build functional profiles for all genes. Each functional profile describes the extent to which a gene's functional influence is focused in one sector of cellular function or is spread among many. We therefore used the entropy of functional profiles as a measurement of gene pleiotropy. This is the first measure of pleiotropy derived from a GI network, and it is notable in that it is based on a highly comprehensive and unbiased data set, identifies functions at a module level, and includes gene functions that are hidden in single-mutant strains but revealed in double mutants. To describe this new pleiotropy score, we compared many gene characteristics between groups of high- and low-pleiotropy genes. Surprisingly, we found that some gene characteristics expected to represent pleiotropic genes corresponded to the low, not high, pleiotropy genes. Gene characteristics that were positively associated with high-pleiotropy genes included high expression variance across environmental conditions, status as a WGD duplicate, high copy number volatility, and high protein abundance.

Pleiotropy has a history of being difficult to define and measure, but is a common and important aspect of the relationship between genotypes and phenotypes. Because of this, pleiotropy is expected to affect gene and network evolution. Classically, pleiotropic genes are expected to be highly constrained. This idea has never before been challenged, thus our results represent a notable departure from current theory. However, principles of network evolution are so far poorly described and many previous measures of pleiotropy are derived from experimental or curated data that contain biases. We hope that this new view of pleiotropy will inspire more investigations into using genome-wide measurements of function at a pathway level.

5.2 Future work

Despite the fact that biclusters mined from the genetic interaction network have significant benefits over modules identified by most clustering methods, they have shortcomings that can be addressed. One challenge we faced is that GI networks are known to have high rates of false negatives, and standard frequent item set mining only identifies complete bipartite structures. The false negatives in genetic interaction data cause fragmentation of network structures that are dense but not complete. Future work

may obtain more accurate functional modules through two approaches. Firstly, integration of other high throughput data sets, such as co-expression or physical interactions, with the GI network could help to fill in connections between genes that have close functional relationships not captured in genetic interactions. Secondly, there are frequent item set mining algorithms that are able to tolerate false negatives, and produce biclusters that are not complete bipartite structures. We investigated some of these algorithms and found them to be too computationally intense for the large size of the latest yeast network. However, these algorithms may be more useful if minimum support and item set size parameters are used to reduce the search space and identify larger modules. Additionally, new frequent item set mining algorithms and implementations may be available in the future.

The exhaustive set of biclusters we have discovered in the GI network remains a rich resource for characterizing the modular nature of cellular functions. Assessing the coherency of different gene characteristics in biclusters may suggest new properties of modules. One of the more thoughtfully designed models of protein network evolution (Kim and Marcotte, 2008) highlighted the physical constraints of proteins and took into account the characteristic of gene age. Consideration of a larger panel of gene characteristics, such as those we have collected, and how they distribute among bicluster modules may suggest further constraints that should be added to network evolution models. For example, we may be able to use evolutionary characteristics of genes to describe the age and evolution of individual modules.

The surprising contrast between the characteristics of highly-pleiotropic genes we have identified and the expectations that pleiotropic genes would show signs of slow, limited evolution should be investigated. In the case of controversy over inter- and intra-modular (date and party) protein interaction hubs, the context of co-expression and the behavior of singlish- and multi-interface proteins provided a convincing explanation. Considering that our set of low-pleiotropy genes had a high tendency to be part of physical modules and that inter-modular PPI hubs had more complex expression patterns that their counterparts, there may be a strong connection between pleiotropy and the PPI hub classes. An interesting first analysis could calculate date and party hubs directly from the GI network, in much the same way as it was done in the original publication, and compare the resulting gene classes to our measure of pleiotropy. We demonstrated here that bicluster modules more accurately reflect gene functions than

individual interactions, so the date-vs-party hub calculations could be modified to describe hub gene co-expression not with individual genes, but with modules.

Lastly, experimental characterizations of high-pleiotropy genes could substantially increase confidence in our pleiotropy measure. Double-mutant phenotypes, such as those observed through high-content screening (Vizeacoumar et al., 2010), could identify specific gene functions that are buffered in single-mutant phenotypes. Further, experiments could confirm the accuracy of gene participation in modules. For example, a gene that appears in a bicluster with a set of functionally-related genes but does not already have documented evidence of this function could be selected for experiments that precisely measure the gene's relationship to the relevant phenotypes.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., and Galle, R.F. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Agaisse, H., Burrack, L.S., Philips, J.A., Rubin, E.J., Perrimon, N., and Higgins, D.E. (2005). Genome-wide RNAi screen for host factors required for intracellular bacterial infection. Science *309*, 1248-1251.

Agarwal, S., Deane, C.M., Porter, M.A., and Jones, N.S. (2010). Revisiting Date and Party Hubs: Novel Approaches to Role Assignment in Protein Interaction Networks. PLOS Computational Biology *6*, e1000817.

Artieri, C.G., and Fraser, H.B. (2014). Evolution at two levels of gene expression in yeast. Genome Research *24*, 411-421.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics *25*, 25-29.

Bader, G.D., and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics *4*, 2.

Barabasi, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat Rev Genet *5*, 101-113.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., and Scholl, C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108-112.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., and Sherman, P.M. (2011). NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic acids research *39*, D1005-D1010.

Baryshnikova, A. (2016). Systematic functional annotation and visualization of biological networks. Cell systems *2*, 412-421.

Baryshnikova, A., Costanzo, M., Dixon, S., Vizeacoumar, F.J., Myers, C.L., Andrews, B., and Boone, C. (2010a). Synthetic genetic array (SGA) analysis in Saccharomyces cerevisiae and Schizosaccharomyces pombe. Methods in enzymology *470*, 145-179.

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.-Y., Ou, J., San Luis, B.-J., Bandyopadhyay, S.*, et al.* (2010b). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nature methods.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., and Cullin, C. (1993). A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. Nucleic Acids Research *21*, 3329-3330.

Bellay, J., Atluri, G., Sing, T.L., Toufighi, K., Costanzo, M., Ribeiro, P.S.M., Pandey, G., Baller, J., VanderSluis, B., Michaut, M.*, et al.* (2011a). Putting genetic interactions in context through a global modular decomposition. Genome research *21*, 1375-1387.

Bellay, J., Han, S., Michaut, M., Kim, T., Costanzo, M., Andrews, B.J., Boone, C., Bader, G.D., Myers, C.L., and Kim, P.M. (2011b). Bringing order to protein disorder through comparative genomics and genetic interactions. Genome biology *12*, R14-R14.

Benjamini, Y., Krieger, A.M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. Biometrika, 491-507.

Berglund, A.-C., Sjölund, E., Ostlund, G., and Sonnhammer, E.L.L. (2008). InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic acids research *36*, D263-266.

Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. Physical review E *67*, 031902.

Bezdek, J.C. (1981). Pattern recognition with fuzzy objective function algorithms (Plenum Press).

Boddy, M.N. (1998). Replication Checkpoint Enforced by Kinases Cds1 and Chk1. Science *280*, 909-912.

Boutros, M., Kiger, A.A., Armknecht, S., Kerr, K., Hild, M., Koch, B., Haas, S.A., Paro, R., Perrimon, N., and Consortium, H.F.A. (2004). Genome-wide RNAi analysis of growth and viability in Drosophila cells. Science *303*, 832-835.

Braun, P., and Gingras, A.C. (2012). History of protein–protein interactions: From egg-white to complex networks. Proteomics *12*, 1478-1498.

Breiman, L. (1996). Bagging predictors. Machine learning *24*, 123-140.

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

Broad Institute (2012). Download Sequence - Schizosaccharomyces group. http://www.broadinstitute.org/annotation/genome/schizosaccharomyces_group/MultiDownloads.html.

Brohée, S., and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. BMC bioinformatics *7*, 488.

Burga, A., Casanueva, M.O., and Lehner, B. (2011). Predicting mutation outcome from early stochastic variation in genetic interaction partners. Nature *480*, 250-253.

Butland, G., Babu, M., Diaz-Mejia, J.J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A.G., Pogoutse, O.*, et al.* (2008). eSGA: E. coli synthetic genetic array analysis. Nat Methods *5*, 789-795.

Byrne, A.B., Weirauch, M.T., Wong, V., Koeva, M., Dixon, S.J., Stuart, J.M., and Roy, P.J. (2007). A global analysis of genetic interactions in Caenorhabditis elegans. Journal of biology *6*, 8-8.

C. elegans Sequencing Consortium (1998). Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science *282*, 2012.

Chang, X., Xu, T., Li, Y., and Wang, K. (2013). Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date'and 'party'hubs. Scientific reports *3*, 1691.

Cheng, Y., and Church, G.M. (2000). Biclustering of expression data. Paper presented at: Ismb.

Chernikova, D., Motamedi, S., Csürös, M., Koonin, E.V., and Rogozin, I.B. (2011). A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. Biology Direct *6*, 26-26.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., and Engel, S.R. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research *40*, D700-D705.

Clauset, A., Moore, C., and Newman, M.E.J. (2008). Hierarchical structure and the prediction of missing links in networks. Nature *453*, 98-101.

Conant, G.C., and Wolfe, K.H. (2006). Functional partitioning of yeast co-expression networks after genome duplication. PLoS Biol *4*.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet *9*, 938-950.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S.*, et al.* (2010). The genetic landscape of a cell. Science (New York, NY) *327*, 425-431.

Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D.*, et al.* (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*.

Cyert, M.S. (2001). Genetic analysis of calmodulin and its targets in Saccharomyces cerevisiae. Annu Rev Genet *35*, 647-672.

De Wulf, P., McAinsh, A.D., and Sorger, P.K. (2003). Hierarchical assembly of the budding yeast kinetochore from multiple subcomplexes. Genes & Development *17*, 2902-2921.

Dixon, S.J., Costanzo, M., Baryshnikova, A., Andrews, B., and Boone, C. (2009). Systematic mapping of genetic interaction networks. Annual review of genetics *43*, 601-625.

Dixon, S.J., Fedyshyn, Y., Koh, J.L.Y., Prasad, T.S.K., Chahwan, C., Chua, G., Toufighi, K., Baryshnikova, A., Hayles, J., Hoe, K.-L.*, et al.* (2008). Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. Proceedings of the National Academy of Sciences of the United States of America *105*, 16653-16658.

Doudna, J.A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096.

Dowell, R.D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D.A., Rolfe, P.A., Heisler, L.E., and Chin, B. (2010). Genotype to phenotype: a complex problem. Science *328*, 469-469.

Du, L.-L., and Novick, P. (2001). Yeast Rab GTPase-activating Protein Gyp1p Localizes to the Golgi Apparatus and Is a Negative Regulator of Ypt1p. Molecular Biology of the Cell *12*, 1215-1226.

Duchaine, T.F., Wohlschlegel, J.A., Kennedy, S., Bei, Y., Conte, D., Pang, K., Brownell, D.R., Harding, S., Mitani, S., and Ruvkun, G. (2006). Functional proteomics reveals the biochemical niche of C. elegans DCR-1 in multiple small-RNA-mediated pathways. Cell *124*, 343-354.

Dudley, A.M., Janse, D.M., Tanay, A., Shamir, R., and Church, G.M. (2005). A global view of pleiotropy and phenotypically derived gene function in yeast. Molecular systems biology *1*, 2005.0001-2005.0001.

Duina, A.A., Miller, M.E., and Keeney, J.B. (2014). Budding Yeast for Budding Geneticists: A Primer on the Saccharomyces cerevisiae Model System. Genetics *197*, 33.

Duncan, M.C., Cope, M.J.T.V., Goode, B.L., Wendland, B., and Drubin, D.G. (2001). Yeast Eps15-like endocytic protein, Pan1p, activates the Arp2/3 complex. Nat Cell Biol *3*, 687-690.

Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.

Dutkowski, J., Kramer, M., Surma, M.A., Balakrishnan, R., Cherry, J.M., Krogan, N.J., and Ideker, T. (2013). A gene ontology inferred from molecular networks. Nat Biotech *31*, 38-45.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research *30*, 207-210.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics *5*, 113.

Ericson, E., Pylvänäinen, I., Fernandez-Ricaud, L., Nerman, O., Warringer, J., and Blomberg, A. (2006). Genetic pleiotropy in Saccharomyces cerevisiae quantified by high-resolution phenotypic profiling. Molecular genetics and genomics : MGG *275*, 605-614.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, pp. 226-231.

Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M.L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. Human Molecular Genetics *23*, 5866-5878.

Ferro-Novick, S., and Brose, N. (2013). Nobel 2013 Physiology or medicine: Traffic control system within cells. Nature *504*, 98-98.

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

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., and Mistry, J. (2013). Pfam: the protein families database. Nucleic acids research *42*, D222-D230.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A.*, et al.* (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research *44*, D279-D285.

Fischer, B., Sandmann, T., Horn, T., Billmann, M., Chaudhary, V., Huber, W., and Boutros, M. (2015). A map of directional genetic interactions in a metazoan cell. Elife *4*, e05464.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics *151*, 1531.

Frost, A., Elgort, Marc G., Brandman, O., Ives, C., Collins, Sean R., Miller-Vedam, L., Weibezahn, J., Hein, Marco Y., Poser, I., Mann, M.*, et al.* (2012). Functional Repurposing Revealed by Comparing S. pombe and S. cerevisiae Genetic Interactions. Cell *149*, 1339-1352.

Gagnon-Arsenault, I., Marois Blanchet, F.-C., Rochette, S., Diss, G., Dubé, A.K., and Landry, C.R. (2013). Transcriptional divergence plays a role in the rewiring of protein interaction networks after gene duplication. Journal of Proteomics *81*, 112-125.

Gasch, A.P., and Eisen, M.B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biology *3*, research0059.0051.

Gasch, A.P., Huang, M., Metzner, S., Botstein, D., Elledge, S.J., and Brown, P.O. (2001). Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. Molecular biology of the cell *12*, 2987-3003.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell *11*, 4241-4257.

Gaudet, P., Škunca, N., Hu, J.C., and Dessimoz, C. (2017). Primer on the gene ontology. The Gene Ontology Handbook, 25-37.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B.*, et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631-636.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., and Andre, B. (2002). Functional profiling of the Saccharomyces cerevisiae genome. nature *418*, 387-391.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., and Shen, Y. (2003). The international HapMap project.

Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. Nat Rev Mol Cell Biol *8*, 645-654.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., and Vitols, E. (2003). A protein interaction map of Drosophila melanogaster. Science *302*.

Girvan, M., and Newman, M.E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences *99*, 7821-7826.

Goffeau, A., Barrell, B.G., Bussey, H., and Davis, R. (1996). Life with 6000 genes. Science *274*, 546.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular biology and evolution *11*, 725-736.

Gout, J.-F., and Lynch, M. (2015). Maintenance and loss of duplicated genes by dosage subfunctionalization. Molecular biology and evolution *32*, 2141-2148.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., and Zeng, Q. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology *29*, 644-652.

Guan, Y., Dunham, M.J., and Troyanskaya, O.G. (2007). Functional Analysis of Gene Duplications in Saccharomyces cerevisiae. Genetics *175*, 933.

Haimovich, G., Medina, Daniel A., Causse, Sebastien Z., Garber, M., Millán-Zambrano, G., Barkai, O., Chávez, S., Pérez-Ortín, José E., Darzacq, X., and Choder, M. (2013). Gene Expression Is Circular: Factors for mRNA Degradation Also Foster mRNA Synthesis. Cell *153*, 1000-1011.

Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G., and Robertson, D.L. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biology *8*, R209.

Hammer, J.A., and Sellers, J.R. (2011). Walking to work: roles for class V myosins as cargo transporters. Nature Reviews Molecular Cell Biology *13*, 13-13.

Han, J.-D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P.*, et al.* (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature *430*, 88-93.

Hardy, S., Legagneux, V., Audic, Y., and Paillard, L. (2010). Reverse genetics in eukaryotes. Biology of the Cell *102*, 561-580.

Hart, G.T., Lee, I., and Marcotte, E.M. (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC bioinformatics *8*, 236.

Hartigan, J.A. (1972). Direct clustering of a data matrix. Journal of the american statistical association *67*, 123-129.

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

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature *402*, C47-C52.

Hatzimanikatis, V., Li, C., Ionita, J.A., and Broadbelt, L.J. (2004). Metabolic networks: enzyme function and metabolite structure. Current Opinion in Structural Biology *14*, 300-306.

Hein, M.Y., Hubner, N.C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I.A., Weisswange, I., Mansfeld, J., and Buchholz, F. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell *163*, 712-723.

Hertz-Fowler, C., Peacock, C.S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., and Rutherford, K. (2004). GeneDB: a resource for prokaryotic and eukaryotic organisms. Nucleic acids research *32*, D339-D343.

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

Hirsh, A.E., and Fraser, H.B. (2001). Protein dispensability and rate of evolution. Nature *411*, 1046-1049.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., and Boutilier, K. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180-183.

Holland, P.W.H., Marlétaz, F., Maeso, I., Dunwell, T.L., and Paps, J. (2017). New genes from old: asymmetric divergence of gene duplicates and the evolution of development. Philosophical Transactions of the Royal Society B: Biological Sciences *372*.

Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell *95*, 717-728.

Horn, T., Sandmann, T., Fischer, B., Axelsson, E., Huber, W., and Boutros, M. (2011). Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. Nat Meth *8*, 341-346.

Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686-691.

Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O.G. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics (Oxford, England) *22*, 2890-2897.

Huttenhower, C., Schroeder, M., Chikina, M.D., and Troyanskaya, O.G. (2008). The Sleipnir library for computational functional genomics. Bioinformatics *24*, 1559-1561.

Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., and Baltier, K. (2015). The BioPlex network: a systematic exploration of the human interactome. Cell *162*, 425-440.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. Nat Genet *31*, 370-377.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences *98*, 4569-4574.

Jones, D.T., and Ward, J.J. (2003). Prediction of disordered regions in proteins from position specific score matrices. Proteins: Structure, Function, and Bioinformatics *53*, 573-578.

Kelley, R., and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. Nature biotechnology *23*, 561-566.

Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature *428*, 617-624.

Khan, I., Chen, Y., Dong, T., Hong, X., Takeuchi, R., Mori, H., and Kihara, D. (2014). Genome-scale identification and characterization of moonlighting proteins. Biology direct *9*, 30-30.

Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G.*, et al.* (2010a). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nature biotechnology.

Kim, D.U., Hayles, J., Kim, D., Wood, V., Park, H.O., Won, M., Yoo, H.S., Duhig, T., Nam, M., Palmer, G.*, et al.* (2010b). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat Biotechnol *28*, 617-623.

Kim, P.M., Lu, L.J., Xia, Y., and Gerstein, M.B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. Science *314*, 1938-1941.

Kim, W.K., and Marcotte, E.M. (2008). Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. PLoS Comput Biol *4*, e1000232.

Kim, Y.-A., and Przytycka, T.M. (2012). Bridging the Gap between Genotype and Phenotype via Network Approaches. Frontiers in genetics *3*, 227-227.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature *267*, 275-276.

Kimura, M., and Ohta, T. (1974). On some principles governing molecular evolution. Proceedings of the National Academy of Sciences *71*, 2848-2852.

Kirschner, M., and Gerhart, J. (1998). Evolvability. Proceedings of the National Academy of Sciences *95*, 8420-8427.

Koch, E.N., Costanzo, M., Bellay, J., Deshpande, R., Chatfield-Reed, K., Chua, G., D'Urso, G., Andrews, B.J., Boone, C., and Myers, C.L. (2012). Conserved rules govern genetic interaction degree across species. Genome Biology *13*, R57.

Kofoed, M., Milbury, K.L., Chiang, J.H., Sinha, S., Ben-Aroya, S., Giaever, G., Nislow, C., Hieter, P., and Stirling, P.C. (2015). An Updated Collection of Sequence Barcoded Temperature-Sensitive Alleles of Yeast Essential Genes. G3: Genes|Genomes|Genetics *5*, 1879-1887.

Kondrashov, F.A., and Kondrashov, A.S. (2006). Role of selection in fixation of gene duplications. Journal of Theoretical Biology *239*, 141-151.

Kosco, K.A., Pearson, C.G., Maddox, P.S., Wang, P.J., Adams, I.R., Salmon, E.D., Bloom, K., and Huffaker, T.C. (2001). Control of Microtubule Dynamics by Stu2p Is Essential for Spindle Orientation and Metaphase Chromosome Alignment in Yeast. Molecular Biology of the Cell *12*, 2870-2880.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P.*, et al.* (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637-643.

Kume, K., Koyano, T., Kanai, M., Toda, T., and Hirata, D. (2011). Calcineurin ensures a link between the DNA replication checkpoint and microtubule-dependent polarized growth. Nature cell biology *13*, 234-242.

Lehner, B., Crombie, C., Tischler, J., Fortunato, A., and Fraser, A.G. (2006). Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nature genetics *38*, 896-903.

Levin, D.E. (2005). Cell wall integrity signaling in Saccharomyces cerevisiae. Microbiology and molecular biology reviews : MMBR *69*, 262-291.

Levy, S.F., and Siegal, M.L. (2008). Network hubs buffer environmental variation in Saccharomyces cerevisiae. PLoS biology *6*, e264-e264.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D.J., Chesneau, A., and Hao, T. (2004). A map of the interactome network of the metazoan C. elegans. Science *303*, 540-543.

Li, Z., Vizeacoumar, F.J., Bahr, S., Li, J., Warringer, J., Vizeacoumar, F.S., Min, R., VanderSluis, B., Bellay, J., and DeVit, M. (2011). Systematic exploration of essential yeast gene function with temperature-sensitive mutants. Nature biotechnology *29*, 361-367.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V.*, et al.* (2009). Population genomics of domestic and wild yeasts. Nature *458*, 337-341.

Longtine, M.S., McKenzie III, A., Demarini, D.J., Shah, N.G., Wach, A., Brachat, A., Philippsen, P., and Pringle, J.R. (1998). Additional modules for versatile and

economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. Yeast *14*, 953-961.

Lovell, S.C., and Robertson, D.L. (2010). An Integrated View of Molecular Coevolution in Protein–Protein Interactions. Molecular Biology and Evolution *27*, 2567-2575.

Luo, J., Emanuele, M.J., Li, D., Creighton, C.J., Schlabach, M.R., Westbrook, T.F., Wong, K.-K., and Elledge, S.J. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. Cell *137*, 835-848.

Lynch-Day, M.A., Bhandari, D., Menon, S., Huang, J., Cai, H., Bartholomew, C.R., Brumell, J.H., Ferro-Novick, S., and Klionsky, D.J. (2010). Trs85 directs a Ypt1 GEF, TRAPPIII, to the phagophore to promote autophagy. Proceedings of the National Academy of Sciences of the United States of America *107*, 7811-7816.

Madden, K., and Snyder, M. (1998). Cell polarity and morphogenesis in budding yeast. Annual review of microbiology *52*, 687-744.

Maji, P., and Paul, S. (2013). Rough-fuzzy clustering for grouping functionally similar genes from microarray data. IEEE/ACM transactions on computational biology and bioinformatics *10*, 286-299.

Malovannaya, A., Lanz, Rainer B., Jung, Sung Y., Bulynko, Y., Le, Nguyen T., Chan, Doug W., Ding, C., Shi, Y., Yucer, N., Krenciute, G.*, et al.* (2011). Analysis of the Human Endogenous Coregulator Complexome. Cell *145*, 787-799.

Mani, R., St Onge, R.P., Hartman, J.L., Giaever, G., and Roth, F.P. (2008). Defining genetic interaction. Proceedings of the National Academy of Sciences of the United States of America *105*, 3461-3466.

Mattenberger, F., Sabater-Muñoz, B., Toft, C., and Fares, M.A. (2017). The Phenotypic Plasticity of Duplicated Genes in Saccharomyces cerevisiae and the Origin of Adaptations. G3: Genes|Genomes|Genetics *7*, 63-75.

Medagli, B., Di Crescenzio, P., De March, M., and Onesti, S. (2016). Structure and Activity of the Cdc45-Mcm2–7-GINS (CMG) Complex, the Replication Helicase. In The Initiation of DNA Replication in Eukaryotes, D.L. Kaplan, ed. (Cham: Springer International Publishing), pp. 411-425.

Middendorf, M., Ziv, E., and Wiggins, C.H. (2005). Inferring network mechanisms: The Drosophila melanogaster protein interaction network. Proceedings of the National Academy of Sciences of the United States of America *102*, 3192-3197.

Mortimer, R.K., and Johnston, J.R. (1986). Genealogy of principal strains of the yeast genetic stock center. Genetics *113*, 35-43.

Musso, G., Costanzo, M., Huangfu, M., Smith, A.M., Paw, J., San Luis, B.-J., Boone, C., Giaever, G., Nislow, C., Emili, A.*, et al.* (2008). The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. Genome Research *18*, 1092-1099.

Myers, C.L., Robson, D., Wible, A., Hibbs, M.A., Chiriac, C., Theesfeld, C.L., Dolinski, K., and Troyanskaya, O.G. (2005). Discovery of biological networks from diverse functional genomic data. Genome biology *6*, R114.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840-846.

Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research *31*, 3812-3814.

Ohno, S. (1970). Evolution by Gene Duplication (London, New York: Unwin, Springer-Verlag).

Ohya, Y., and Botstein, D. (1994). Diverse Essential Functions Revealed by Complementing Yeast Calmodulin Mutants. Science *263*, 963-966.

Ohya, Y., Sese, J., Yukawa, M., Sano, F., Nakatani, Y., Saito, T.L., Saka, A., Fukuda, T., Ishihara, S., Oka, S.*, et al.* (2005). High-dimensional and large-scale phenotyping of yeast mutants. Proceedings of the National Academy of Sciences of the United States of America *102*, 19015-19020.

Orr, H.A. (2000). Adaptation and the cost of complexity. Evolution *54*, 13-20.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res *38*, D196-203.

Ow, Y.-L.P., Green, D.R., Hao, Z., and Mak, T.W. (2008). Cytochrome c: functions beyond respiration. Nat Rev Mol Cell Biol *9*, 532-542.

Paaby, A.B., and Rockman, M.V. (2013). The many faces of pleiotropy. Trends in genetics : TIG *29*, 66-73.

Page, B.D., Satterwhite, L.L., Rose, M.D., and Snyder, M. (1994). Localization of the Kar3 kinesin heavy chain-related protein requires the Cik1 interacting protein. The Journal of Cell Biology *124*, 507.

Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature *435*, 814-818.

Park, S., and Lehner, B. (2013). Epigenetic epistatic interactions constrain the evolution of gene expression. Molecular Systems Biology *9*.

Pastor-Satorras, R., Smith, E., and Solé, R.V. (2003). Evolving protein interaction networks through gene duplication. Journal of Theoretical Biology *222*, 199-210.

Pawson, T., and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. science *300*, 445-452.

Peter, J., and Schacherer, J. (2016). Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. Yeast *33*, 73-81.

Peters, C., and Mayer, A. (1998). Ca2+/calmodulin signals the completion of docking and triggers a late step of vacuole fusion. Nature *396*, 575-580.

Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nature Genetics *48*, 709-717.

Pritykin, Y., Ghersi, D., and Singh, M. (2015). Genome-Wide Detection and Analysis of Multifunctional Genes. PLoS computational biology *11*, e1004467-e1004467.

Pritykin, Y., and Singh, M. (2013). Simple Topological Features Reflect Dynamics and Modularity in Protein Interaction Networks. PLOS Computational Biology *9*, e1003243.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R.*, et al.* (2005). Global analysis of protein phosphorylation in yeast. Nature *438*, 679-684.

Pulverer, B. (2001). Trio united by division as cell cycle clinches centenary Nobel. Nature *413*, 553-553.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., and Clements, J. (2012). The Pfam protein families database. Nucleic acids research *40*, D290-D301.

Rapoport, I. (1940). Mnogokratnye linejnye povtoreniya uchastkov khromosom i ikh evolyucionnoe znachenie.[Multiple linear repeats of chromosome segments and their evolutionary significance]. Zh Obshchej Biologii *1*, 235-270.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. science *297*, 1551-1555.

Rees, J.S., Lowe, N., Armean, I.M., Roote, J., Johnson, G., Drummond, E., Spriggs, H., Ryder, E., Russell, S., and St Johnston, D. (2011). In vivo analysis of proteomes and interactomes using Parallel Affinity Capture (iPAC) coupled to mass spectrometry. Molecular & Cellular Proteomics *10*, M110. 002386.

Reiss, D.J., Baliga, N.S., and Bonneau, R. (2006). Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC bioinformatics *7*, 280.

Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of molecular biology *314*, 1041-1052.

Rhind, N., Chen, Z., Yassour, M., Thompson, D.A., Haas, B.J., Habib, N., Wapinski, I., Roy, S., Lin, M.F., Heiman, D.I.*, et al.* (2011). Comparative functional genomics of the fission yeasts. Science (New York, NY) *332*, 930-936.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*, 276-277.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotech *17*, 1030-1032.

Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S.R., Qu, H., Shales, M., Park, H.-O., Hayles, J.*, et al.* (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. Science (New York, NY) *322*, 405-410.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., and Mosca, R. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212-1226.

Rutherford, S.L. (2003). Between genotype and phenotype: protein chaperones and evolvability. Nat Rev Genet *4*, 263-274.

Rutherford, S.L., and Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. Nature *396*, 336-342.

Ryan, C.J., Krogan, N.J., Cunningham, P., and Cagney, G. (2013). All or nothing: protein complexes flip essentiality between distantly related eukaryotes. Genome biology and evolution *5*, 1049-1059.

Rzhetsky, A., and Gomez, S.M. (2001). Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. Bioinformatics *17*, 988-996.

Scholl, C., Fröhling, S., Dunn, I.F., Schinzel, A.C., Barbie, D.A., Kim, S.Y., Silver, S.J., Tamayo, P., Wadlow, R.C., and Ramaswamy, S. (2009). Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. Cell *137*, 821-834.

Schroer, T.A. (2004). Dynactin. Annual review of cell and developmental biology *20*, 759-779.

Schuldiner, M., Collins, S.R., Thompson, N.J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., and Greenblatt, J.F. (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell *123*, 507-519.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature genetics *34*, 166-176.

Service, R.F. (2006). Solo Winner Detailed Path From DNA to RNA. Science *314*, 236.

SGD Project (2010). http://www.yeastgenome.org/download-data/.

Sharan, R., and Shamir, R. (2000). CLICK: a clustering algorithm with applications to gene expression analysis.

Sheng, Q., Moreau, Y., and De Moor, B. (2003). Biclustering microarray data by Gibbs sampling. Bioinformatics *19*, ii196-ii205.

Shih, Y.-K., and Parthasarathy, S. (2012). Identifying functional modules in interaction networks through overlapping Markov clustering. Bioinformatics *28*, i473-i479.

Shou, C., Bhardwaj, N., Lam, H.Y., Yan, K.-K., Kim, P.M., Snyder, M., and Gerstein, M.B. (2011). Measuring the evolutionary rewiring of biological networks. PLoS computational biology *7*, e1001050.

Sipiczki, M. (2000). Where does fission yeast sit on the tree of life. Genome Biol *1*, 1011-1011.

Skelly, D.A., Merrihew, G.E., Riffle, M., Connelly, C.F., Kerr, E.O., Johansson, M., Jaschob, D., Graczyk, B., Shulman, N.J., Wakefield, J.*, et al.* (2013).

Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Research *23*, 1496-1504.

Smits, A.H., and Vermeulen, M. (2016). Characterizing protein–protein interactions using mass spectrometry: challenges and opportunities. Trends in biotechnology *34*, 825-834.

Sonnhammer, E.L.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: Multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Research *26*, 320-322.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular biology of the cell *9*, 3273-3297.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res *34*.

Stearns, F.W. (2010). One hundred years of pleiotropy: a retrospective. Genetics *186*, 767-773.

Stelling, J., Sauer, U., Szallasi, Z., Doyle Iii, F.J., and Doyle, J. (2004). Robustness of Cellular Functions. Cell *118*, 675-685.

Stern, S., Dror, T., Stolovicki, E., Brenner, N., and Braun, E. (2007). Genome-wide transcriptional plasticity underlies cellular adaptation to novel challenge. Molecular Systems Biology *3*, 106.

Stevens, R.C., and Davis, T.N. (1998). Mlc1p Is a Light Chain for the Unconventional Myosin Myo2p in Saccharomyces cerevisiae. The Journal of Cell Biology *142*, 711.

Strambio-De-Castillia, C., Niepel, M., and Rout, M.P. (2010). The nuclear pore complex: bridging nuclear transport and gene regulation. Nature Reviews Molecular Cell Biology *11*, 490-501.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. science *302*, 249-255.

Sugiura, R., Sio, S.O., Shuntoh, H., and Kuno*, T. (2001). Molecular genetic analysis of the calcineurin signaling pathways. Cellular and Molecular Life Sciences *58*, 278-288.

Sundberg, H.A., Goetsch, L., Byers, B., and Davis, T.N. (1996). Role of calmodulin and Spc110p interaction in the proper assembly of spindle pole body compenents. The Journal of Cell Biology *133*, 111.

Suzuki, M., Igarashi, R., Sekiya, M., Utsugi, T., Morishita, S., Yukawa, M., and Ohya, Y. (2004). Dynactin is involved in a checkpoint to monitor cell wall synthesis in Saccharomyces cerevisiae. Nature cell biology *6*, 861-871.

Taipale, M., Jarosz, D.F., and Lindquist, S. (2010). HSP90 at the hub of protein homeostasis: emerging mechanistic insights. Nat Rev Mol Cell Biol *11*, 515-528.

Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. Bioinformatics *18*, S136-S144.

The Gene Ontology Consortium (2012). GO Downloads. http://www.geneontology.org/page/downloads.

Thieffry, D., Huerta, A.M., Pérez-Rueda, E., and Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays *20*, 433-440.

Tirosh, I., Weinberger, A., Carmi, M., and Barkai, N. (2006). A genetic signature of interspecies variations in gene expression. Nat Genet *38*, 830-834.

Tischler, J., Lehner, B., and Fraser, A.G. (2008). Evolutionary plasticity of genetic interaction networks. Nature genetics *40*, 390-391.

Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., and Chang, M. (2004). Global mapping of the yeast genetic interaction network. Science *303*.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics *17*, 520-525.

Typas, A., Nichols, R.J., Siegele, D.A., Shales, M., Collins, S.R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B.L.*, et al.* (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. Nat Methods *5*, 781-787.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P.*, et al.* (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623-627.

Ungermann, C., Nichols, B.J., Pelham, H.R., and Wickner, W. (1998). A vacuolar v-t-SNARE complex, the predominant form in vivo and on isolated vacuoles, is disassembled and activated for docking and fusion. J Cell Biol *140*, 61-69.

Van Dongen, S.M. (2001). Graph clustering by flow simulation.

VanderSluis, B., Bellay, J., Musso, G., Costanzo, M., Papp, B., Vizeacoumar, F.J., Baryshnikova, A., Andrews, B., Boone, C., and Myers, C.L. (2010). Genetic interactions reveal the evolutionary trajectories of duplicate genes. Molecular Systems Biology *6*.

Vázquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Modeling of protein interaction networks. Complexus *1*, 38-44.

Veraksa, A., Bauer, A., and Artavanis-Tsakonas, S. (2005). Analyzing protein complexes in Drosophila with tandem affinity purification–mass spectrometry. Developmental Dynamics *232*, 827-834.

Vidal, M., Cusick, Michael E., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. Cell *144*, 986-998.

Vizeacoumar, F.J., Arnold, R., Vizeacoumar, F.S., Chandrashekhar, M., Buzina, A., Young, J.T.F., Kwan, J.H.M., Sayad, A., Mero, P., Lawo, S.*, et al.* (2013). A negative genetic interaction map in isogenic cancer cell lines reveals cancer cell vulnerabilities. Molecular Systems Biology *9*, 696-696.

Vizeacoumar, F.J., van Dyk, N., S.Vizeacoumar, F., Cheung, V., Li, J., Sydorskyy, Y., Case, N., Li, Z., Datti, A., Nislow, C.*, et al.* (2010). Integrating high-throughput genetic interaction mapping and high-content screening to explore yeast spindle morphogenesis. The Journal of Cell Biology *188*, 69.

Vo, T.V., Das, J., Meyer, M.J., Cordero, N.A., Akturk, N., Wei, X., Fair, B.J., Degatano, A.G., Fragoza, R., and Liu, L.G. (2016). A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. Cell *164*, 310-323.

Wach, A., Brachat, A., Pöhlmann, R., and Philippsen, P. (1994). New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. Yeast *10*, 1793-1808.

Wagner, A. (2005). Robustness and Evolvability in Living Systems (Princeton University Press).

Wagner, G.P., and Zhang, J. (2011). The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nat Rev Genet *12*, 204-213.

Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. Science *287*, 116-122.

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., and Bezginov, A. (2015). Panorama of ancient metazoan macromolecular complexes. Nature.

Wapinski, I. (2009). Fungal Orthogroups Repository. https://portals.broadinstitute.org/regev/orthogroups/.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007a). Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics *23*, i549-i558.

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

Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F., and Jones, D.T. (2004a). The DISOPRED server for the prediction of protein disorder. Bioinformatics (Oxford, England) *20*, 2138-2139.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004b). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. Journal of molecular biology *337*, 635-645.

Welch, J.J., Waxman, D., and Houle, D. (2003). Modularity and the cost of complexity. Evolution *57*, 1723-1734.

Wong, S., Butler, G., and Wolfe, K.H. (2002). Gene order evolution and paleopolyploidy in hemiascomycete yeasts. Proceedings of the National Academy of Sciences of the United States of America *99*, 9272-9277.

Wood, V. (2006). Schizosaccharomyces pombe comparative genomics; from sequence to systems. In Comparative genomics (Springer), pp. 233-285.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution *24*, 1586-1591.

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular biology and evolution *17*, 32-43.

Yoshida, T., Toda, T., and Yanagida, M. (1994). A calcineurin-like gene ppb1+ in fission yeast: mutant defects in cytokinesis, cell polarity, mating and spindle pole body positioning. Journal of cell science *107 (Pt 7*, 1725-1735.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Yu, M.K., Kramer, M., Dutkowski, J., Srivas, R., Licon, K., Kreisberg, J., Ng, C.T., Krogan, N., Sharan, R., and Ideker, T. (2016). Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems. Cell Syst *2*, 77-88.

Zaki, M.J., Parthasarathy, S., Ogihara, M., and Li, W. (1997). New Algorithms for Fast Discovery of Association Rules. Paper presented at: KDD.

Zarin, T., and Moses, A.M. (2014). Insights into molecular evolution from yeast genomics. Yeast *31*, 233-241.

Zhang, A. (2006). Advanced analysis of gene expression microarray data, Vol 1 (World Scientific Publishing Co Inc).

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., and Houfek, T. (2001). Global analysis of protein activities using proteome chips. science *293*, 2101-2105.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. Genes & Development *21*, 1010-1024.

Zotenko, E., Mestre, J., O'Leary, D.P., and Przytycka, T.M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. PLoS Comput Biol *4*, e1000140.

Appendix 1: Term definitions

Initializations and acronyms

AP-MS...affinity purification followed by mass spectrometry

BLAST...Basic Local Alignment Search Tool

CAI .....codon adaptation index

DAmP....decreased abundance by mRNA perturbation

DD .....duplication and divergence

DMA .....deletion mutant array

dN/dS ....ratio of the rate of nonsynonymous mutations to the rate of synonymous mutations

DNA.....deoxyribonucleic acid

FCM.....fuzzy c-means

GEO .....Gene Expression Omnibus

GFP .....green fluorescent protein

GI.....genetic interaction

GO.....Gene Ontology

HR .....homologous recombination

MA .....manual annotations (annotation gold standard)

MCL.....Markov clustering

N_c.....effective number of codons

NETO ....new-end take off (phase of *S. pombe* growth)

PCC.....Pearson's correlation coefficient

PCR.....polymerase chain reaction

PPI.....protein-protein interaction

RNA.....ribonucleic acid

SAFE.....spatial analysis of functional enrichment

SGA.....synthetic genetic array (high-throughput method of mating yeast strains, notably used to construct GI networks through calculation of SGA scores)

SM .....single mutant

SNP .....single-nucleotide polymorphism

SSD.....small-scale duplication

TS.....temperature sensitive

TSA .....temperature sensitive array

UTR.....untranslated region (of mRNA molecules)

WGD.....whole-genome duplication

Y2H .....yeast two-hybrid

Glossary

Also see Appendix 4, section A4.3 for definitions of gene characteristics.

Allele A version of a gene.

Array gene A gene that is represented by a single-mutant SGA array strain.

Array strain In the context of SGA, a yeast strain in a fixed collection, such as the deletion mutant array or the temperature sensitive array, that contains an array-specific selectable marker and is typically mated with a query strain.

Bicluster In a matrix, a subset of rows paired with a subset of columns such that the elements in the intersection are meaningful.

Bipartite Consisting of two disjoint sets of nodes and a set of edges such that no edge occurs between two nodes in the same set.

Cluster A set of data points that show higher similarity with each other than with other data points, e.g. densely connected nodes in a network.

Connected For two nodes, linked with a network edge; for more nodes, linked with edges. (The graph-theoretic definition involving a path is not used in this document; here, and typically in genomics, the term refers to a single edge, or multiple edges that may or may not be incident.)

Conservation Retention through evolution, often observed through identification in two diverged species; often said of a sequence, structure, pattern or relationship.

DAmP A method in which a marker inserted into the 3' UTR of a gene destabilizes the gene transcripts, causing low protein levels.

Degree The number of edges a node participates in; equivalent to the number of neighbors the node has.

Deletion collection For *S. cerevisiae* or *S. pombe*, a set of mutant strains, each of which has a single gene removed from its genome; nearly all nonessential genes are represented by a strain.

Deletion mutant A strain in which a known gene has been removed.

Deletion mutant array The fixed set of deletion-mutant yeast strains that are each mated with every query strain in an SGA experiment.

Edge One of the two basic types of units that compose a graph or network: a relationship between two nodes in a graph; e.g. an interaction in an interaction network.

Entropy A measure of diversity; specifically in this document, Shannon entropy; given a distribution, the average information gained from a single data point. See Appendix 4, section A4.1.

Epsilon score The interaction score for a pair of strains screened in an SGA genetic interaction experiment.

Essential Required for viability of an organism (typically said of a gene).

Fitness In yeast, the observed growth rate relative to that of a wild type strain.

Fitness defect Decrease of fitness in comparison to wild type fitness, i.e. 1 – fitness.

Gene A region of DNA that is transcribed as a unit; the RNA transcript may be used directly or translated into a protein.

Graph A mathematical structure composed of a set of objects (nodes) and a set of relationships (edges) that each link two objects (nodes) together.

Homolog An ortholog or paralog.

Hub A node with a high degree, used as a general description or in relation to a defined minimum degree.

Mutant A strain or organism that contains mutations; not wild type.

Mutation A change in DNA sequence, usually relative to that of another strain or species, such as a standard laboratory strain or ancestral species.

Network Real-world phenomenon or specific concept that can be modeled as objects with relationships (i.e. represented as a graph) or its representative graph data set; graph terminology can be used for networks.

Node One of the two basic types of units that compose a graph or network, it may be joined to another node with an edge; e.g. usually a gene or protein in this document.

Nonessential Not required for viability of an organism, but possibly causing a reduction in fitness if mutated (said of a gene).

Orthologs Genes of different species that evolved from one ancestral gene.

Paralogs Genes in a single genome that evolved from one ancestral gene and have detectable sequence similarity with each other.

Profile Array (ordered list) of all interaction measurements collected for a single node in a network.

Query gene A gene that is represented by a single-mutant SGA query strain.

Query strain In the context of SGA, a yeast strain that contains a query-specific selectable marker and is mated with an array strain.

Small-scale duplicate Gene that derives from a "parent" gene that was duplicated by a mutation event that did not affect the entire genome.

Strain A version of a species, as defined by its genome, which may have a known or unknown sequence. In the context of the yeast GI network, strains are experimentally constructed and each represents a single gene; for ease of language in this context, a strain may be referred to as a gene occasionally.

Subgraph A subset of a graph's nodes and all or some of the edges between them.

Temperature sensitive Said of a mutant allele or strain harboring a mutant allele, has a phenotype that occurs only at high temperature.

Temperature sensitive array The fixed set of temperature-sensitive yeast strains that are each mated with every query strain in an SGA experiment.

Whole-genome duplicate In yeast, one of two paralogs created by the whole genome duplication event that occurred approximately 100 million years ago in an ancestor of *S. cerevisiae* and its closest relatives in six genera.

Wild type Of the standard reference version of a strain or species, which is typically standard in terms of its DNA sequence; not a mutant.

Appendix 2: Supplementary items for Chapter 2

A2.1 Supplementary figures

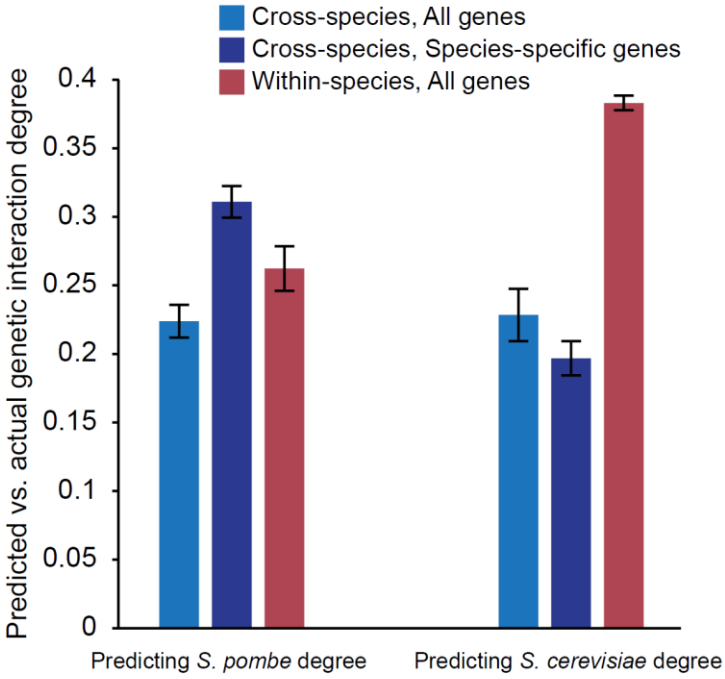


Figure A2.1. Evaluation of prediction performance excluding the SM fitness defect feature from bagged regression tree models. Models were trained on all features listed in Table 2.1 except for SM fitness defect. Pearson correlation coefficients between predicted and actual negative genetic interaction degrees were averaged across 25 repetitions of model construction, shown here with error bars of standard deviation. The left set of bars shows the performance of predictions made for ~550 *S. pombe* genes and the right set of bars shows the performance of predictions made for all nonessential deletion mutants in *S. cerevisiae*. For each scenario, models were trained both on data from the same species (red bar) as well as data from the other species (blue bars). The light blue bars correspond to predicting degrees of all genes in the test species, while the dark blue bars correspond to predicting the subset of genes lacking orthologs in the training species.

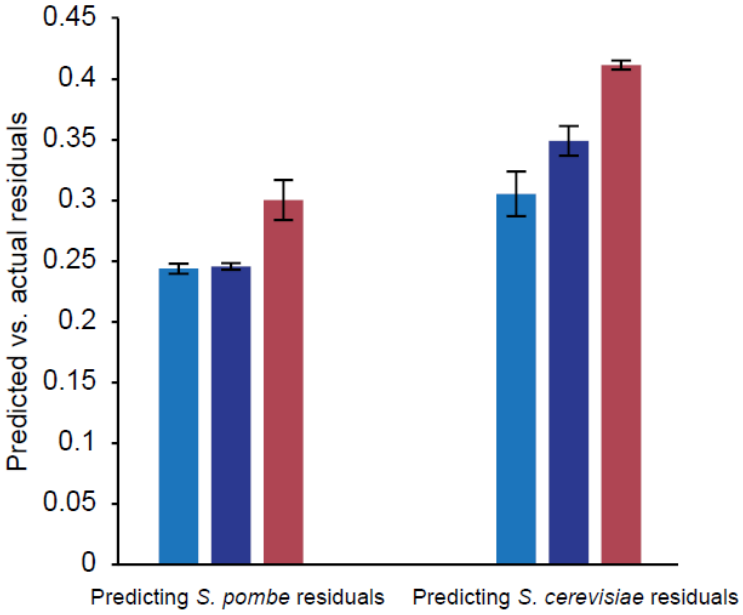
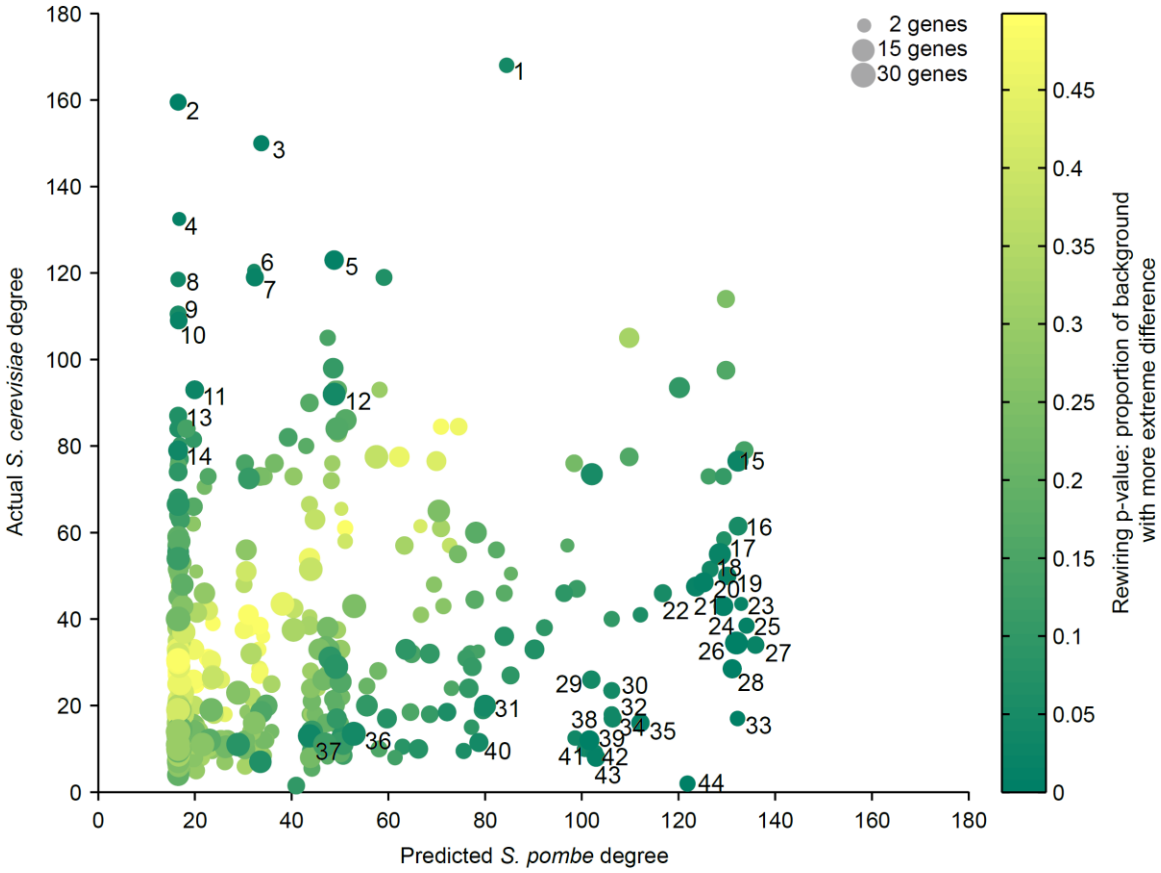


Figure A2.2. Patterns in gene characteristics other than SM fitness defect show predictive ability not captured by SM fitness defect alone. Using all features except SM fitness defect, models were trained to predict the residual negative genetic interaction degree that remained after subtracting degree predictions made from a regression tree model that was trained on the single feature SM fitness defect. Details of the bar chart are the same as those specified for Figure A2.1.



1 (c) histone ubiquitination
2 (p) phospholipid dephosphorylation
3 (p) cytokinesis checkpoint
4 (c) mitochondrial DNA metabolic process
5 (c) histone exchange
6 (p) cellular response to osmotic stress
7 (p) regulation of calcium−mediated signaling
8 (p) proton−transporting ATP synthase complex assembly
9 (p) positive regulation of transcription from RNA polymerase II promoter, meiotic
10 (p) protein insertion into membrane raft
11 (c) establishment of mitotic spindle orientation
12 (c) histone deacetylation
13 (c) acetyl−CoA biosynthetic process from pyruvate
14 (c) hyperosmotic response
15 (p) CVT pathway
16 (p) meiotic gene conversion
17 (c) double−strand break repair via homologous recombination
18 (p) positive regulation of histone methylation
19 (c) mitochondrial electron transport, ubiquinol to cytochrome c
20 (p) mating type switching
21 (c) ribosomal small subunit assembly
22 (p) regulation of cellular response to stress
23 (c) regulation of nuclease activity
24 (c) double−strand break repair via single−strand annealing
25 (c) meiotic DNA double−strand break processing
26 (c) CVT pathway
27 (c) C−terminal protein lipidation
28 (c) meiotic DNA double−strand break formation
29 (c) nucleotide−excision repair, DNA damage recognition
30 (c) L−serine biosynthetic process
31 (c) adaptation of signaling pathway by response to pheromone involved in conjugation with cellular fusion
32 (c) donor selection
33 (c) positive regulation of lipid metabolic process
34 (c) regulation of lipid biosynthetic process
35 (c) response to hexose stimulus
36 (c) lipid transport
37 (p) response to reactive oxygen species
38 (p) regulation of intracellular protein kinase cascade
39 (p) positive regulation of protein modification process
40 (c) maintenance of cell polarity
41 (p) phospholipid translocation
42 (c) activation of protein kinase activity
43 (p) chromatin silencing at rDNA
44 (c) pyrimidine salvage

Figure A2.3. Global analysis of rewiring based on whole-genome predictions in *S. pombe*. Points in the scatter plot each represent groups of between two and 23 genes that are annotated with the same GO term (section 2.8.3). Darker color represents complexes that are predicted to have significant rewiring. Generally, genes in GO-term

groups that fall on the diagonal are predicted to have conserved degrees, while those that fall far off-diagonal show evidence for large degree differences between the two species. Significantly rewired groups are labeled by their GO terms.

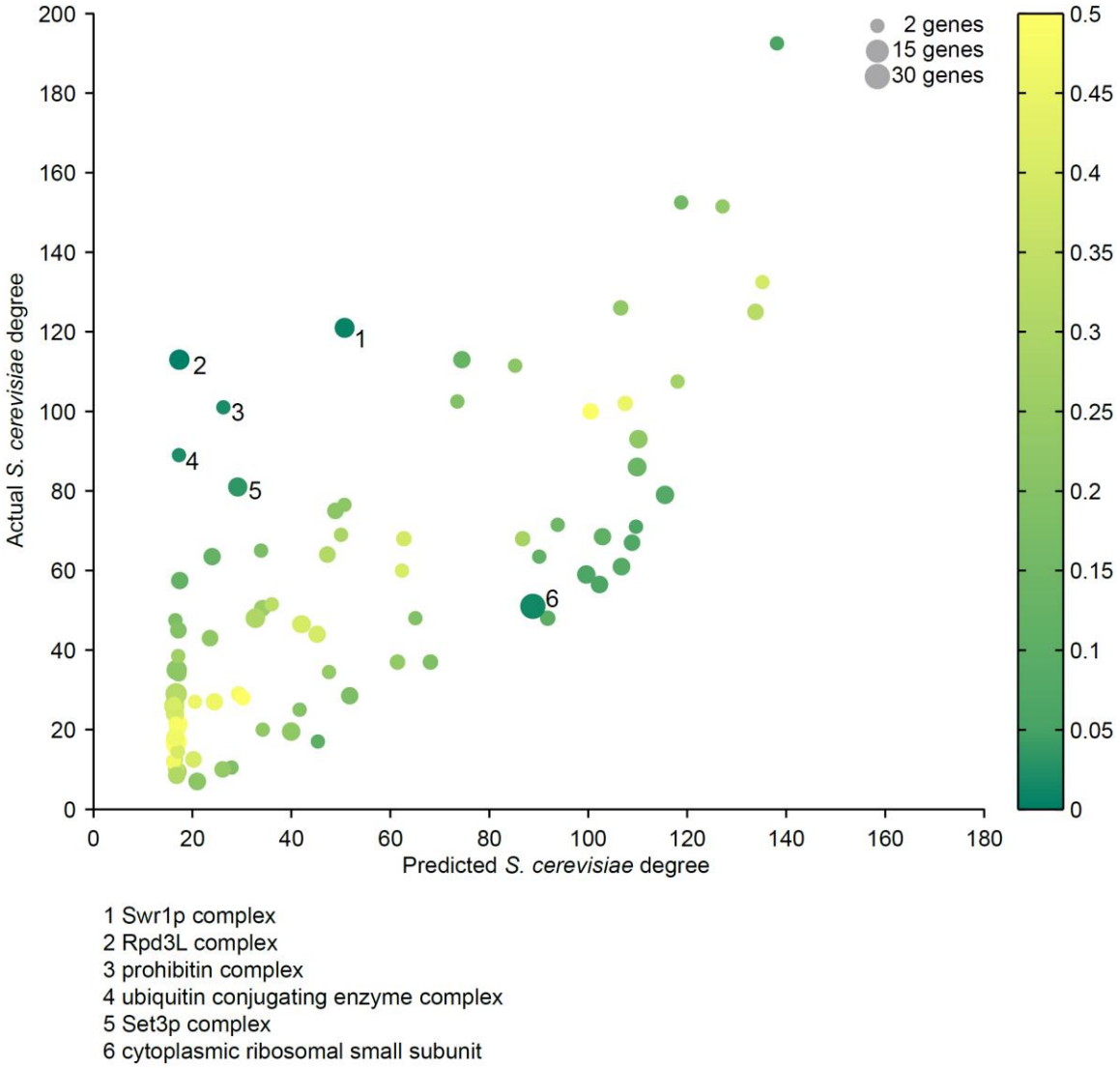
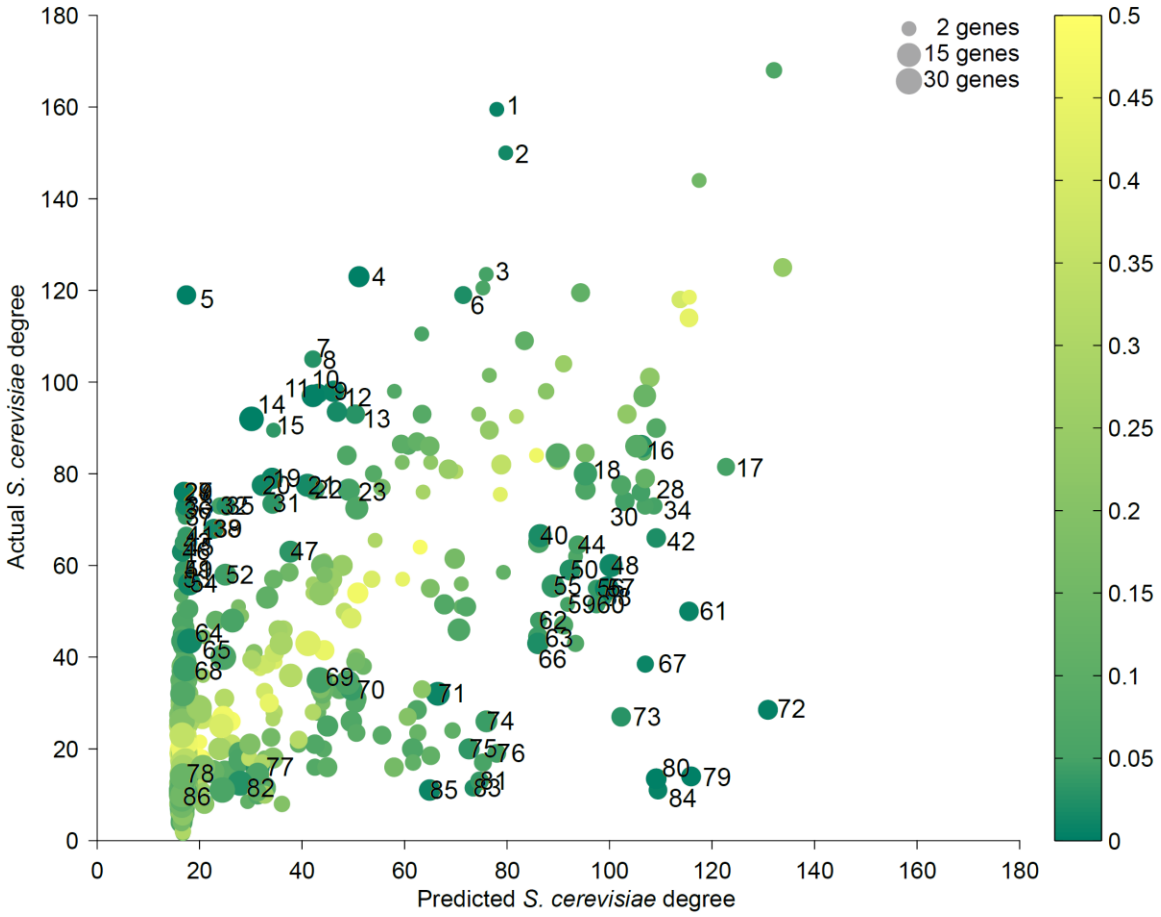


Figure A2.4. Within-species control for cross-species rewiring analysis. The rewiring-discovery procedure was applied to *S. cerevisiae* genes and their predicted and actual genetic interaction degrees (substituting *S. cerevisiae* predictions for *S. pombe* predictions in the pipeline). **(A)** This within-species evaluation revealed six out of 91 complexes that appeared significantly rewired ($p < 0.05$). While this is fewer than was identified in the *S. pombe-S. cerevisiae* comparison, this is more than expected by

chance, which likely reflects complexes for which we are systematically over- or under-predicting actual degrees. Of the 11 rewired complexes (Figure 2.4A), 4 of these are among the six complexes significant in the control experiment. **(B, figure below)** 14 of the 44 predicted rewired GO term in the *S. pombe-S. cerevisiae* comparison (Figure A2.3) also showed significance in the within-species control. We suggest that these cases should be excluded from further analysis, as they likely reflect systematic prediction errors, not true cases of cross-species differences.



1 (p) phospholipid dephosphorylation
2 (p) cytokinesis checkpoint
3 (p) mitotic spindle elongation
4 (c) histone exchange
5 (p) regulation of calcium−mediated signaling
6 (c) barrier septum formation
7 (c) positive regulation of intracellular protein kinase cascade
8 (c) regulation of MAPKKK cascade
9 (p) positive regulation of cell communication
10 (c) positive regulation of response to stimulus
11 (p) protein amino acid deacetylation
12 (p) chromatin silencing at telomere
13 (c) establishment of mitotic spindle orientation
14 (c) histone deacetylation
15 (c) phosphatidylinositol biosynthetic process
16 (p) retrograde transport, endosome to Golgi
17 (c) polyamine biosynthetic process
18 (c) retrograde transport, endosome to Golgi
19 (c) hyperosmotic response
20 (p) regulation of transcription, mitotic
21 (p) chromatin silencing at centromere
22 (p) protein insertion into membrane
23 (p) CVT pathway
24 (p) mRNA export from nucleus in response to heat stress
25 (p) stress−activated MAPK cascade
26 (p) cellular sodium ion homeostasis
27 (p) response to arsenic
28 (p) DNA synthesis involved in DNA repair
29 (c) autophagic vacuole assembly
30 (p) protein targeting to peroxisome
31 (p) chromatin silencing at silent mating−type cassette
32 (p) negative regulation of transcription from RNA polymerase II promoter, mitotic
33 (p) endoplasmic reticulum unfolded protein response
34 (p) meiotic DNA double−strand break formation
35 (c) DNA replication−independent nucleosome assembly
36 (p) regulation of cell shape
37 (c) protein homooligomerization
38 (c) protein insertion into ER membrane
39 (c) response to arsenic
40 (p) peroxisome organization

Continued on next page.

41 (c) regulation of ubiquitin homeostasis
42 (p) pantothenate metabolic process
43 (c) septin checkpoint
44 (c) protein import into peroxisome matrix, docking
45 (c) misfolded or incompletely synthesized protein catabolic process
46 (c) protein insertion into membrane
47 (c) nuclear migration along microtubule
48 (c) double−strand break repair via nonhomologous end joining
49 (p) protein retention in ER lumen
50 (p) double−strand break repair via homologous recombination
51 (c) protein retention in ER lumen
52 (p) microtubule−based movement
53 (p) autophagic vacuole assembly
54 (c) endoplasmic reticulum unfolded protein response
55 (c) protein import into peroxisome matrix
56 (p) DNA replication−dependent nucleosome assembly
57 (c) double−strand break repair via homologous recombination
58 (p) negative regulation of translation
59 (p) positive regulation of histone methylation
60 (c) DNA replication−dependent nucleosome assembly
61 (c) mitochondrial electron transport, ubiquinol to cytochrome c
62 (p) regulation of chromatin silencing at centromere
63 (c) heteroduplex formation
64 (c) postreplication repair
65 (c) cellular response to organic substance
66 (c) double−strand break repair via single−strand annealing
67 (c) meiotic DNA double−strand break processing
68 (c) response to biotic stimulus
69 (c) microautophagy
70 (p) protein deubiquitination
71 (c) protein deubiquitination
72 (c) meiotic DNA double−strand break formation
73 (c) flocculation via cell wall protein−carbohydrate interaction
74 (c) response to salt stress
75 (c) karyogamy involved in conjugation with cellular fusion
76 (c) phospholipid dephosphorylation
77 (p) dicarboxylic acid metabolic process
78 (p) cellular amino acid derivative metabolic process
79 (c) polyamine metabolic process
80 (c) pantothenate metabolic process
81 (c) sulfur amino acid transport
82 (c) cellular amino acid derivative biosynthetic process
83 (c) FAD transport
84 (c) isocitrate metabolic process
85 (c) aromatic amino acid family biosynthetic process
86 (p) nicotinamide nucleotide metabolic process

Figure A2.4B. See entry in caption above.

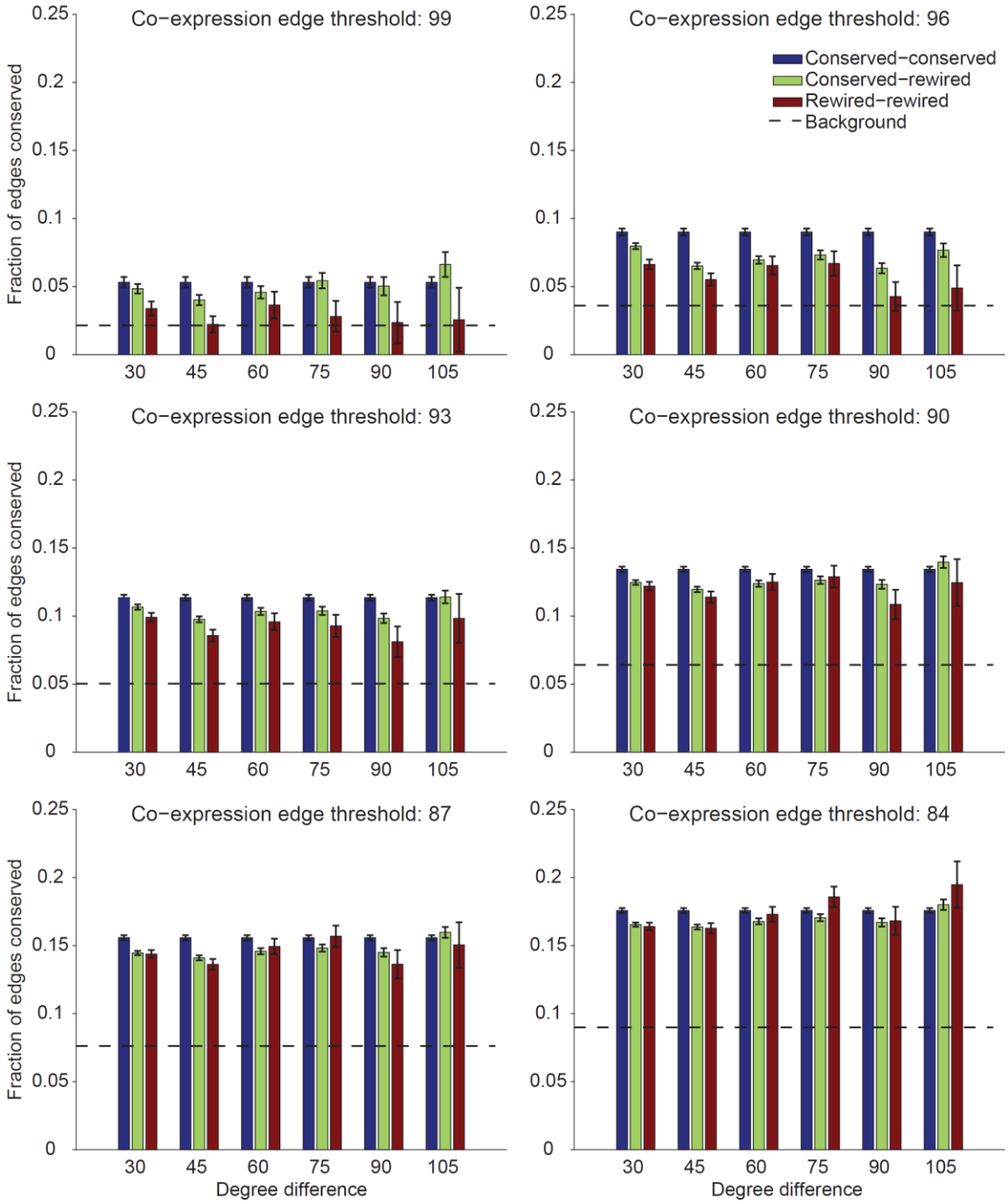


Figure A2.5. Validation of rewiring predictions is robust across a range of co-expression percentile thresholds used to define networks. As described in the main text and Figure 2.4, we constructed networks of co-expression relationships among genes for each yeast species, then labeled genes according to our rewiring designation. Edges in the co-expression network were classified by whether involved genes were

both rewired, only one was rewired, or neither was rewired. Bars show fractions of conserved co-expression relationships between species within each of these classes and error bars are 95% confidence intervals for the binomial proportion test. Panels show conservation results from co-expression networks that differ in their confidence and density, both of which are affected by placing a threshold, which is a percentile, on co-expression levels and retaining only edges corresponding to levels that exceed the threshold. Note that we observe a significant difference between the conserved-conserved and rewired-rewired classes for a range of cutoffs. Also, the significance of the difference diminishes for weaker thresholds, likely due to an abundance of spurious co-expression edges allowed at the cutoffs.

A2.2 Gene characteristics

Yeast conservation

Yeast conservation is a count of how many of 23 different species of Ascomycota fungi possess an ortholog of a given gene. This measure was first described in Wapinski et al. (2007b), and ortholog data was downloaded from the Fungal Orthogroups Repository (Wapinski, 2009). The 23 species are an expanded set of the 17 species described in the study, with the additions of *S. octosporus*, *S. japonicus*, *L. elongosporus*, *C. parasilosis*, *C. tropicalis*, and *C. guilliermondii*.

Broad conservation

Similar, though complementary, to yeast conservation, broad conservation is a count of how many out of a set of 86 non-yeast species possess an ortholog of a given gene. To count this, we obtained orthogroup designations from InParanoid (Ostlund et al., 2010). For each gene, we considered it to have an ortholog in another species only if it appeared in a cluster with the other species and was given a score of 1.0 by the InParanoid clustering method; that is, we considered a yeast gene to have an ortholog in species *x* if it was a seed gene for a gene cluster that had an orthologous cluster in species *x*. Although Ostlund et al. (2010) considered 100 species, we disregarded the yeast species since the yeast conservation measure already captures information from these species.

CAI

The Codon Adaptation Index, a measure of bias in the usage of synonymous codons, was calculated with the cai tool in the EMBOSS suite (Rice et al., 2000). For each gene, the index is based on a comparison between codon frequencies in the gene and frequencies observed in a set of highly expressed genes; for both *S. pombe* and *S. cerevisiae*, EMBOSS included a default codon usage table that was used.

Copy number

Copy number is a count of the number of paralogs a gene has. This was determined from clusters identified by the InParanoid algorithm (Berglund et al., 2008) run on *S. cerevisiae* and *S. pombe*. All genes that appear in the same cluster were considered copies.

Disorder

The protein disorder measure is the percent of unstructured residues in a gene's protein product as predicted by the Disopred2 software (Ward et al., 2004a).

dN/dS

dN/dS is the ratio between nonsynonymous and synonymous mutations in coding regions of genes. For *S. pombe* genes, dN/dS was calculated twice, using *S. japonicus*, *S. octosporus* as out-group species, and averaged to produce a final dN/dS estimate. Orthologous protein sequences were globally aligned with EMBOSS (Rice et al., 2000) using default parameters. For each *S. pombe* gene, only the out-group ortholog that produced the highest alignment score was used for dN/dS calculations; dN/dS ratios were calculated with the PAML package's implementation of the Yang and Nielsen method for estimating substitution rates (Yang, 2007; Yang and Nielsen, 2000).

Similarly, we computed the average dN/dS ratio for *S. cerevisiae* in comparison to *sensu stricto* yeast species (*S. paradoxus*, *S. bayanus* and *S. mikatae*). Protein sequences were aligned using MUSCLE (Edgar, 2004) and dN/dS ratios were computed using PAML (Yang, 2007).

Number of domains

The number of domains for a gene is the number of regions that Pfam has identified as domains in the protein sequence of the gene. Domain matches for each protein were obtained online from the Pfam database (Punta et al., 2012).

Number of unique domains

Since the same domain is often repeated multiple times in a single protein, this feature modifies number of domains by counting the number of unique domains present in each protein.

Nc

This measure is a simple statistic of codon usage bias and expresses the effective number of codons used in a gene. The chips tool of EMBOSS (Rice et al., 2000) was used to calculate this feature.

Protein length

Protein length is simply the number of amino acids in the corresponding protein.

Co-expression degree

This measure is derived from the co-expression network, the construction of which is described in section 2.8.4. The network contains a level of co-expression for all pairs of genes. We therefore sparsified the network by considering only edges between gene pairs whose co-expression levels were above the 95^{th} percentile. The co-expression degree of a gene is the number of genes with which its co-expression value is retained in this restricted network.

Expression level

Expression levels of all *S. cerevisiae* genes were downloaded from (Holstege et al., 1998). Expression levels of all *S. pombe* genes are measured RNAseq abundance that corresponds to Grabherr et al. (2011) and were downloaded from the Broad Institute's Fungal Genome Initiative website (Broad Institute, 2012).

Expression variation

We estimated the amount of variability in a gene's expression level by measuring the variance of its expression across a number of different microarray experiments, which included microarray data from different growth conditions and replicates. Within each study, we found each gene's percentile of variation. The final value assigned to each gene is its average percentile across all studies. These datasets were obtained from a number of different studies that deposited data in the Gene Expression Omnibus (GEO) (Edgar et al., 2002). *S. pombe* data used in this analysis is the same as those used in construction of the *S. pombe* co-expression network.

Fitness defect

S. pombe fitness defect measurements were obtained by conducting a series of control SGA experiments as described elsewhere (Baryshnikova et al., 2010a; Dixon et al., 2009; Dixon et al., 2008). Briefly, a *S. pombe* SGA query strain harboring a dominant drug-resistance marker (natMX4) inserted at a neutral genomic locus (*h*- leu1∆::*natMX4 ade6-M210 ura4-∆18 leu1-32*), was crossed against the *S. pombe* nonessential deletion mutant collection (*h+ geneX*∆::*kanMX4 ade6-M210 ura4-∆18 leu1-32*). Following mating and sporulation, haploid meiotic progeny harboring both the *kanMX4* and *natMX4* markers are selected and colony sizes are measured after applying standard normalization procedures. We have previously shown that colony sizes derived from these control screens reflect fitness defect of the *kanMX4*-marked single mutant strains that comprise the deletion mutant array. Fitness estimates were based on four control screens as described above and combined with five mutant screens (prz1, res2, SPAC1687.22c, SPCC1682.08, and SPAC6G9.14), which contained the dominant drug-resistance marker (*natMX4*) (Dixon et al., 2008).

S. cerevisiae fitness defect values, defined quantitatively in Baryshnikova et al. (2010b), were published in Costanzo et al. (2010) and experimental procedures are detailed in Baryshnikova et al. (2010a). As in the *S. pombe* protocol described above, SGA was used to insert a neutral query marker into mutant strains so that we could observe colony growth for each mutant in the deletion collection under the effects of only the single deletion. Fitness estimates are based on a large number of replicate screens.

Protein-protein interaction degree

The protein-protein interaction degree of each gene's protein is the number of physical interactions reported in BioGRID, version 2.0.58 (Stark et al., 2006). Interactions considered physical were restricted to those identified by the following terms: Affinity Capture-MS, Affinity Capture-RNA, Affinity Capture-Western, Biochemical Activity, Co-crystal Structure, Co-fractionation, Co-localization, Co-purification, Far Western, FRET, PCA, Protein-peptide, Protein-RNA, Reconstituted Complex, and Two-hybrid.

Multifunctionality

Multifunctionality is a measure of the number of GO terms that are annotated to a gene (Ashburner et al., 2000). From GeneDB (Hertz-Fowler et al., 2004) and Saccharomyces Genome Database (SGD Project, 2010) gene association files (download in November 2009) for *S.pombe* and *S. cerevisiae*, respectively, redundant terms—one term from pairs of terms that are considered "alternative ids"—were removed before totaling the number of GO term annotations for each gene.

A2.3 Genetic interaction degrees

Negative genetic interaction degrees of *S. pombe* genes were derived from interactions reported in Roguev et al. (2008). Only those interactions with S-scores <= -2.5 were considered. This dataset contains 551 genes that are involved in chromosome function; intentionally included are ~100 genes that participate in processes present in both *S. pombe* and human, but importantly, are not present in *S. cerevisiae* (e.g. RNAi machinery).

Negative genetic interaction degrees of *S. cerevisiae* genes were collected from the measurements reported in Costanzo et al. (2010), which screened for interactions involving 3456 array genes, 1438 of which have *S.pombe* orthologs. As suggested by the authors, only negative interactions with an epsilon value of <= -0.08 and a p-value cutoff < 0.05 were considered. This dataset includes degree measurements for most nonessential genes.

A2.4 Orthologs

Orthology mappings (Additional files 4 and 5) are from the InParanoid eukaryotic ortholog database (Berglund et al., 2008). Although the InParanoid algorithm produces clusters, our analysis depends on ortholog pairs. To calculate correlations between *S. cerevisiae* and *S. pombe* for each of the gene features (Figure 2.2A), only genes in one-to-one orthology mappings were used. When holding out orthologs for degree prediction in a set of "species-specific" genes (Figure 2.2C), all genes that had any number of orthologs were removed. Since InParanoid may not report orthologs that other algorithms have detected, we took a conservative approach by additionally removing any genes that had an ortholog in the pombe database GeneDB (Wood, 2006), which includes manually curated orthologs.

Appendix 3: Supplementary items for Chapter 3

Table A3.1. Bicluster size preference tables. Each element is the median of all maximum Jaccard similarities between GO terms and biclusters, calculated for the set of biclusters with specified dimensions.

DMA biclusters, query side

	4-6 query strains	6-7	8-9	10-11	12-13	14-15	16-18	>=19
4-6 array strains	0.0833	0.1176	0.1333	0.129	0.129	0.125	0.1176	0.119
6-7	0.087	0.12	0.1429	0.1389	0.1348	0.1333	0.1333	0.1304
8-9	0.087	0.1319	0.1429	0.1538	0.1429	0.15	0.1429	0.1325
10-11	0.087	0.1429	0.1538	0.1603	0.1538	0.1613	0.118	0.1615
12-13	0.0952	0.1538	0.1923	0.1818	0.1667	0.1333	0.1493	
14-15	0.0945	0.1667	0.1538	0.1667	0.1603	0.1613		
16-18	0.1071	0.1905	0.2	0.1818				
>=19	0.125	0.2174	0.2					

TSA biclusters, query side

	4-6 query strains	6-7	8-9	10-11	12-13	14-15	16-18	>=19
4-6 array strains	0.12	0.1538	0.1538	0.1579	0.15	0.15	0.1538	0.1515
6-7	0.125	0.1538	0.1739	0.1667	0.1739	0.1667	0.1724	0.1667
8-9	0.1176	0.1667	0.1765	0.1739	0.1923	0.1852	0.1724	0.1511
10-11	0.1176	0.1667	0.1818	0.1818	0.245	0.2	0.1765	
12-13	0.1333	0.1818	0.2273	0.2495	0.298	0.1938		
14-15	0.1348	0.2	0.2	0.2174	0.25			
16-18	0.1538	0.2	0.2327	0.2451				
>=19	0.1818	0.2	0.2327					

DMA biclusters, array side

	4-6 query strains	6-7	8-9	10-11	12-13	14-15	16-18	>=19
4-6 array strains	0.1111	0.1053	0.1053	0.1111	0.1053	0.0952	0.0909	0.0769
6-7	0.2	0.1818	0.1667	0.1429	0.1333	0.1379	0.1364	0.12
8-9	0.25	0.2308	0.2	0.2	0.1791	0.1667	0.15	0.131
10-11	0.2857	0.25	0.2308	0.2143	0.1875	0.2	0.1695	0.1387
12-13	0.3333	0.2857	0.25	0.2143	0.1905	0.2083	0.1737	
14-15	0.3333	0.3125	0.2727	0.25	0.2273	0.2174		
16-18	0.3846	0.3333	0.2667	0.3158				
>=19	0.4286	0.35	0.25					

TSA biclusters, array side

	4-6 query strains	6-7	8-9	10-11	12-13	14-15	16-18	>=19
4-6 array strains	0.2222	0.1818	0.1538	0.1304	0.1176	0.1053	0.1	0.0833
6-7	0.2727	0.2105	0.1818	0.1667	0.1538	0.1429	0.125	0.119
8-9	0.3	0.25	0.2222	0.2	0.1875	0.16	0.1659	0.1328
10-11	0.3333	0.2941	0.25	0.2273	0.2183	0.2105	0.2	
12-13	0.4	0.3333	0.2857	0.2667	0.25	0.2106		
14-15	0.4286	0.375	0.3077	0.2727	0.2727			
16-18	0.4444	0.4286	0.3333	0.3				
>=19	0.5385	0.4808	0.375					

Table A3.2. Terms used in the manual annotation (MA) scheme for yeast gene biological-process annotation.

DNA replication & repair/HR/cohesion
Chromatin/transcription
RNA processing
Nuclear-cytoplasmic transport
Ribosome/translation
ER-Golgi traffic
Golgi/endosome/vacuole sorting
Protein folding & glycosylation/cell wall
Protein degradation/proteasome
Signaling/stress response
Cell cycle progression/meiosis
Cell polarity/morphogenesis
Chrom. seg./kinetoch./spindle/microtub.
Metabolism/mitochondria
Amio acid biosynth & transport
Drug/ion transport
Peroxisome
Lipid/sterol/fatty acid biosynth & transport
Autophagy
Highly pleiotropic

Table A3.3. Terms used in the SAFE scheme for yeast gene biological-process annotation.

Cell Polarity and Morphogenesis
Cell polarity and Morphogenesis – Cytokinesis
DNA Replication and Repair
Glycosylation and Protein folding/targeting
Metabolism
Mitochondrial transport, morphology, translation and respiration
Mitosis and Chromosome Segregation
mRNA and tRNA processing
Nuclear-Cytoplasmic Transport
Peroxisome biogenesis and transport
Protein degradation/turnover
Translation - ribosome biogenesis
Translation - rRNA/ncRNA processing
tRNA Wobble Modification
TS Transcription and chromatin organization
Vesicle Traffic
Vesicle Traffic - MVB Sorting and RIM Signaling Pathways

Appendix 4: Supplementary items for Chapter 4

A4.1 Pleiotropy (entropy) scores

Entropy of a strain was calculated from its functional profile as $-\sum_{i=1}^{k} p_i \log_2 p_i$, where p_i is the fraction of biclusters annotated with the i-th process and k is the total number of terms in the annotation scheme. Pleiotropy scores were not assigned to genes that had fewer than 10 annotated biclusters.

A4.2 Supplementary figures and tables

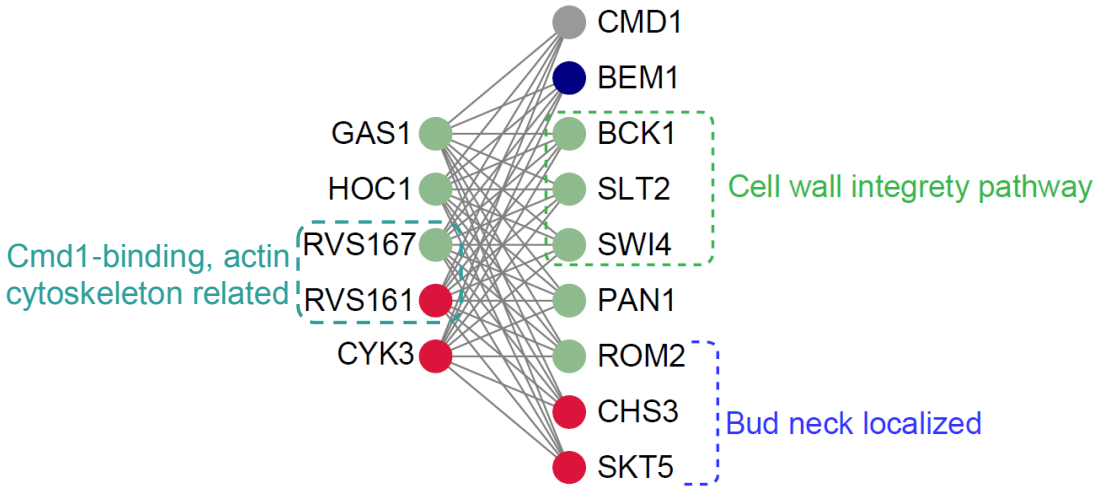


Figure A4.1. Example bicluster that contains CMD1 and was annotated by the bioprocess term "Cell polarity/morphogenesis". Due to a physical interaction with MYO2p, a myosin required for polarized growth, Cmd1 localizes to the bud neck and tip.

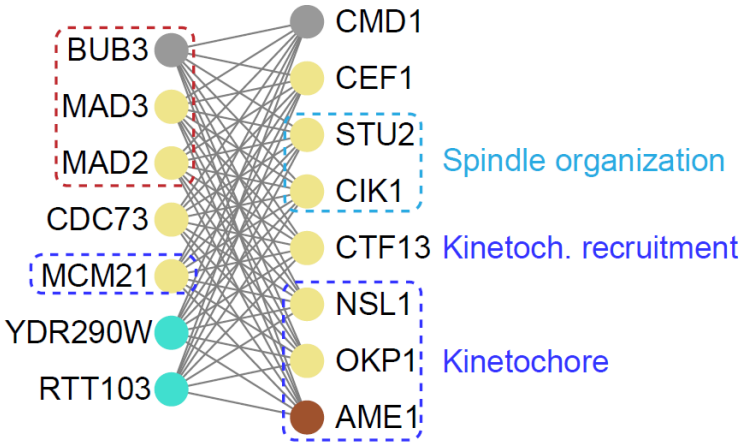


Figure A4.2. Example bicluster that contains CMD1 and was annotated by the bioprocess term "Chrom. seg/kinetoch./etc". Cmd1 is involved in attachment of microtubules to the SPB and is required for correct spindle function.

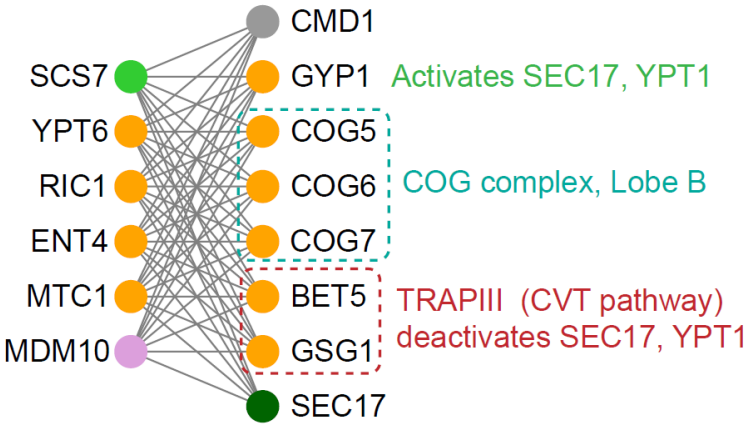


Figure A4.3. Example bicluster that contains CMD1 and was annotated by the bioprocess term "Golgi/endosome/vacuole". Cmd1 is thought to regulate the final stages of vacuolar fusion.

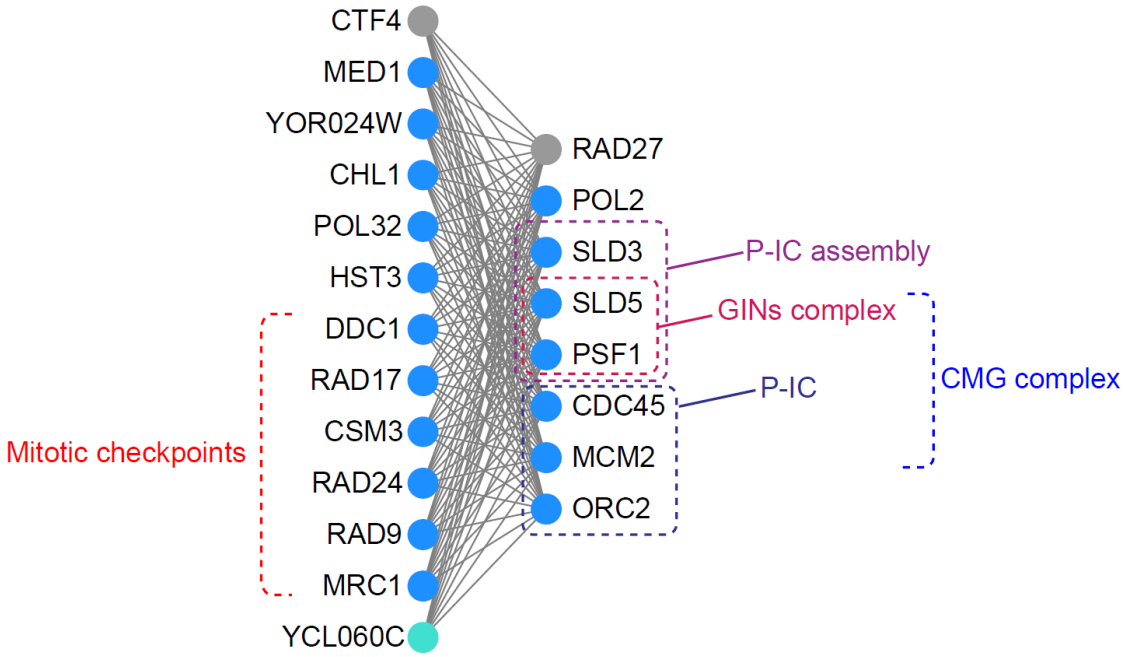


Figure A4.4. Example bicluster that contains RAD27 and complexes involved in the DNA replication fork.

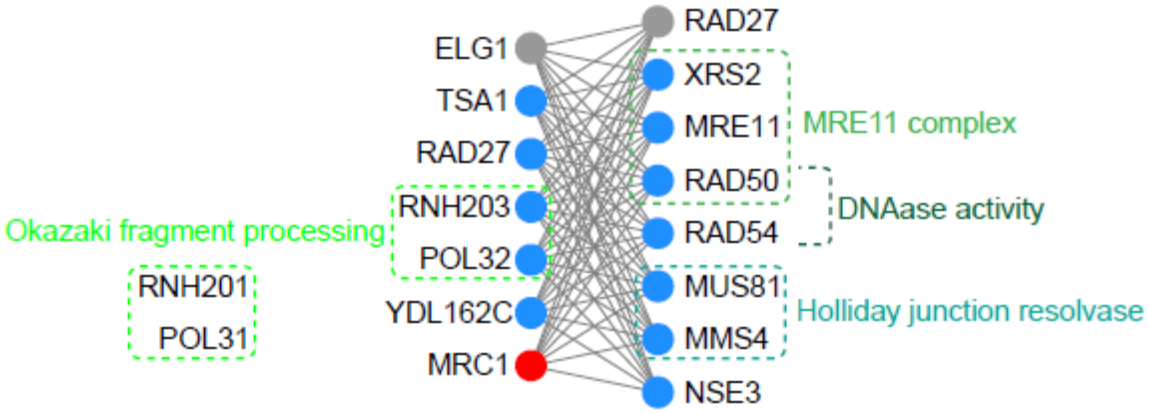


Figure A4.5. Example bicluster that contains RAD27 and complexes involved in Okazaki fragment processing and double-strand break repair. Another RAD27-associated bicluster (not shown) contained the two genes listed on the left.

Table A4.1. Summary of gene characteristics associated with high- and low-pleiotropy genes for "TSA, array" scoring configurations. Tests were performed for pleiotropy scores derived from different pleiotropy scoring configurations (columns). Values shown are the number of rank-sum tests that yielded a significant p-value, out of 12 variations performed (see section 4.9 and Table 4.2). Blank cells indicate zero tests with significant results. Asterisks indicate characteristics that were associated strongly enough with both pleiotropy classes that they are listed in two rows. The significance of p-values from rank-sum tests was determined using the FDR-control procedure described in Benjamini et al. (2006), counting tests for 37 gene properties as a family.

Gene/protein properties	Pleiotropy class with positive association	TSA		
		Array		
		Associate		Adjacent
		MA	SAFE	MA
Copy number volatility	High	1		
Protein abundance in stress	High			4
Copy number	High			
WGD duplicate	High			
Expression var., environ.	High		1	
Protein abundance	High			4
Expression var., genetic-A	High	7		
CAI	High			1
Transcription level	High			4
Expression level	High			4
Protein length	High			
Number of domains	High		4	
Number of unique domains	High		4	
Single mutant fitness defect*	High	5	12	10
Originated in Saccharomyces	High			
Essential*	High			
Expression var., genetic-B*	High			
Complex member	Low	9		
Number of complexes	Low	10		
Curated phenotypes	Low			
Multifunctionality	Low			
Phenotypic capacitance	Low			
PPI degree, Tap MS	Low	12		
Single mutant fitness defect*	Low			
Chemical-genetic degree	Low			
Effective number of codons	Low			
Essential*	Low			
Yeast conservation	Low			
dN/dS	Low			
Broad conservation	Low			
Gene age	Low			
Expression var., genetic-B*	Low			

Table A4.2. Summary of gene characteristics with non-robust associations with high- and low-pleiotropy genes. Details are as in Table 4.1.

Gene/protein properties	Pleiotropy class with positive association	DMA						TSA					
		Query			Array			Query			Array		
		Associate		Adjacent	Associate		Adjacent	Associate		Adjacent	Associate		Adjacent
		MA	SAFE	MA	MA	SAFE	MA	MA	SAFE	MA	MA	SAFE	MA
Coexpression degree	High		4	9						6			
Curated phenotypes	High			11			6						
SSD duplicate	High	2		2	3	2		1		2	2		2
Yeast conservation	High			15									
PPI degree, Tap MS	High			13									
Broad conservation	High			8									
Gene age	High			5									
Deleterious SNP rate of strains	High			1									
Multifunctionality	High			1									
PPI degree, Y2H	High			1									
log2(Distance from telomere)	Low			1				1	12	4			
Copy number volatility	Low			1									
Protein disorder	Low							1		8			
Coexpression degree	Low							8					
Pleiotropy sum	Low			1			6						
Originated in Saccharomyces	Low			4									
Deleterious SNP rate	Low					1							

A4.3 Gene characteristics

Gene age indicates the phylogenetic distance of the most distantly related species with an identified ortholog to a given yeast gene. Genes only found in *S. cerevisiae* are assigned the age of 0 and genes with orthologs appearing in more distant species are assigned higher ages up to 14. Two phylogenetic trees were used in this analysis: one obtained from Ostlund et al. (2010) contains 99 animal, plant, and fungi species and one obtained from Wapinski et al. (2007) contains 23 yeast species.

Broad conservation is a count of how many non-yeast species, out of a set of 86, have an ortholog of a given gene. To count this, we obtained orthogroup designations from InParanoid (Ostlund et al., 2010). For each gene, we considered it to have an ortholog in another species only if it appeared in a cluster with the other species and was given a score of 1.0 by the InParanoid clustering method; that is, we considered a yeast gene to have an ortholog in species x if it was a seed gene for a gene cluster that had an orthologous cluster in species x. Note that this measure is similar, though complementary, to the “yeast conservation” measure described below, which focuses on conservation within the yeast clade.

CAI, codon adaptation index, is a sequence-based measure of bias in usage of synonymous codons as compared to usage in highly expressed genes. It was calculated using the cai tool and the default codon usage table in the EMBOSS suite (Rice et al., 2000).

Chemical-genetic degree is a count of drug and environmental conditions to which a homozygous diploid gene-deletion mutant strain is significantly sensitive (Hillenmeyer et al., 2008).

Co-expression degree is a measure derived from a co-expression network based on integration of a large collection of expression datasets (Huttenhower et al., 2006). The network was sparsified by considering only edges between gene pairs whose co-expression levels were above the 95th percentile. The co-expression degree of a gene is the number of genes with which its co-expression value is retained in this restricted network.

Complex member is a binary feature that indicates whether the corresponding protein is a component of at least one complex based on the complex standard provided in Costanzo et al. (2016).

Copy number is a count of the number of paralogs each gene has. This was determined from clusters identified by the InParanoid algorithm (Ostlund et al., 2010). All genes that appeared in the same cluster were considered paralogs.

Copy number volatility is the number of times that a gene is lost or gained among 23 Ascomycete fungi species, as defined by Wapinski et al. (2007).

Curated phenotypes is the number of mutant phenotypes associated with a nonessential gene’s deletion strain. Mutant phenotypes were downloaded from the Saccharomyces Genome Database (SGD) on January 31, 2013. The list of phenotypes was filtered to include only those related to deletion mutants of verified or uncharacterized open reading frames (mutant type = 'null', feature type = 'ORF'). Phenotypes were further filtered to only include increased or decreased phenotype

expression compared to a wild-type strain. Finally, the number of non-wild-type phenotypes was counted for each gene. Unclear descriptions of phenotypes, such as "abnormal", were ignored.

Deleterious SNP rate is the number of predicted deleterious SNPs observed for a given gene in a diverse set of sequenced *S. cerevisiae* strains (Liti et al., 2009) normalized by gene length. **Deleterious SNP rate of strains** is similar, but counts strains containing deleterious SNPs. These SNP features were derived from identification and analysis of SNPs in 19 strains as described in (Jelier et al., 2011). Briefly, SNPs were identified from sequence alignments of all strains to the S288C reference sequence. The SIFT algorithm, with some modifications, was used to predict which nonsynonymous SNPs are likely to have functional consequences. We applied the recommended threshold to SIFT scores, calling any SNP with a score of <= 0.05 deleterious.

dN/dS is the ratio of the number of nonsynonymous to synonymous mutations in a given gene. We computed the average dN/dS ratio for *S. cerevisiae* in comparison to the sensu stricto yeast species (*Saccharomyces paradoxus*, *Saccharomyces bayanus*, and *Saccharomyces mikatae*). Protein sequences were aligned using MUSCLE and dN/dS ratios were computed using PAML (Edgar, 2004; Yang, 2007).

Effective number of codons is a measure of codon usage bias and is an alternative to CAI that does not require a pre-defined set of highly expressed genes. This measure was computed using PAML (Yang, 2007).

Essential, a binary feature, is true for any gene that is required for viability under standard laboratory conditions.

Expression level is a measurement of the mRNA expression level of a gene (Holstege et al., 1998).

Expression variance, environ. is the variance in a gene's expression across all measurements in the Gasch et al. (2000) dataset. This study subjected yeast to many environmental conditions and measured expression of nearly all yeast genes with

microarrays. Environments included heat shock, hydrogen peroxide, superoxide generated by menadione, diamide, dithiothreitol, hyper-osmotic shock, amino acid starvation, nitrogen source depletion, and progression into stationary phase, as well as alternative carbon sources and variable temperatures. The data contain multiple time points and temperatures for the environments listed.

Expression variance, genetic-A is the variance of expression for each yeast gene measured across a set of strains including BY4716, RM11-1a, and 112 segregants from crosses between BY4716 and RM11-1a. This reflects variation that occurs in genetically diverse genomic backgrounds. The expression data set was produced by Brem and Kruglyak (2005) using DNA microarrays.

Expression variance, genetic-B is the variance of expression for each yeast gene measured across 22 strains from geographically and environmentally diverse locations. This reflects expression variation that occurs in genetically diverse genomic backgrounds. The expression data set was produced by Skelly et al. (2013) using RNA-seq.

log2(Distance from telomere) is the distance, in nucleotides and log-transformed, between a gene and the start of a its closest telomere.

Multifunctionality is a count of annotations to “biological process” terms of the Gene Ontology (Ashburner et al., 2000). Specifically, it is the total number of annotations across a set of functionally distinct GO terms described in Myers et al. (2006).

Number of complexes is a count of the number of complexes by which a given gene is annotated in the protein complex standard provided in Costanzo et al. (2016).

Number of domains is the number of domains, counting repeated domains, present within a given protein, as identified by PFAM (Finn et al., 2013)(downloaded July 2015). **Number of unique domains** is the same but does not count repeated domains.

Originated in Saccharomyces is a binary value that is true if a gene originated in the *Saccharomyces* clade of the phylogenetic tree, which is assumed if the most distant species with an ortholog is a *Saccharomyces* yeast species. Specifically, we consulted the species tree from Wapinski et al. (2007) and identified all genes that appear only in *S. cerevisiae*'s closest relatives: species up to and including *Saccharomyces bayanus*. Note that although some more distant species (*Naumovozyma castellii*, *Lachancea kluyveri*) were originally placed in the genus *Saccharomyces* and may still be referred to with this name as described in (Wapinski et al., 2007b), these have subsequently been associated with different genera.

Phenotypic capacitance was computed by Levy and Siegal (2008) and captures variability across a range of morphological phenotypes upon deletion of a nonessential gene.

PPI degree, Tap MS is the total number of protein-protein interactions in the union of the two data sets from Gavin et al. (2006) and Krogan et al. (2006), which both performed experiments using Tandem Affinity Purification coupled with Mass Spectrometry.

PPI degree, Y2H is the total number of binary, physical interactions detected using yeast two-hybrid analysis (Yu et al., 2008).

Protein abundance was measured by fluorescence of GFP-tagged proteins grown in liquid rich media; **protein abundance under stress** was measured by fluorescence of GFP-tagged proteins grown in liquid minimal media (Newman et al., 2006).

Protein disorder is the percent of unstructured residues as predicted by the Disopred2 software (Ward et al., 2004).

Protein length is the number of amino acids in a gene's encoded protein.

Single mutant fitness defect was calculated by Costanzo et al. (2016) and is the decrease in the growth of a single-gene mutant strain as compared to wild type.

SSD duplicate, a binary feature, is true for genes with one or more paralogs that resulted from small scale duplication (SSD) events. To identify pairs of genes that emerged from SSD events, VanderSluis et al. (2010) searched for gene pairs that meet the following criteria: the gene pair must have a sufficiently high sequence similarity score (FASTA Blast, E = 10), sufficient protein alignment length (> 80% of the longer protein), an amino acid level identity of at least 30% for proteins with aligned regions longer than 150 amino acids or greater than [0.01n + 4.8L^(-0.32(1 + exp(-L/1000)))] with L defined as the aligned length and n = 6 for shorter proteins (Gu et al., 2002; Rost, 1999).

Transcription level is the average measured number of mRNA copies of each transcript per cell (Holstege et al., 1998).

WGD duplicate, a binary feature, is true for any gene that has a paralog that resulted from the whole genome duplication event. The WGD event designation was reconciled from several sources (Byrne and Wolfe, 2005).

Yeast conservation counts how many of 23 different species of Ascomycota fungi possess an ortholog of a gene. This measure was described by Wapinski et al. (2007) and ortholog data were downloaded from the associated website http://www.broadinstitute.org/regev/orthogroups/.