The Dissertation Committee for Kriston Lyle McGary Certifies that this is the approved version of the following dissertation:

The Functional Network in Predictive Biology: Predicting Phenotype from Genotype and Predicting Human Disease from Fungal Phenotype

Committee:

Edward M. Marcotte, Supervisor

James J. Bull

Vishwanath R. Iyer

Arlen W. Johnson

Scott W. Stevens

The Functional Network in Predictive Biology: Predicting Phenotype from Genotype and Predicting Human Disease from Fungal Phenotype

by

Kriston Lyle McGary, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2008

Dedication

To my wife, Eryn, who shares my hopes and visions
and has kept them strong in the face of life's challenges.

To my parents, who taught me that hope and vision are worth having
and, perhaps more importantly, worth sharing.

To my Wise mentor, who helped me develop a vision for biology
and the study of the rest of life, too.

Acknowledgements

I thank the past and present members of the Marcotte lab who have contributed to my work, particularly those whose contribution have been most direct: Ram Narayanaswamy and Wei Niu, for help with microscopy, and Tae Joo Park, for the frog expertise that got me out the door. I also thank the many in the lab that have given me thoughtful feedback about my work or suggestions for thriving in academia: Insuk Lee, Christine Vogel, Dan Boutz, Arun Ramani, Mark Carlson, John Prince and Traver Hart. I also thank Mark Tsechansky, whose ideas about aggregation will stick with me.

I thank Arlen Johnson and his lab for their help and support while struggling through yeast molecular biology. Even computational biologists need to get their hands dirty occasionally and you helped me learn a lot. I also thank Scott Stevens and his lab for help and bench space while I was working on protein purification on a project that seems so very long ago. I thank Vishwanath Iyer for providing a summer home to a wandering first-year graduate student.

I thank Edward Marcotte, who loves research and passes that on to those around him. In particular, I thank him for giving me room to fail and opportunities to succeed. I appreciated the chance to pursue projects just because I found them interesting. And, as this dissertation shows some of those interesting projects made it to publication, too.

The Functional Network in Predictive Biology: Predicting Phenotype from Genotype and Predicting Human Disease from Fungal Phenotype

Publication No._____

Kriston Lyle McGary, Ph.D.

The University of Texas at Austin, 2008

Supervisor: Edward M. Marcotte

The ability to predict is one of the hallmarks of successful theories. Historically, the predictive power of biology has lagged behind disciplines like physics because the biological world is complex, challenging to quantify, and full of exceptions. However, in recent years the amount of available data has expanded exponentially and biological predictions based on this data become a possibility. The functional gene network is a quantitative way to integrate this data and a useful framework for making biological predictions. This study demonstrates that functional networks capture real biological insight and uses the network to predict both subcellular protein localization and the phenotypic outcome of gene knockouts. Furthermore, I use the functional network to evaluate genetic modules shared between diverse organisms that lead to orthologous phenotypes, many that are non-obvious. I show that the successful predictions of the functional network have broad applicability and implications that range from the design of large-scale biological experiments to the discovery of genes with potential roles in human disease.

Table of Contents

List of Tables x

List of Figures xi

Chapter 1: Introduction to predictive biology and functional networks 1

Predicting the nuclear export adaptor of the ribosome small subunit. 2

Method for predicting the SSU adaptor 3

Approach to experimental testing of SSU adaptor predictions 4

Experimental results 5

Evaluation of first predictive effort 5

New quantitative, predictive approach with the functional network 6

Applications of the functional network and new predictive approaches 7

Chapter 2: Predicting protein localization to the yeast shmoo 11

Introduction 11

Methods 14

Classifier Construction 14

Testing predicted shmoo genes 15

Results 16

Shmoo localized proteins can be predicted from their functional linkage to known shmoo proteins. 16

Adaptive re-use of polarization proteome 19

Conclusion 26

References 27

Chapter 3: Predicting yeast knockout phenotypes with a functional network 29

Introduction 29

Background 29

Methods 33

Assembling the set of non-redundant loss-of-function phenotypes 33

Prediction of phenotypes and evaluation of prediction quality 34

Prediction of human disease gene sets....36
Generation of random phenotype sets....36
Yeast strains, media, and growth....37
Results....38
Computational Results....38
Using Guilt-By-Association to Predict Essentiality....38
A yeast gene network predicts varied, specific loss-of-function phenotypes....40
Integration of functional genomics and proteomics data is important for phenotype prediction....50
Prediction of quantitative cell morphology phenotypes....57
Genes increasing cell-to-cell variation are less functionally coherent than those decreasing variation....62
The functional network predicts yeast orthologs of human disease genes....68
Experimental Results....71
Extending a genetic screen by network-guided reverse genetics....71
Discussion....76
Why are loss-of-function phenotypes predictable?....77
AUC is a useful measure of gene functional coherence....79
Recapitulation of the classic mutator phenotype in the yeast knockout collection....80
Applying network-based phenotype prediction to humans and other organisms....81
Conclusions....82
References....83
Chapter 4: Predicting and testing human disease genes in model organisms by finding equivalent phenotypes between species....91
Introduction....91
Identifying and expanding models of human disease....93
Identifying equivalent phenotypes (phenologs)....93
Finding genes in model organism for orthologous phenotype....97
Testing candidate genes....98

Methods....100
Collection of phenotypes....100
Identification of non-redundant phenotype sets....101
Calculating Orthologs....102
Calculation of phenologs....103
Tests of sub-network modularity....104
Treatment of animals....106
Xenopus laevis embryo manipulations....106
Confocal imaging....107
Morpholino oligonucleotides and cDNA clones....107
Results....109
Computational Results....109
Phenologs identifies obviously equivalent phenotypes....113
Techniques developed for identifying homologous genes can be applied to phenologs....114
Phenologs identify dense subnetworks in functional network..114
Experimental Results....120
Experimental confirmation of a yeast model for vertebrate angiogenesis....120
Experimental confirmation of a worm model for neural tube defects....127
Discussion....132
Conclusion....134
References....136
Chapter 5: Putting the pieces together....140
References....142
Vita....154
Publications....154

List of Tables

TABLE 2.1: MANUALLY VERIFIED SHMOO TIP LOCALIZED GENES IDENTIFIED BY THE CELL CHIP 17

TABLE 2.2: MANUALLY VERIFIED SHMOO TIP LOCALIZED PROTEINS IDENTIFIED BY THE CLASSIFIER 18

TABLE 3.1. PREDICTABILITY OF 100 YEAST GENE DELETION PHENOTYPES 43

TABLE 4.1 EXAMPLES FROM THE >6,000 SIGNIFICANT PHENOLOGS DETECTED 117

List of Figures

FIGURE 2.1 SCHEMATIC OF CELL CHIP METHOD WITH PREDICTIONS 13

FIGURE 2.2 A. MANUAL VALIDATION BASED ON CLASSIFIER DOUBLES COVERAGE OF KNOWN SHMOO GENES. B. GENES INVOLVED IN SHMOO FORMATION OVERLAP SIGNIFICANTLY WITH BUDDING GENES 21

FIGURE 2.3 IMAGES OF VARIOUS SHMOO LOCALIZED CELLULAR COMPONENTS 22

FIGURE 2.4 EXOCYTOTIC PROTEINS ARE MORE DISTALLY LOCATED THAN ENDOCYTOTIC PROTEINS IN THE SHMOO 23

FIGURE 2.5 THE FUNCTIONAL NETWORK SUGGESTS POSSIBLE ROLES FOR UNKNOWN PROTEINS LOCALIZED TO THE SHMOO 25

FIGURE 3.1 OVERVIEW OF GUILT-BY-ASSOCIATION PHENOTYPE PREDICTION. 32

FIGURE 3.2 DIVERSE YEAST GENE LOSS-OF-FUNCTION PHENOTYPES ARE PREDICTABLE USING GUILT-BY-ASSOCIATION IN A FUNCTIONAL GENE NETWORK 41

FIGURE 3.3 LOSS-OF-FUNCTION PHENOTYPES ARE PREDICTED SIGNIFICANTLY BETTER THAN RANDOM EXPECTATION 48

FIGURE 3.4 A PLOT OF SEED SET SIZE VERSUS PREDICTABILITY OF THE PHENOTYPE SHOWS NO SIGNIFICANT CORRELATION. 49

FIGURE 3.5 FUNCTIONAL NETWORKS HAVE GREATER PREDICTIVE POWER FOR PHENOTYPE THAN PHYSICAL PROTEIN NETWORKS. 51

FIGURE 3.6 PREDICTIVE POWER OF FUNCTIONAL NETWORK RELIES ON PHYSICAL AND FUNCTIONAL INFORMATION. 54

FIGURE 3.7 LOWER PROBABILITY LINKAGES CONTINUE TO IMPROVE PREDICTIVE ACCURACY, ALBEIT WITH DIMINISHING RETURNS 55

FIGURE 3.8 ITERATIONS OF THE FUNCTIONAL NETWORK IMPROVE PHENOTYPE PREDICTION. 56

FIGURE 3.9 NETWORK-BASED PREDICTIONS OF QUANTITATIVE CELL MORPHOLOGY PHENOTYPES. A WIDE VARIETY OF PHENOTYPES BASED UPON QUANTITATIVE YEAST CELL SHAPE AND INTRACELLULAR FEATURES ARE PREDICTABLE 60

FIGURE 3.10 PREDICTIONS OF QUANTITATIVE CELL MORPHOLOGY PHENOTYPES ARE SIGNIFICANTLY BETTER THAN RANDOM. 61

FIGURE 3.11 GENES WHOSE DISRUPTION DECREASES POPULATION CO-EFFICIENT OF VARIANCE (CV) ARE ESSENTIALLY RANDOM. 64

FIGURE 3.12 GENES KNOCKOUTS THAT INCREASE VARIANCE ACROSS MANY MORPHOLOGICAL TRAITS TYPICALLY AFFECT GENOMIC STABILITY 67

FIGURE 3.13 YEAST GENES WHOSE HUMAN ORTHOLOGS ARE LINKED TO THE SAME DISEASES ARE PREDICTED SIGNIFICANTLY BETTER THAN RANDOM EXPECTATION 70

FIGURE 3.13 NETWORK-GUIDED EXTENSION OF A GENETIC SCREEN 73

FIGURE 3.14 NETWORK CONNECTIVITY PREDICTS GENES INVOLVED IN CELL ELONGATION 75

FIGURE 4.1 THE RATE OF ASSOCIATING GENES TO ORGANISM-LEVEL PHENOTYPES IN MODEL ORGANISMS GREATLY EXCEEDS THAT IN HUMANS 95

FIGURE 4.2 PHENOLOGS CAN BE IDENTIFIED BASED ON SIGNIFICANTLY OVERLAPPING SETS OF ORTHOLOGOUS GENES 96

FIGURE 4.3 AN EXAMPLE OF A PHENOLOG MAPPING HIGH INCIDENCE OF MALE C. ELEGANS PROGENY TO HUMAN BREAST/OVARIAN CANCERS 99

FIGURE 4.4 SYSTEMATIC IDENTIFICATION OF PHENOLOGS 110

FIGURE 4.5 MANY MORE ORTHOLOGOUS PHENOTYPES ARE OBSERVED THAN EXPECTED BY RANDOM CHANCE 111

FIGURE 4.6 COUNT OF PHENOLOGS ABOVE A FALSE DISCOVERY RATE THRESHOLD 112

FIGURE 4.7 GENES INVOLVED IN PHENOLOGS SHOW ENHANCED INTERCONNECTIVITY IN GENE NETWORKS 119

FIGURE 4.8 EXAMPLE OF A NON-OBVIOUS DISEASE MODEL REVEALED BY PHENOLOGS: YEAST MUTANTS SENSITIVE TO THE HYPERCHOLESTEROLEMIA DRUG LOVASTATIN PREDICT MAMMALIAN ANGIOGENESIS DEFECTS 121

FIGURE 4.9 IN SITU HYBRIDIZATION SHOWS XSOX12 EXPRESSION IN VEINS AND DEVELOPING HEART OF A STAGE 32 XENOPUS EMBRYO 124

FIGURE 4.10 MORPHOLINO (MO) KNOCKDOWN OF XSOX12 INDUCES DEFECTS IN VASCULATURE 125

FIGURE 4.11 HEMORRHAGING IS APPARENT IN STAGE 45 XENOPUS EMBRYOS DUE TO DYSFUNCTIONAL VASCULATURE FOLLOWING XSOX12 MORPHOLINO KNOCKDOWN 126

FIGURE 4.12 SCHEMATICALLY REPRESENTATION OF THE VALIDATION OF TWO NEW NEURAL TUBE DEFECT GENES PREDICTED BY PHENOLOGS AND GENE NETWORKS 129

FIGURE 4.13 MORPHOLINO KNOCKDOWNS OF XENOPUS GENES RFX2 AND IFT140 SHOW STRONG NEURAL TUBE DEFECTS 130

FIGURE 4.14 RFX2-MO KNOCKDOWN ANIMALS SHOWS THAT CILIATED CELLS ARE INTACT, BUT LACK CILIA 131

FIGURE 4.15 PROPOSED MODEL TO EXPLAIN GREATER FUNCTIONAL COHERENCE AMONG ORTHOLOGS INVOLVED IN BOTH PHENOTYPES RELATIVE TO ORTHOLOGS INVOLVED IN A SINGLE PHENOTYPE 133

Chapter 1: Introduction to predictive biology and functional networks

In science, predictive, quantitative theories and models are preferred over post-hoc, qualitative theories, because they are both useful for guiding further research and indicate that our models are not just rationalizing data, but providing genuine insight into the workings of the system under question. Historically, physics has been the field of science best known for its ability to predict the phenomena in its domain; however, biologists also value predictive models and in recent years a renewed emphasis on prediction has accompanied the rise of systems biology [1]. Some efforts have used explicit modeling of physics and chemistry, using diffusion and reaction rates to model biological systems. One classic example is the von Dassow et al. computational model of the formation of the segment polarity stripes during early Drosophila development, which incorporated known molecular details and showed that the model recapitulated the biology over a wide range of values for parameters that were unknown [2]. Unfortunately, this type of explicit modeling is limited to well studied systems. Perhaps the longest running collaborative effort for biologically relevant predictions is in the area of predicting protein structure. Since 1994, a series of competitions, Critical assessment of techniques for protein Structure Prediction (CASP), have evaluated the performance of many different approaches to predicting protein structure including: *ab initio* approaches using molecular dynamic simulations, homology based structure prediction, and fold recognition [3]. More recently, Scott et al. attempted to predict subcellular protein localization using a customized Bayesian classifier [4] and previously known data [5, 6]. Finally, the entire field of functional genomics sprung up to predict the function of the large number of new genes identified by genome sequencing by extrapolating from known functional data.

Predictions in biology remain challenging. Part of the challenge is to identify the areas of biology that are tractable for prediction and that will make the greatest contribution to our understanding of life. As with other disciplines, it is very important to identify the optimal scope of the research, aim for incremental progress, and make readily testable hypotheses. In the following, I will discuss my first effort to test a reasonably ambitious prediction, which failed, and evaluate it by these criteria. I will end the chapter with a discussion of the conclusions that guided my future attempts at prediction.

PREDICTING THE NUCLEAR EXPORT ADAPTOR OF THE RIBOSOME SMALL SUBUNIT.

The formation of the ribosome is a complicated, multi-step process that starts in the nucleus and finishes in the cytoplasm. The ribosome's two subunits, the large subunit (LSU) and the small subunit (SSU), start as a single nuclear RNA transcript, which is cleaved into two pieces. Each RNA molecule separately undergoes further processing and the integration of structural proteins within the nucleus. After key steps are finished, the subunits are exported to the cytoplasm for further processing; however, until recently, the nuclear export adaptors for both subunits were unknown. Nuclear export adaptors physically link their cargo to the export machinery of the nuclear pore, typically CRM1, to provide directed transport across the nuclear membrane. In 2000, Ho *et al*. reported the discovery of the yeast nuclear export adaptor for the ribosome large subunit (LSU), NMD3 [7]. Thus far, the nuclear export adaptor for the ribosome SSU has not been identified. So, I decided to test the hypothesis that the ribosomal SSU nuclear adaptor can be predicted based on properties of the LSU adaptor. I collaborated with Arlen Johnson, who had originally identified the LSU adaptor, on this project.

Method for predicting the SSU adaptor

My prediction strategy was based on the assumption that key properties of the LSU nuclear adaptor, NMD3, can serve as a guide to identify SSU export adaptor candidate genes. Several properties of NMD3 serve as logical starting points for finding the SSU adaptor. NMD3 is essential, as would be expected of a gene critical to the function of core cellular machinery. NMD3p has a classical nuclear export sequence (NES), which is known to interact with CRM1 during export. NMD3 is a conserved protein across archeal and eukaryotic organisms, with an additional domain in eukaryotes containing the NES that is not present in the archea, which lack a nuclear membrane. Furthermore, NMD3 contained a nuclear localization sequence (NLS), which allows it to recycle to the nucleus after delivering the nascent SSU to the cytoplasm.

I created a flexible scoring scheme that integrated multiple criteria, since it was unclear which criteria would be most useful for identifying the SSU: essentiality, presence of a nuclear export sequence, links to the SSU and/or nuclear export machinery in the functional network, the distribution of the protein across eukarya and archea, and haploinsufficiency. I integrated the scoring scheme into a web-based application to simplify the exploration of various combinations and weighting of criteria and to facilitate collaboration. In addition, I created a simple tool to compare the results from various weightings of the evidence. The underlying data was obtained from the following: essentiality and haploinsufficiency from SGD [8], nuclear export sequence prediction from NetNES [9], functional links from a pre-publication version of YeastNet v.2. [10], and protein conservation using Blast (data provided by Insuk Lee).

I explored various combinations of requirements and scoring schemes. Across a wide range of parameters, NMD3 was recovered as a top candidate, which suggested that our approach could identify the SSU adaptor if it resembles the LSU adaptor in its mode of action.

Approach to experimental testing of SSU adaptor predictions

After comparing the results from multiple criteria weightings, we hand selected five top candidates, including NOP1, PNO1, SUI3, for initial testing, with plans to screen additional candidate genes once the screening approach was validated and tested for scalability. We assayed each gene for nuclear export activity in the following manner. We engineered dominant negative versions of each gene with a defective nuclear export sequences. If the candidate is the SSU adaptor, over-expression of the dominant negative version is expected to block export of the SSU, shifting the distribution of GFP labeled SSUs from the nucleolus and cytoplasm to the nucleus.

For each gene, the dominant negative was constructed in two PCR steps, followed by cloning into an inducible over-expression vector. In the first step, the 5' end of each gene was amplified using a forward primer containing an enzyme restriction site and a reverse primer that mutated nucleotides in the NES to mutate functionally important codons from leucine to alanine. The 3' end was amplified in a similar manner, with the forward primer designed to match the mutant nucleotides at the NES and the reverse primer containing a restriction site compatible with the expression vector. In the second step, the PCR products of the first reactions were fused by amplifying with the forward primer of 5' reaction and the reverse primer of the 3' reaction. Only molecules resulting from the overlap of the mutated NES region are amplified. The mutant gene was then

cloned into a galactose inducible over-expression vector and co-transformed into yeast with a plasmid expressing a GFP tagged SSU protein. Transformed strains were incubated in galactose media prior to microscopy to activate over-expression of the dominant negative. A strain transformed with mutant NMD3 and a GFP tagged LSU protein served as a positive control.

Experimental results

Four of the most promising candidates were screened and found not to affect the distribution of the SSU. The mutant form of NOP1 appeared to affect growth, so a rescue experiment was performed by adding a potent NES to the N-terminus of the protein. The exogenous NES failed to rescue the growth phenotype, which suggests that the cause of the mutant phenotype is the destabilization of protein structure rather than the disruption of nuclear export.

Evaluation of first predictive effort

Upon reviewing the data, the methodological challenges and timeline, I decided that extending the screen to a larger number of genes would be a large investment with an ambiguous conclusion. The predictions were highly dependent on the assumption that the SSU adaptor would operate with the same basic mechanism as the LSU adaptor, which was reasonable, but not guaranteed. Unfortunately, in the case, the approach would not work unless we precisely identified both the specific gene and specific mechanism. Interestingly, recent work has suggested that one of the predicted genes, PNO1 (also known as DIM2), may be involved in nuclear export, but its role as the adaptor has not yet been proven [11].

Given the challenge of this approach, it become clear that the predictive power of the functional network should be tested in system that is easy to assay, which has more than a single target, and, requires a minimal number of assumptions and auxiliary data. The final requirement was needed so that the method developed would be broadly applicable rather than being rendered obsolete by its own success.

NEW QUANTITATIVE, PREDICTIVE APPROACH WITH THE FUNCTIONAL NETWORK

In my attempt to find a more broadly applicable approach to prediction, I chose one of the most generally applicable tools for inferring the function of unknown genes, which was first developed by Lee *et al*. in 2004 [12]. Their probabilistic functional network uses a Bayesian framework to integrate many types of biological datasets (e.g. protein interactions, transcriptional co-regulation, and gene fusions). Abstracting the specific nature of interactions allows network to report the probability that two genes are functionally related without specifying the precise nature of the association. Each data set integrated into the network is evaluated for quality by calculating the likelihood that two genes in a pathway will be linked by the evidence. The cumulative evidence from multiple data sets can lead to well supported linkages that are poorly supported in any one experiment. The network has already proven useful for predicting the function of genes involved in chromatin modification and ribosome biogenesis [12-14]. Furthermore, the functional network provides an intuitive conceptual framework with the potential for novel application. In order to make other predictions with the network, we took advantage of an established principle for inferring gene function from network connections, the principle of guilt-by-association (GBA). In GBA, the function of uncharacterized genes is inferred from the functions of characterized neighbors in the network ([15-17]; reviewed in [18]). The functional network is weighted, so, given a set

of genes with a property of interest, e.g. subcellular localization or knockout phenotype, other genes can be predicted based on the sum of their weights to that set. I will provide greater detail for this general approach and specific applications in the following chapters.

Applications of the functional network and new predictive approaches

I have successfully tested two different applications of the functional network (discussed in chapter 2 and 3) and have leveraged it to help understand a novel predictive method which I will present in chapter 4.

In chapter 2, I show that the functional network is a useful predictor of protein localization. After a large genome wide screen for proteins localized to the yeast shmoo tip missed a large number of known genes, I developed a classifier for predicting additional proteins localized to the shmoo tip. The initial set of 37 proteins from large screen was used to train a classifier and the top predictions were tested by hand in a small manual screen. The classifier-guided retesting strategy doubled the coverage of the screen and remaining false negatives could be rationalized based upon low protein abundance.

In chapter 3, I show proof-of-principle that genes linked in a functional network are likely to give rise to the same loss-of-function phenotype, demonstrating efficacy for predicting yeast mutant phenotypes. I show that diverse yeast gene loss-of-function phenotypes are predictable, from biochemical to morphological to fitness effects. The approach I describe provides a rational and quantitative foundation for targeted reverse genetic studies, which I demonstrate by predicting, then verifying, essential genes whose

disruption produces elongated yeast cells. The breadth of applicability suggests that this approach could be implemented to identify genes likely to lead to human disease by leveraging extensive functional genomics data to expand sets of known disease genes by predicting new candidate disease genes.

Finally, in chapter 4, I predict phenotype through a novel method that identifies equivalent phenotypes between species. Mapping between genotype and phenotype is often non-obvious, complicating prediction of genes underlying specific phenotypes. I address this problem through comparative analyses of phenotypes. I define orthologous phenotypes between organisms (phenologs) based upon overlapping sets of orthologous genes associated with each phenotype. Genes known to have a phenotype in one organism become predictions for having the orthologous phenotype in the other organism. Comparisons of >189,000 human, mouse, yeast, and worm gene-phenotype associations reveal many significant phenologs, including novel non-obvious human disease models. For example, phenologs suggest a yeast model for mammalian angiogenesis defects and an invertebrate model for vertebrate neural tube birth defects. With collaborators, we use the former to discover that SOX13 regulates vertebrate angiogenesis; with the latter, we demonstrate that IFT140 and RFX2 knockdowns cause neural tube defects. Phenologs create a rich framework for comparing mutational phenotypes, identifying adaptive reuse of gene systems, and predicting new disease genes.

Bibliography

1. Liu ET: Systems biology, integrative biology, predictive biology. *Cell* 2005, 121:505-506.

2. von Dassow G, Meir E, Munro EM, Odell GM: The segment polarity network is a robust developmental module. *Nature* 2000, 406:188-192.

3. John Moult JTPRJKF: A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Genetics* 1995, 23:ii-iv.

4. Scott MS, Calafell SJ, Thomas DY, Hallett MT: Refining protein subcellular localization. *PLoS Comput Biol* 2005, 1:e66.

5. Kumar A, Cheung KH, Ross-Macdonald P, Coelho PS, Miller P, Snyder M: TRIPLES: a database of gene function in Saccharomyces cerevisiae. *Nucleic Acids Res* 2000, 28:81-84.

6. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: Global analysis of protein expression in yeast. *Nature* 2003, 425:737-741.

7. Ho JH, Kallstrom G, Johnson AW: Nmd3p is a Crm1p-dependent adapter protein for nuclear export of the large ribosomal subunit. *J Cell Biol* 2000, 151:1057-1066.

8. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M *et al*: SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 1998, 26:73-79.

9. la Cour T, Kiemer L, Molgaard A, Gupta R, Skriver K, Brunak S: Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng Des Sel* 2004, 17:527-536.

10. Lee I, Li Z, Marcotte EM: An Improved, Bias-Reduced Probabilistic Functional Gene Network of the Baker's Yeast, Sacchromyces cerevisiae. *PLOS One* 2007.

11. Vanrobays E, Leplus A, Osheim YN, Beyer AL, Wacheul L, Lafontaine DL: TOR regulates the subcellular distribution of DIM2, a KH domain protein required for cotranscriptional ribosome assembly and pre-40S ribosome export. *Rna* 2008, 14:2061-2073.

12. Lee I, Date SV, Adai AT, Marcotte EM: A probabilistic functional network of yeast genes. *Science* 2004, 306:1555-1558.

13. Combs DJ, Nagel RJ, Ares M, Jr., Stevens SW: Prp43p is a DEAH-box spliceosome disassembly factor essential for ribosome biogenesis. *Mol Cell Biol* 2006, 26:523-534.

14. Lee I, Li Z, Marcotte EM: An Improved, Bias-Reduced Probabilistic Functional Gene Network of Baker's Yeast, Saccharomyces cerevisiae. *PLoS ONE* 2007, 2:e988.

15. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: A combined algorithm for genome-wide prediction of protein function. *Nature* 1999, 402:83-86.

16. Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T: Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res* 1999, 9:1198-1203.

17. Schwikowski B, Uetz P, Fields S: A network of protein-protein interactions in yeast. *Nat Biotechnol* 2000, 18:1257-1261.

18. Sharan R, Ulitsky I, Shamir R: Network-based prediction of protein function. *Mol Syst Biol* 2007, 3:88.

Chapter 2: Predicting protein localization to the yeast shmoo

INTRODUCTION

Cells coordinate a large scale re-arrangement of their internal machinery when they undergo polarized growth. Polarized growth is a fundamental, highly conserved, cellular process that is necessary for both basic cell division and a number of specialized growth patterns, for example, during development [1] and the formation of neuronal processes. The budding yeast, *S. cerevisae*, is a common model organism for studying polarized growth, which adaptively re-uses the polarization machinery for vegetative growth, mating, and filamentous growth [2]. During normal cell division by budding, yeast orient their growth relative to the site of their last bud. However, during mating, the direction of growth is determined by a pheromone gradient that allows cells to extend a mating projection, the shmoo, toward cells of the opposite mating type. In each case, the cell's morphological rearrangement is accompanied by a number of other changes, including changes in protein localization.

When exposed to pheromone, yeast cells arrest in the G1 phase of the cell cycle and extend a shmoo up the gradient of the pheromone. As the cell extends the shmoo, it re-orients various components of the cellular machinery along the new axis of polarization. Many of the processes involved in budding are also used in shmoo formation; however, there are a number of differences as well. Morphologically, the shmoo neck does not become as constricted as the bud neck and a number of proteins involved in pheromone sensing are not part of the budding process. Furthermore, the functional outcomes, the nuclear segregation and cytokinesis of budding versus the cell fusion and karyogamy of mating, entail a number of other different processes.

Although a number of shmoo related genes have been localized to the shmoo (e.g. shmoo marker, Fus1 [3]), previous proteome-wide screens have not yet characterized the shmoo-dependent re-localization of proteins. My collaborators on this project [4] developed a cell micro-array based imaging assay that can characterize the spatial distribution of proteins throughout the cell by simultaneously surveying several thousand yeast strains with GFP tagged proteins. The cell chip, a new technology for high-throughput microscopy, was recently developed by a collaborative effort of several labs at the University of Texas at Austin [5]. The method involves the parallel treatment and fixation of thousands of yeast strains, which are then printed on a microscope slide using technology similar to the Stanford style microarray. In this project, the library of yeast strains with GFP tagged proteins were exposed to alpha factor, a yeast mating pheromone. The method is outlined in **Figure 2.1** and reported in Narayanaswamy et al [5]. The initial genome-wide screen identified 37 genes localized to the shmoo; however, the screen only identified 6 of 47 genes annotated in the literature (as reported by the Saccharomyces Genome Database, SGD [6]) as localized to either the shmoo or the polarisome (part of the polarization machinery), which indicated a significant false-negative rate.

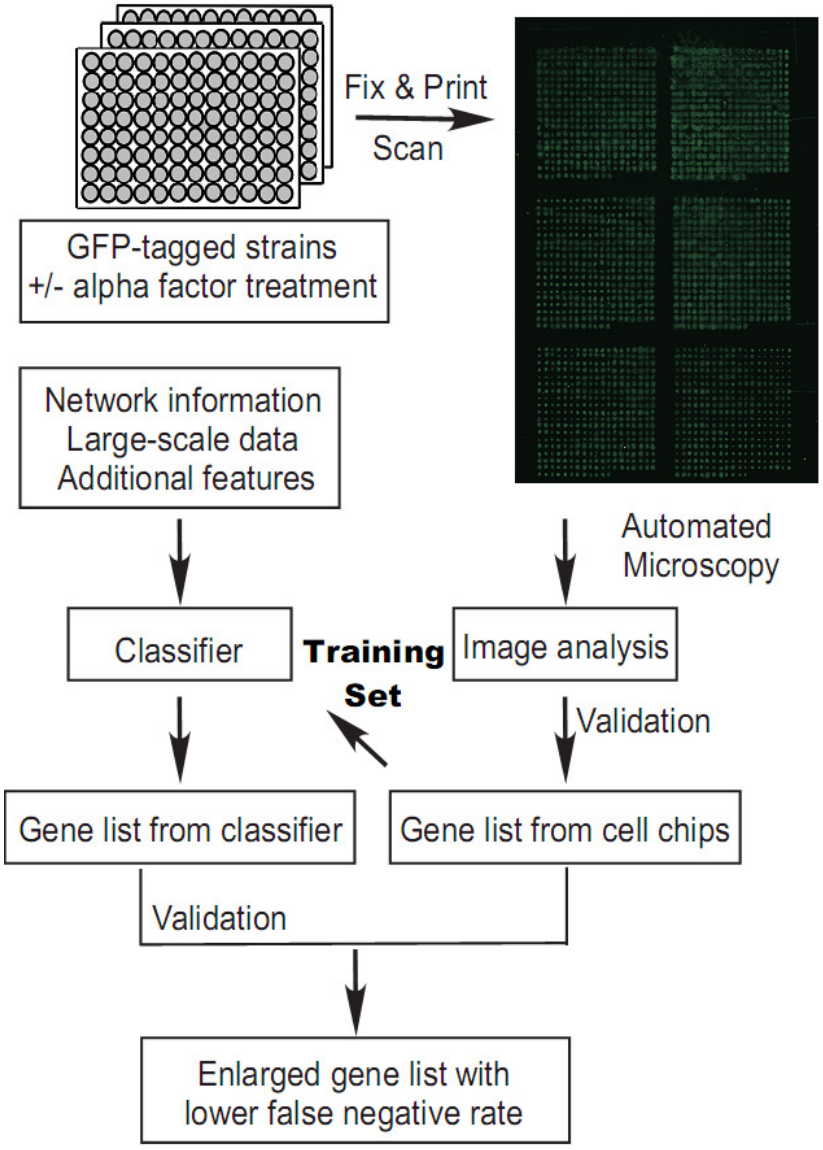


FIGURE 2.1 SCHEMATIC OF CELL CHIP METHOD WITH PREDICTIONS. First, the cell chip is screened to identify shmoo genes and the functional network is constructed independently. Second, genes identified in the screen are used to train a network based classifier. Finally, a manual screen of predicted genes recovers additional shmoo localized genes. Adapted from [4].

High throughput genome-wide screens are notoriously prone to false negatives [7], usually due to technical reasons and the possibility of human error when thousands of yeast strains are being assayed. In this case, the high false negative rate may be due to fixation induced auto-fluorescence, which reduces the signal-to-noise ratio and obscures lower abundance proteins. When we realized that the false negative rate of the automated screen was high, we asked whether a targeted manual screen might recover additional shmoo localized proteins more efficiently than re-screening the entire library. We created a classifier to predict additional shmoo genes using the genes identified in the cell chip assay as a training set and manually screened the set of predicted genes. The follow up screen identified an additional 37 shmoo localized proteins and together established a clear ordering of cellular organelles along the axis of polarization.

METHODS

Classifier Construction

The genes identified in the high throughput screen were used to train a naïve Bayesian classifier (using the machine learning tool, Weka). Six genes annotated as mitochondrial proteins were manually removed from the set to avoid training on them, since they were potentially false positives. Features were aggregated from data from the UCSF GFP screen [8] and the pre-publication YeastNet v.2 functional network by Lee *et al.* [9]. The features included for each gene were: the sum of log likelihood scores (LLS) to the set of shmoo genes, the ratio of the LLS sum linking genes to the shmoo set divided by the LLS sum of the gene's links to all genes in the network, estimated protein abundance (molecules per cell), and cell location during growth in rich media (in the absence of pheromone). The test set of 5804 genes, labeled as shmoo or not-shmoo was

also used as the both the training set and test set. Ten-fold cross validation had very similar results. The area under the ROC curve was 0.843, which indicates that it is a reasonably accurate classifier. After training, the classifier recovered 20 of the 37 shmoo genes (cross-validation: 19). An additional 151 (cross-validation: 153) genes not identified in the initial screen were also classified as shmoo genes using a 0.5 probability cutoff. The classifier was constructed in collaboration with Matt Davis.

Testing predicted shmoo genes

For the manual secondary screen, proteins predicted to be shmoo localized were tested in strains from the *S. cerevisiae* GFP tagged clone collection (Invitrogen). The collection consists of strains derived from the strain EY0986 (ATCC 201388: MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 (S288C)) by chromosomally tagging ~4200 genes with Aequorea Victoria GFP (S65T) at the carboxy-terminal end of an open reading frame [10]. We inoculated GFP strains from -80°C stocks into YPD in 96 well plates, grew them overnight, exposed them to alpha factor for three hours, and imaged them on a fluorescent scope. Two graders independently scored cell images for shmoo localization of the GFP signal. Proteins were considered shmoo localized when both grader's agreed. Microscopy for classifier predicted genes was performed in collaboration with Ram Narayanaswamy.

Additional related methods, which I was not involved in, are published in Narayanaswamy, *et al* [4].

RESULTS

Shmoo localized proteins can be predicted from their functional linkage to known shmoo proteins.

In order to expand our recovery of yeast proteins, we combined genome-wide datasets of protein localization in yeast cells growing in the absence of pheromone [10] and the integrated probabilistic gene network [9, 11] to develop a classifier for predicting additional proteins localized to the shmoo tip (for example, proteins that had been missed because of low expression or penetrance). As detailed in the Methods, our initial set of 37 proteins from the cell chip screen was used to train a *naïve* Bayesian classifier (**Table 2.1**). Application of the classifier identified 151 proteins exceeding a 50% probability score threshold. With a limited set of candidate genes, we could individually assay them in the absence of fixative. 118 of the 151 proteins were already present in the extant GFP library. We manually re-tested each of these 118 GFP fusion strains for protein localization to the shmoo tip. From this set, 37 additional proteins (~31% of those tested) were confirmed to be shmoo-localized (**Table 2.2**).

TABLE 2.1: MANUALLY VERIFIED SHMOO TIP LOCALIZED GENES IDENTIFIED BY THE CELL CHIP. As in [4].

Gene name	ORF name	Human ortholog*	Gene Ontology biological process annotation
ABP1	YCR088W	DBNL	establishment of cell polarity (sensu Fungi)
AIP1	YMR092C	WDR1	response to osmotic stress
BEM3	YPL115C	-	pseudohyphal growth
CAP1	YKL007W	CAPZA2	barbed-end actin filament capping
CAP2	YIL034C	CAPZB	filamentous growth
CAR1	YPL111W	ARG1	arginine catabolism to ornithine
CBK1	YNL161W	STK38L	regulation of exit from mitosis
CDC10	YCR002C	SEPT9	cell wall organization and biogenesis
CDC11	YJR076C	-	cell wall organization and biogenesis
CDC48	YDL126C	VCP	ubiquitin-dependent protein catabolism
EDE1	YBL047C	EPS15	endocytosis
END3	YNL084C	-	endocytosis
ENT1	YDL161W	EPN3	endocytosis
EXO70	YJL085W	-	cytokinesis
EXO84	YBR102C	EXOC8	exocytosis
FUS1	YCL027W	-	conjugation with cellular fusion
INP52	YNL106C	SYNJ2	cell wall organization and biogenesis
KEL1	YHR158C	RABEPK	cell morphogenesis
LSG1	YGL099W	GNL1	ribosome biogenesis
MID2	YLR332W	-	cell wall organization and biogenesis
PEA2	YER149C	-	pseudohyphal growth
POP2	YNR052C	CNOT8	regulation of transcription from RNA polymerase II promoter
SEC10	YLR166C	EXOC5	establishment of cell polarity (sensu Fungi)
SEC18	YBR080C	NSF	ER to Golgi vesicle-mediated transport
SEC2	YNL272C	-	exocytosis
SEC3	YER008C	-	cytokinesis
SEC5	YDR166C	EXOC2	cytokinesis
SEC6	YIL068C	EXOC3	cytokinesis
SEC8	YPR055W	EXOC4	cytokinesis
SHM2	YLR058C	SHMT1	one-carbon compound metabolism
SHR3	YDL212W	-	ER to Golgi vesicle-mediated transport
SLA1	YBL007C	GRAP	cell wall organization and biogenesis
SLG1	YOR008C	-	cell wall organization and biogenesis
SMY1	YKL079W	-	exocytosis
YCR043C	YCR043C	-	biological process unknown
YMR295C	YMR295C	-	biological process unknown
YOR304C-A	YOR304C-A	-	biological process unknown

* as calculated by InParanoid [12], listing only the top-scoring inparalog

TABLE 2.2: MANUALLY VERIFIED SHMOO TIP LOCALIZED PROTEINS IDENTIFIED BY THE CLASSIFIER. As in [4].

Gene name	ORF name	Human ortholog*	Gene Ontology biological process annotation
ABP140	YOR239W	METTL2B	actin cytoskeleton organization and biogenesis
ARK1	YNL020C	AAK1	protein amino acid phosphorylation
BCK1	YJL095W	-	protein amino acid phosphorylation
BEM1	YBR200W	-	establishment of cell polarity (sensu Fungi)
BNI1	YNL271C	DIAPH1	pseudohyphal growth
BOI1	YBL085W	-	establishment of cell polarity (sensu Fungi)
BSP1	YPR171W	-	actin cortical patch distribution
BUD6	YLR319C	-	actin filament organization
BZZ1	YHR114W	TRIP10	endocytosis
CHS3	YBR023C	-	cytokinesis
CHS5	YLR330W	-	spore wall assembly (sensu Fungi)
ENT2	YLR206W	EPN3	endocytosis
KEL2	YGR238C	RABEPK	conjugation with cellular fusion
LAS17	YOR181W	WASL	endocytosis
MYO2	YOR326W	MYO5B	vesicle-mediated transport
MYO5	YMR109W	MYO1E	cell wall organization and biogenesis
PAN1	YIR006C	-	endocytosis
PRK1	YIL095W	AAK1	protein amino acid phosphorylation
RGD1	YBR260C	ARHGAP21	response to acid
RVS161	YCR009C	BIN3	endocytosis
RVS167	YDR388W	-	endocytosis
SAC6	YDR129C	PLS3	endocytosis
SEC15	YGL233W	EXOC6	cytokinesis
SEC31	YDL195W	SEC31A	ER to Golgi vesicle-mediated transport
SFB3	YHR098C	-	ER to Golgi vesicle-mediated transport
SHS1	YDL225W	-	establishment of cell polarity (sensu Fungi)
SLA2	YNL243W	HIP1R	actin filament organization
SMI1	YGR229C	-	regulation of fungal-type cell wall biogenesis
SRV2	YNL138W	CAP1	pseudohyphal growth
SYP1	YCR030C	-	biological process unknown
TWF1	YGR080W	TWF1	bipolar bud site selection
VRP1	YLR337C	WIPF1	endocytosis
WSC2	YNL283C	-	cell wall organization and biogenesis
WSC3	YOL105C	-	cell wall organization and biogenesis
YDR348C	YDR348C	-	biological process unknown
YER071C	YER071C	-	biological process unknown
YIR003W	YIR003W	-	biological process unknown

* as calculated by InParanoid [12], listing only the top-scoring inparalog

The classifier-guided retesting strategy doubled the coverage of the screen, and remaining false negatives could be rationalized based upon low protein abundance. The targeted secondary screen had a much higher success rate than the initial unguided screen (1% vs. 31%). Of the 37 genes identified in the second round of screening, 7 overlapped with previous literature, bringing the combined total of identified genes to 13 of 47 genes reported in the literature (**Figure 2.2A**). Of genes in the GFP library, we could identify 63% of the known shmoo tip-localized proteins that were present at >2500 molecules/cell, but less than 21% of known proteins with <2500 molecules/cell, which suggest that the high false negative rate may primarily be due to the insensitivity of the microscope and camera. On average, the shmoo tip proteins identified via the classifier method were less abundant than those recovered via the cell chip method, which suggests that using a network guided approach to expand an initial list of seed genes works well in cases with high false negative rates that can be reduced in lower-throughput assays.

Adaptive re-use of polarization proteome

Unsurprisingly, the set of 74 shmoo-localized proteins (37 each from initial screen and network identified), showed a marked enrichment for Gene Ontology functional categories related to polarized growth (with $p < 10^{-6}$ being the threshold of probability calculated using a hypergeometric distribution [13] that the intersection of given list with any functional category occurs by chance), with the strongest enrichment observed for the GO Biological Process annotation *establishment of cell polarity* ($p < 10^{-35}$), followed by annotations including *anatomical structure morphogenesis* ($p < 10^{-32}$), *cellular bud site selection* ($p < 10^{-29}$), *cytokinetic process* ($p < 10^{-28}$), *vesicle-mediated transport* ($p < 10^{-22}$), *reproduction* ($p < 10^{-20}$), *endocytosis* ($p < 10^{-18}$), *actin filament organization* ($p <$

10^{-13}), *exocytosis* ($p < 10^{-12}$), and *conjugation* ($p < 10^{-6}$). Furthermore, there appears to be very broad reuse of the proteins during formation of the shmoo and buds (**Figure 2.2B**).

In addition, 41 of the 74 proteins have human orthologs (**Tables 2.1 & 2.2**), which implies that there is a broad conservation of these processes across eukaryotes.

Figure 2.3 indicates a number of complexes involved in the polarization process. The exocyst, an octomeric complex, helps dock vesicles to the bud site during cell growth. Together, the initial and classifier screens identified all eight members of the complex, which strongly implies a conserved role for the complex in both budding and shmooing. Members of the septin ring, a pentameric complex, were also identified, are characteristically present at the shmoo neck, forming a collar like structure. One of the members CDC3, was absent from the GFP library and could not be identified. Additionally, several components of the actin cortical patch were identified as present in the shmoo, which illustrates the role of the actin cytoskeleton in supporting the polarized outgrowth. Interestingly, we identified a number of genes which are actively involved in endocytosis, RVS161, RVS167, SAC6, and ENT2, which were broadly distributed throughout the shmoo, often in punctuates. **Figure 2.4** We observed that proteins involved in exocytosis were localize at the extreme tip of the shmoo while proteins involved in endocytosis localized more broadly around the shmoo tip, which may indicate a great deal of membrane turnover in the shmoo as it grows.

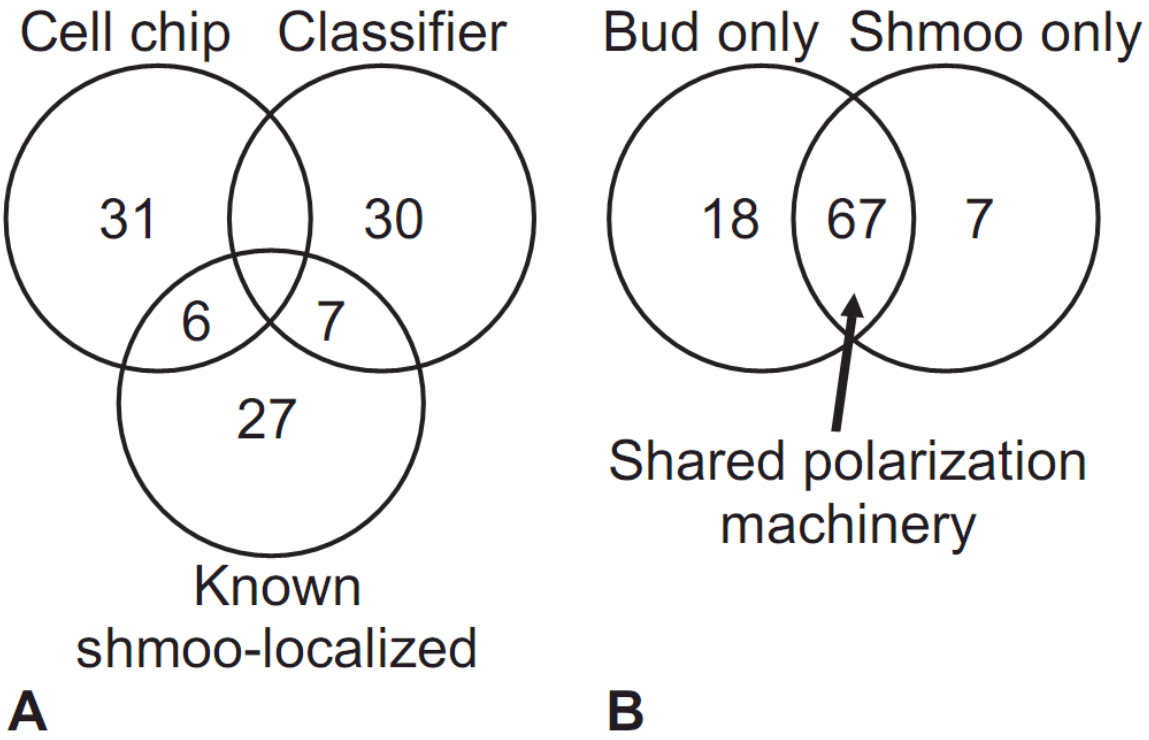


FIGURE 2.2 A. MANUAL VALIDATION BASED ON CLASSIFIER DOUBLES COVERAGE OF KNOWN SHMOO GENES. B. GENES INVOLVED IN SHMOO FORMATION OVERLAP SIGNIFICANTLY WITH BUDDING GENES and indicate significant reuse of polarization machinery. Adapted from [4].

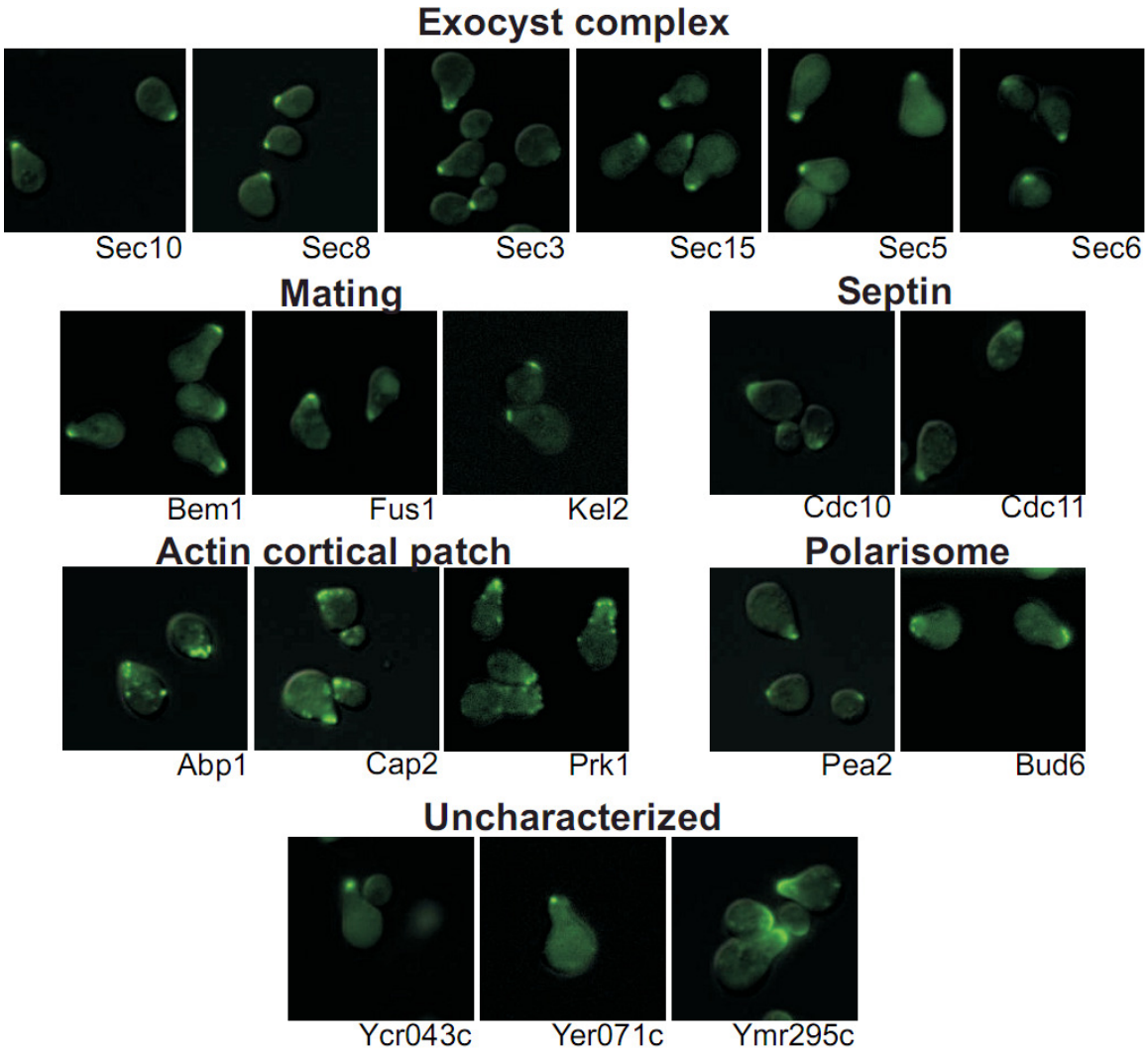


FIGURE 2.3 IMAGES OF VARIOUS SHMOO LOCALIZED CELLULAR COMPONENTS.

GFP-tagged proteins are identified by microscopy. Adapted from [4].

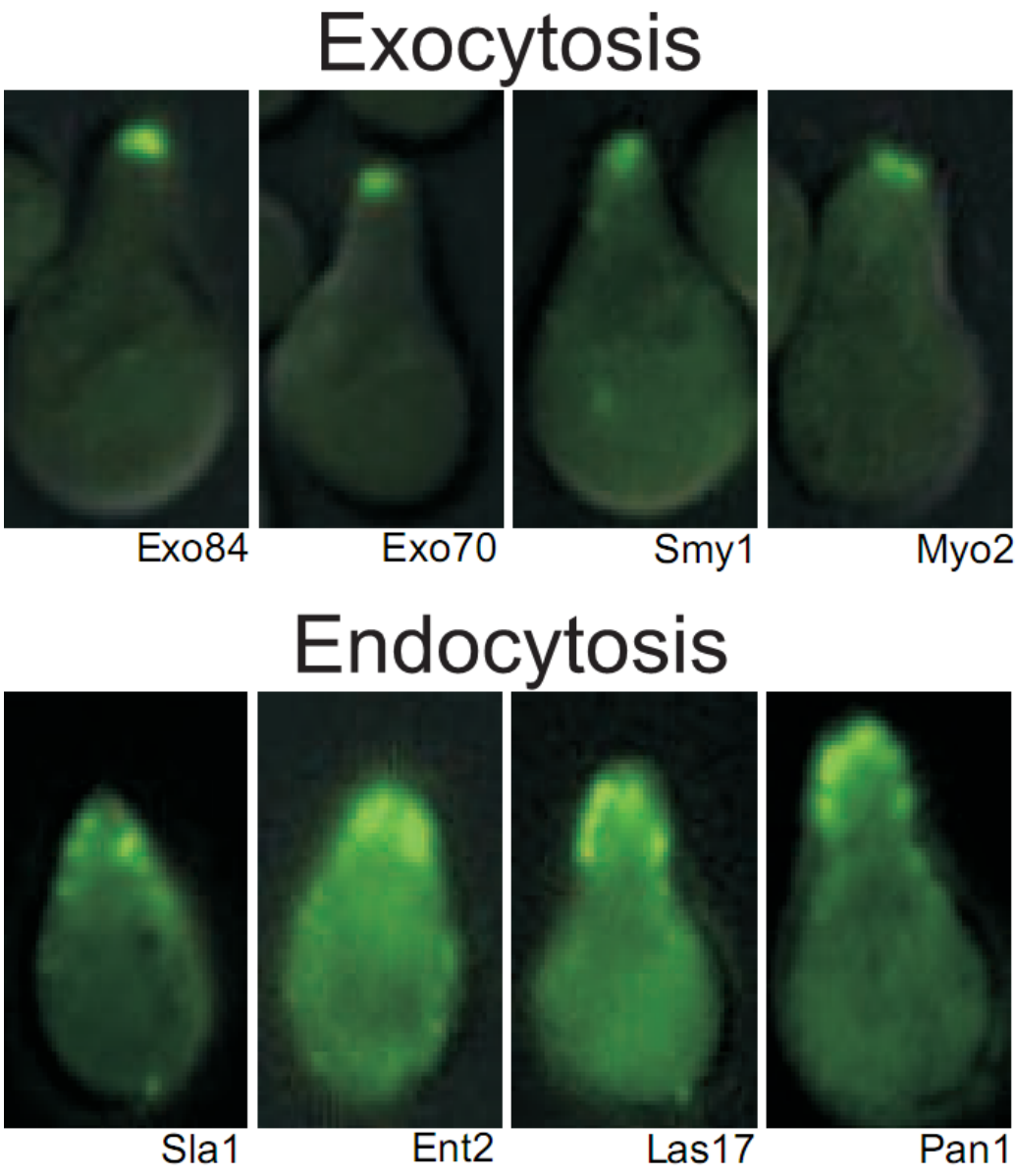


FIGURE 2.4 EXOCYTOTIC PROTEINS ARE MORE DISTALLY LOCATED THAN ENDOCYTOTIC PROTEINS IN THE SHMOO. Adapted from [4].

In addition to characterized complexes, the screen and network based classifier found several uncharacterized proteins, such as YMR295C [14], YDR348C, and YOR304C-A, localize to both the bud [10] and shmoo tips which implies general reuse in polarization. Examination of these proteins' functional relationships in the yeast functional gene network illustrates their potential involvement in specific aspects of polarized cell growth (**Figure 2.5**): YOR304C-A is linked with Bud6, a key protein in polarization signaling (and also found in the screen) and Duo1, a cytoskeletal protein. YMR295C and YDR348C are network neighbors, with the former also linked to glycolytic transcription factor Gcr1, and the latter tied to the cell cycle progression genes Clb2 and Cdc28. Therefore, these genes may help connect polarization with other cellular systems such as cell cycle control and energy metabolism. The functional network, therefore, helps predict the localization of uncharacterized genes and provides a starting point by which to evaluate their role.

In line with the adaptive re-use of the polarization machinery, the primary predictive strength that prior localization data contributed to the classifier was bud localization. However, only 6 of the 37 genes identified in the second round were bud localized, which suggests that the primary predictive power of the classifier came from the features derived from the functional network. This demonstrates the power of the functional network to find likely false negatives from genome-wide screens of gene localization and can significantly expand the number of genes recovered with minimal re-screening.

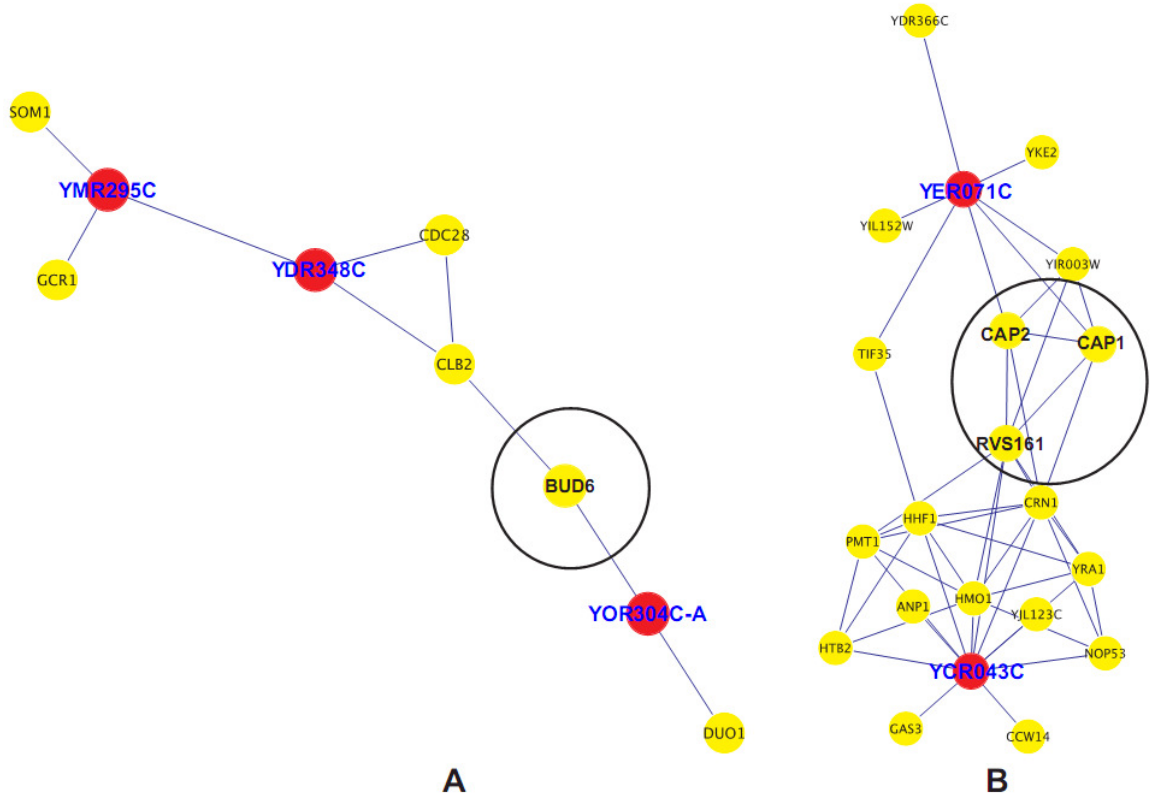


FIGURE 2.5 THE FUNCTIONAL NETWORK SUGGESTS POSSIBLE ROLES FOR UNKNOWN PROTEINS LOCALIZED TO THE SHMOO. Characterized proteins are labeled in black and represented by yellow circles. Uncharacterized proteins are in labeled in blue with red circles. Bold genes are identified shmoo genes. See text for additional details. Adapted from [4].

CONCLUSION

Polarized growth is a fundamental cellular process that has been repurposed for multiple functions and the spatial distribution of the shmoo proteome demonstrates that a significant fraction of genes involved in polarized growth are shared by differing cellular processes in *S. cerevisiae*. Of the 67 proteins shared between the mating and budding polarization processes (**Figure 2.2**) there is significantly greater conservation between human and yeast than expected by chance (39 of the 67 yeast genes have human orthologs [12], $p < 0.023$, chi-square test), suggesting that the functions of these core polarization components are consistent across eukaryotes. Recognition of this significant overlap helps contribute to our understanding of phenologs, a topic I will be covering in detail in chapter 4.

The functional network, which integrates a broad variety of data, but does not explicitly include location data, can predict protein localization even for uncharacterized genes. Functional networks were developed to predict gene function [11] but this project illustrates that they are capable of broader application because they capture information beyond the originally intended application.

The process of screening followed by prediction could be iterated by adding the newly identified shmoo genes to the training set. This would then re-weight their neighboring genes in the network, which could potentially identify additional genes that are related to the shmoo and pheromone response pathways.

This chapter has been abstracted and reworked from a paper that is in press [4].

REFERENCES

1. Muller HA: Genetic control of epithelial cell polarity: lessons from Drosophila. *Dev Dyn* 2000, 218:52-67.

2. Madden K, Snyder M: Cell polarity and morphogenesis in budding yeast. *Annu Rev Microbiol* 1998, 52:687-744.

3. Proszynski TJ, Klemm R, Bagnat M, Gaus K, Simons K: Plasma membrane polarization during mating in yeast cells. *J Cell Biol* 2006, 173:861-866.

4. Narayanaswamy R, Moradi E, Niu W, Hart G, Davis M, McGary K, Ellington A, Marcotte EM: Systematic definition of protein constituents along the major polarization axis reveals an adaptive re-use of the polarization machinery in pheromone treated budding yeast. *Journal of Proteome Research* 2008.

5. Narayanaswamy R, Niu W, Scouras AD, Hart GT, Davies J, Ellington AD, Iyer VR, Marcotte EM: Systematic profiling of cellular phenotypes with spotted cell microarrays reveals mating-pheromone response genes. *Genome Biol* 2006, 7:R6.

6. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M *et al*: SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 1998, 26:73-79.

7. Edmonds D, Breitkreutz BJ, Harrington L: A genome-wide telomere screen in yeast: the long and short of it all. *Proc Natl Acad Sci U S A* 2004, 101:9515-9516.

8. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: Global analysis of protein expression in yeast. *Nature* 2003, 425:737-741.

9. Lee I, Li Z, Marcotte EM: An Improved, Bias-Reduced Probabilistic Functional Gene Network of Baker's Yeast, Saccharomyces cerevisiae. *PLoS ONE* 2007, 2:e988.

10. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: Global analysis of protein localization in budding yeast. *Nature* 2003, 425:686-691.

11. Lee I, Date SV, Adai AT, Marcotte EM: A probabilistic functional network of yeast genes. *Science* 2004, 306:1555-1558.

12. Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL: InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res* 2008, 36:D263-266.

13. Reimand J, Kull M, Peterson H, Hansen J, Vilo J: g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* 2007, 35:W193-200.

14. Fleischer TC, Weaver CM, McAfee KJ, Jennings JL, Link AJ: Systematic identification and functional screens of uncharacterized proteins associated with eukaryotic ribosomal complexes. *Genes Dev* 2006, 20:1294-1307.

Chapter 3: Predicting yeast knockout phenotypes with a functional network

INTRODUCTION

After the successful application of the classifier to protein localization and expanding the coverage of genome wide screens, I looked for a related application where the network could guide research by providing targeted predictions. I also sought to reduce the complexity and obscurity of the classifier by adopting a simpler, more broadly applicable, and intuitively understandable algorithm for prediction. In this chapter and the associate paper [1], I demonstrate that loss-of-function yeast phenotypes are predictable by simple guilt-by-association in functional gene networks. By computational testing of more than one thousand loss-of-function phenotypes from genome-wide assays of yeast I show that diverse phenotypes are predictable, spanning cellular morphology, growth, metabolism, and quantitative cell shape features. I apply the method to (1) extend a genome-wide screen by predicting, then verifying, genes whose disruption elongates yeast cells, and (2) computationally predict human disease genes. To facilitate network-guided screens, I have established a web server at http://www.yeastnet.org which provides network based predictions based on submitted genes.

BACKGROUND

Historically, genetic relationships between mutations were inferred when the mutations resulted in a shared phenotype. Similar phenotypic outcomes were typically interpreted as representing a functional relationship between the two genes and these relationships were represented as genetic pathways and later as gene networks. With high throughput technologies being integrated into functional networks, it is now possible to ask whether the inverse inference is possible. Are functionally linked genes likely to share a common phenotype? If so, it is possible to predict the phenotypic outcomes of gene disruption by extrapolating from known phenotypic data. Of particular interest, the method could be applied to identify candidate genes that are likely to cause a specific disease when

mutated, based on their linkage to a known disease gene. In this project, I show that a wide range of yeast phenotypes can be predicted using a functional network and demonstrate that the method is likely to be broadly applicable to human disease.

Advances over the past decade in both forward and reverse genetics mean that the predictability I find can be applied in a simple way to correctly associate genes with phenotypes of interest. For forward genetics, genome-wide association studies (reviewed in [2]) are starting to identify candidate genes associated with human traits and diseases, such as recent studies correlating variants in the ORMDL3 gene to risk of childhood asthma [3]. At the same time, reverse genetics by rapid testing of candidate genes has become more routine with the creation of mutant strain collections (e.g., yeast deletion strain collections [4, 5]) and the development of RNA interference (RNAi) for down-regulation of genes (e.g., as for genome-wide RNAi screens of *C. elegans* [6, 7] or human cell lines, reviewed in [8]). With the ability to predict loss-of-function phenotypes, I suggest utilizing the two aspects of genetics synergistically: with a starting set of genes linked to a specific phenotype by a forward genetic screen, computational predictions of additional genes associated with that phenotype can be evaluated using reverse genetics, expanding on the original screen, much like were able to do for shmoo localized proteins in the previous chapter. Furthermore, given the polygenic nature of many diseases and phenotypes, this approach will improve the characterization of the network of genes affecting a trait of interest.

Functional networks have been successfully used to annotate unknown genes using the principle of guilt-by-association (GBA), which assumes that the function of a gene is closely related to its neighbors in the network [9]. I applied GBA to predict yeast phenotypes using YeastNet v.2 [10] by asking if the genes linked to a seed set of genes associated with a particular loss-of-function phenotype might also be more likely to result in the same phenotype upon disruption. This probabilistic functional gene network has 102,803 functional links among 5,483 yeast genes, where the probability of a link indicates the likelihood that two genes will have the same Gene Ontology biological

process annotations [11] relative to the background expectation. Genes are ordered by their connectivity to the initial set; the genes linked most strongly to the seed set become candidate genes for the same phenotype.

An illustration of the functional network and GBA is displayed in **Figure 3.1** as a graph with circles (genes) connected by edges (functional links). Blue circles represent seed genes that lead to the target phenotype upon knockout. Red circles represent tightly linked neighbors that are candidate genes that are predicted to give rise to the phenotype upon disruption. The sums of the genes' probabilistic linkages to the seed set are used to rank the phenotype predictions. As a consequence, genes tightly linked to multiple seed genes score more highly than genes weakly linked to only a single seed gene.

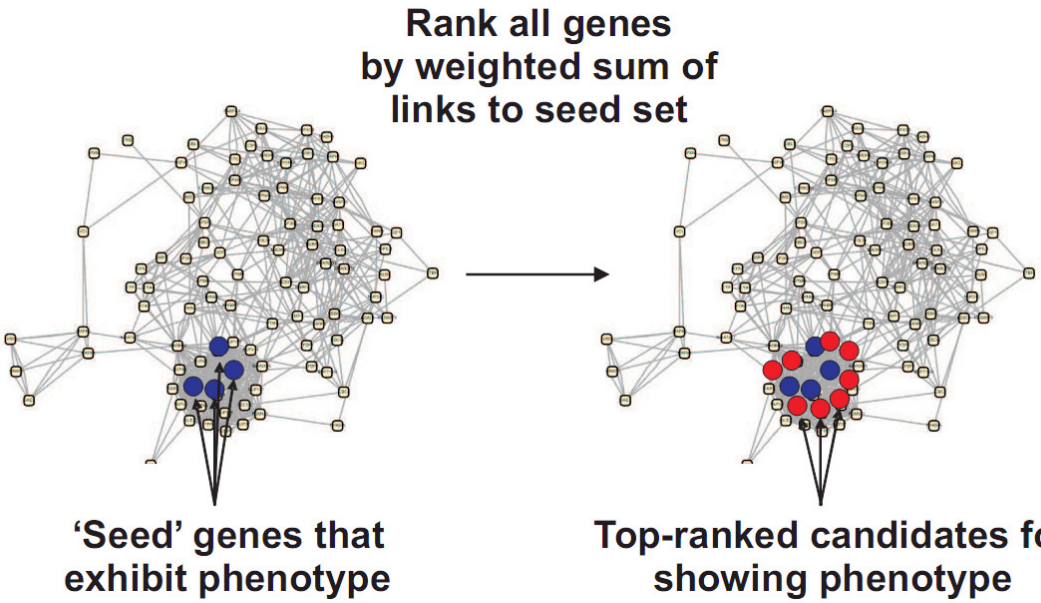


FIGURE 3.1 OVERVIEW OF GUILT-BY-ASSOCIATION PHENOTYPE PREDICTION. Guilt-by-association phenotype prediction employs a functional gene network, represented here as circles (genes) connected by lines (functional linkages), and a seed set of genes (blue circles) whose disruption is known to give rise to the phenotype of interest. Neighboring genes in a functional gene network (red circles) are candidates for also giving rise to the phenotype. Candidates are prioritized by the sum of their network linkage weights to the set of seed genes. A gene strongly linked to multiple seed genes will thus rank more highly than a gene weakly linked to a single seed gene. Network drawn with Cytoscape [12]. Figure used by permission [1].

METHODS

As published in [1].

Assembling the set of non-redundant loss-of-function phenotypes

A literature search was conducted to find genome-scale studies of yeast gene knockout phenotypes. Datasets were compiled from studies that systematically examined a large fraction of the yeast genome. No effort was made to minimize redundancy among the gene sets themselves. Nonetheless, only one set is a strict subset of another (genes that have changed levels of transposon cDNA upon knockout are a subset of the genes that reduce retrotransposition). Most studies were conducted using one or more of the following strain collections: haploid [4], homozygous diploid [4], heterozygous diploid [4], tetracycline-titratable [13]. The reported data were a mix of qualitative, pseudo-quantitative, and quantitative results. Pseudo-quantitative data (often reported as "+", "++", "-", "--", etc.) were thresholded at the most stringent reported value (except for the small set of genes conferring the phenotype “branched cells” [4]; all genes with this morphology were included). Quantitative data were arbitrarily thresholded using cutoffs that appeared consistent with the sensitivity of the assay. Predictability was not used as a criterion for selecting thresholds. In some cases, thresholds less stringent than those selected result in more predictable phenotype sets (data not shown). In cases where an uncharacterized open reading frame overlapped a known gene on the chromosome and both shared the same phenotype (e.g. Axial budding [14]; the dubious open reading frame YOR300W overlaps BUD7), the uncharacterized gene was removed from the phenotype set. Additional phenotypes were collected from the *Saccharomyces* Genome Database (SGD) [15]; phenotypes extracted from SGD used the threshold determined by SGD.

For the 281 quantitative phenotypes reported by SCMD [16], the 40 knockout strains with either the highest or lowest values for each SCMD feature were selected (resulting in 562 seed gene sets). Similarly, 440 CV phenotypes were generated by considering the 40 knockout strains with either the higher or lowest CV for each SCMD CV feature (220 total features).

Prediction of phenotypes and evaluation of prediction quality

For each gene in the network, I calculated the sum of its link weights to genes with the phenotype in question (the seed set), *i.e.*, assigning each gene i the score $S_i = \sum_{j \in seed} LLS_{ij}$, where j is a gene in the seed gene set, and LLS_{ij} is the log likelihood score for the linkage between gene i and gene j, as reported in [11] except where explicitly analyzing other networks. Genes were then rank-ordered by their S_i scores, with the highest scoring genes the most likely to share the phenotype with the seed set. For networks reporting only binary linkages (MIPS [17], DIP [18]), I considered all linkages to be of weight 1. For calculation of **Figure 3.5**, YeastNet v.2, DIP and PICO [19] were each evaluated at two different confidence levels. For analyses of protein interaction networks, the following networks were analyzed: YeastNet v. 2, which corresponds to all interactions reported in [11]; physical protein interactions (PPI) from the Database of Interacting Proteins (DIP) [18] (downloaded on February 4, 2007) selecting as the core set those interactions reported by [20]; Collins *et al* [21], using their reported threshold; PICO E-0 and E-2, PPI sets from [19]; MIPS, all PPI in physical complexes reported by [19] derived from [17]. In all cases, self interactions were removed.

For each phenotype, the predictability was evaluated by generating a ROC curve based upon the gene ranking and calculating the area under the curve (AUC). The ROC

curve indicates the relative rate of true and false positive predictions as a function of the score S_i, plotting the true positive rate (TP/(TP+FN)) versus false positive rate (FP/(FP+TN)). In calculating S_i, self-self links were not permitted, and each gene in the seed set was withheld in turn from the seed set for evaluation (*i.e.*, leave-one-out cross-validation). TP, true positives, was defined (for a specific threshold) as the number of genes from the seed set ranked above a given S_i; FP, false positives, as the number of genes above the threshold but not in the seed set; FN, false negatives, as the number of seed genes ranked below the threshold; and TN, true negatives, as the number of non-seed genes ranked below the threshold. The AUC ranges from 0 to 1, with 0.5 indicating random performance and 1.0 indicating perfect classification. Note that AUC is calculated using only seed genes represented in the network (*i.e.*, the network is not penalized for partial coverage of the seed set), allowing the predictive capacity of networks of differing sizes to be compared. For the purposes of calculating a ROC curve, all genes not linked to the phenotype seed set were treated as being of the same rank. Note that none of the phenotypes have been tested for all genes (most tested only non-essential genes). Due to differences in the reporting of genes tested, ROC curves for the set of 100 phenotypes were calculated over the entire set of yeast genes in the network being tested (5,483 genes for the functional network). Thus, the measures of predictability (AUC) are likely to be underestimates, since all untested genes are considered false positives.

As an alternative test for functional enrichment, I used ArrayPlex [22] to calculate the hypergeometric probability of the enrichment for each GO annotation within a given gene set.

Prediction of human disease gene sets

For the test of human disease gene prediction, I collected sets of yeast genes whose human orthologs were linked to the same OMIM disease [23]. Human disease phenotypes from OMIM were collapsed into major categories (*i.e.*, variants of each disease were collapsed into a single category, such as collapsing "Cataract, polymorphic and lamellar" and "Cataract, crystalline aculeiform" into a single category of cataract defects). Each human disease gene was mapped to one of 2,151 human-yeast orthology groups using InParanoid [24], and seed sets of yeast genes linked to the same disease were selected such that at least 4 of the yeast genes were present in YeastNet. Calculation of predictability and measurement of AUC was performed as for yeast phenotypes, considering linkages in YeastNet between human-yeast orthology groups rather than between individual yeast genes.

Generation of random phenotype sets

In order to estimate the random distribution of AUC scores for literature phenotypes, sets of genes of the same sizes as the real phenotype seed sets were drawn from the complete set of yeast genes and tested for predictability, using as the background set of genes those designated by SGD as "verified" or "uncharacterized" (not dubious or pseudogenes) (as of January 29, 2007). For SCMD morphology phenotypes [16], 1000 sets of 40 genes were drawn randomly from the complete set of genes analyzed by SCMD, then tested for predictability in order to generate the null expectation for the AUC distribution. For human disease phenotypes, random gene sets were generated for comparison by randomly drawing from the set of network annotated human-yeast orthologs such that the set size distribution of the random sets matched the size distribution of the actual OMIM disease seed sets.

Yeast strains, media, and growth

For predicting elongation mutants, I employed a seed set of 77 non-essential genes identified by Giaever *et al.* [4] as "Elongate 3" in a screen of the homozygous diploid yeast deletion collection. Using GBA with this seed set, I predicted additional genes likely to give rise to elongated cells, and selected for assay the 35 top-ranked essential genes with strains available in the tetracycline-downregulatable library of yeast strains [13]. A negative set of 17 strains from the same library was randomly selected from those genes not linked to any of the known elongated genes. The corresponding strains were obtained from Open Biosystems. Each strain was grown to saturation at 30°C in YPD, inoculated into fresh YPD with 10 ng/ml doxycycline, grown 16 hours and imaged [13] to evaluate cell morphology. Two biologists evaluated the images for each strain (with strain names hidden) for elongated cell morphologies using a simple qualitative scoring scheme (0-2), assigning a final score to each strain as the sum of the independent evaluations. Strains scoring >2 were selected as elongated, which minimized false positives, yet recovered NUT2, previously reported to be elongated [13]. Wei Niu helped train me in microscopy and helped handle the yeast library. Edward Marcotte helped with image analysis.

In order to predict gene-phenotype associations (see prediction testing below for phenotypes tested), I calculated the sum of links between a given gene and a seed set of genes known to lead to a specific phenotype upon knockout, i.e., each gene in the network, i was given the score $S_i = \sum_{j \in seed} LLS_{ij}$, where j is a gene with the given phenotype, and LLS_{ij} is the log likelihood score for the linkage between gene i and gene j. Self-self links were not permitted, effectively making the approach equivalent to a classifier using leave-one-out cross-validation. The genes with the highest S_i score are

most likely to share the phenotype with the seed set. Genes were rank ordered by their S_i score for each phenotype, which was used to generate a ROC curve and calculate the area under the curve (AUC). The ROC curve plots the relative true positive rate (TP/(TP+FN)) to the false positive rate (FP/(FP+TN)) as a function of S_i. True positives, TP, were defined (for a given S_i threshold) as the number of genes with the specific phenotype that scored above the threshold; false positives, FP, as the number of phenotype related genes below the threshold; false negatives, FN, as the number of genes in the set below the threshold; and true negatives, TN, as the number of genes not known to have the phenotype below the threshold. The AUC can range from 0 to 1, with 0.5 indicating that the predictions are random. The appropriate background set of genes varies between phenotype screens, since none of the knockout libraries are complete; however, the ROC curves were calculated assuming that the entire genome has been screened. It is likely that the real AUC is underestimated for many phenotypes, since untested, predicted genes are treated as false positives.

RESULTS

To evaluate the utility of applying the GBA concept to phenotype prediction, I initially tested predictions computationally with a large assortment of phenotypes and then experimentally validated a set of predictions for a specific morphological trait.

Computational Results

Using Guilt-By-Association to Predict Essentiality

I first investigated whether the network could distinguish viable from non-viable yeast gene deletion strains. Essential genes of both yeast and humans are known to be more highly connected in protein physical interaction networks than non-essential genes

[25-27], and there is evidence that essential proteins may also be enriched in the same physical complexes [19, 28]. I asked if essential genes could be predicted on the basis of their connections to other essential genes in a functional gene network. I employed the guilt-by-association approach, using as the seed set the 1,027 known essential yeast genes [4, 29], then scoring each gene in yeast for its likelihood to be essential as a function of connectivity to this seed set. As described in the methods above, each gene in the seed set was withheld in turn from the seed set in order to evaluate it; *i.e.*, performing leave-one-out cross-validation. As the prediction score for each gene, I calculated the sum of the weights of linkages connecting the query gene to genes in the seed set. Given that each linkage's weight in this network corresponds to the log likelihood of the linked genes belonging to the same pathway [11] the sum of linkage weights therefore represents the *naïve* Bayesian combination of evidence that the query gene belongs to the same pathway as the seed set genes. I expect genes in the same pathway to often exhibit the same loss-of-function phenotypes. Thus, this score should also serve to identify genes that share phenotypes with the seed set genes.

To evaluate prediction quality, I calculated the true positive rate (sensitivity, TP/(TP+FN)) and the false positive rate (1-specificity, FP/(FP+TN)), as a function of the prediction score, plotting the resulting receiver operating characteristic (ROC) curve. As **Figure 3.2** shows, the essential genes are strongly predictable on the basis of their network neighbors. Therefore, in addition to the previous observations that essential genes have larger numbers of physical interaction partners, I demonstrate that essential yeast genes are also preferentially connected to each other in a functional network.

A yeast gene network predicts varied, specific loss-of-function phenotypes

In order to further test phenotypic predictability, I collected an additional set of 99 yeast knockout phenotypes that had been generated in large scale genetic screens that assayed a substantial fraction of the genome (typically, all non-essential genes). These reverse genetic screens are made possible by the recent creation of libraries of yeast knockout strains [4, 5] and are reported either in the *Saccharomyces* Genome Database (SGD; [30]) or in one of 32 additional publications in the literature, listed in full in **Table 3.1**. In these collections, a single yeast gene is deleted in each yeast strain; a phenotypic assay on the complete set of knockout strains thereby associates that phenotype with those deleted genes that gave rise to it. These screens are ideal for addressing the general question of whether or not specific loss-of-function phenotypes are predictable. Crucially, the phenotypic data is not integrated into the functional network [11], which allows for an independent, unbiased assessment of the functional networks predictive accuracy.

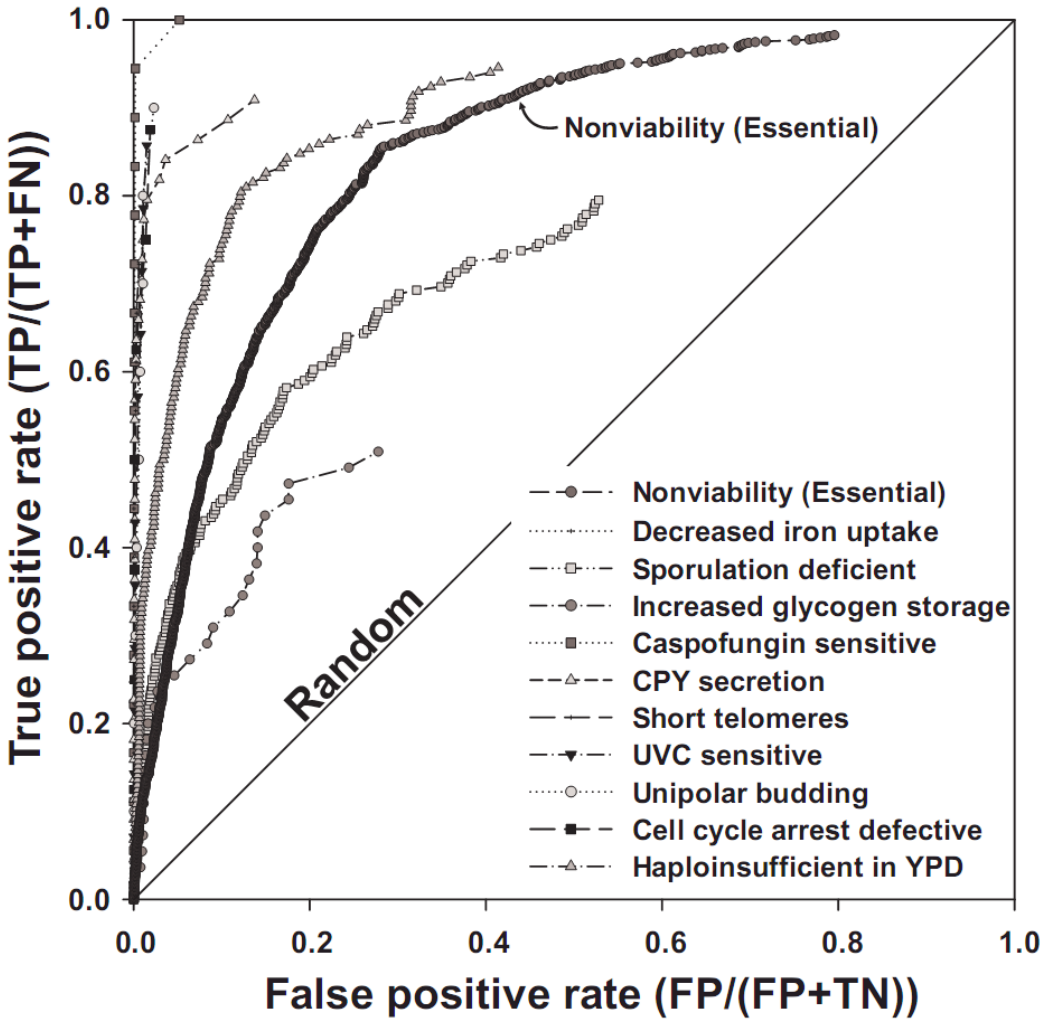


FIGURE 3.2 DIVERSE YEAST GENE LOSS-OF-FUNCTION PHENOTYPES ARE PREDICTABLE USING GUILT-BY-ASSOCIATION IN A FUNCTIONAL GENE NETWORK. Predictability is measured in a ROC plot of the true positive rate (sensitivity) versus false positive rate (1-specificity) for predicting genes giving rise to 10 specific loss-of-function phenotypes, as well as for essential genes whose disruption produces nonviable yeast [4]. For each phenotype, each gene in the yeast genome was prioritized by the sum of the weights of its network linkages to the seed genes associated with the phenotype. Genes with higher scores are more tightly linked to the seed set and therefore more likely to give rise to the phenotype. Each phenotype was evaluated using leave-one-out cross-validation, omitting genes from the seed set for the purposes of evaluation. More predictable phenotypes tend towards the top-left corner of the graph; random predictability is indicated by the diagonal. For clarity, the line connecting the

final point of each graph to the top right corner has been omitted. Figure used by permission [1].

TABLE 3.1. PREDICTABILITY OF 100 YEAST GENE DELETION PHENOTYPES. Table used by permission from [1].

Phenotype[a]	AUC	# seed genes with phenotype	# seed genes in network	Cite
caspofungin sensitive	0.996	20	18	[31]
increased resistance to calcofluor white	0.982	10	10	[32]
unipolar budding	0.941	10	10	[14]
CPY secretion (3)	0.937	46	44	[33]
cell cycle arrest defective	0.930	8	8	[34]
UVC sensitive (high)	0.919	15	14	[35]
sensitivity at 15 generations in galactose	0.908	17	14	[4]
CANR mutator (high)	0.904	18	18	[36]
haploinsufficient in rich medium (YPD)	0.898	184	184	[37]
cellular chitin level increased (3)	0.873	22	21	[32]
bleomycin resistant (3)	0.871	5	4	[38]
morphology: branched (diploid)	0.870	5	5	[4]
sensitivity at 15 generations in 1.5 M sorbitol	0.867	6	4	[4]
caspofungin resistant	0.866	8	8	[31]
inviable (essential)	0.845	1100	1027	[4, 29]
shortened telomeres (3)	0.843	20	18	[39]
sensitivity at 15 generations in minimal +his +leu +ura medium	0.843	77	70	[4]
MMS sensitive (3)	0.837	78	73	[40]
cellular chitin level reduced (2)	0.835	17	17	[32]
Petite	0.833	179	166	[41]
sensitivity at 5 generations in minimal +his +leu +ura medium	0.827	62	51	[4]
long telomeres (3)	0.824	6	6	[39]
decreased calcofluor white resistance	0.814	65	63	[37, 42]
Growth defect on a fermentable carbon source	0.812	257	249	[43]
transposon cDNA expression changed (high)	0.810	27	26	[44]
morphology: clumpy (3)(diploid)	0.802	18	18	[4]
gamma radiation sensitive (3)	0.793	31	31	[45]
cell cycle arrest defective and defective shmoo	0.782	30	29	[34]
sensitivity at 5 generations in galactose	0.781	11	10	[4]
small (haploid)	0.778	215	192	[46]
retrotransposition reduced	0.772	99	89	[44]
K1 killer toxin sensitive (40%)	0.770	72	72	[42]
increased iron uptake	0.757	76	70	[47]
Growth defect on a non-fermentable carbon source	0.755	498	448	[43]

gentamycin sensitive (high)	0.754	11	11	[48]
proteasome inhibitor sens (high)	0.753	22	22	[49]
reduced fitness in rich medium (YPD)	0.748	891	872	[37]
mycophenolic acid sensitive	0.746	38	33	[50]
axial budding	0.745	4	4	[14]
morphology: elongate (3) (diploid)	0.739	77	73	[4]
sporulation deficient	0.738	261	244	[51]
random budding (high)	0.737	74	72	[14]
large (haploid)	0.728	227	205	[46]
reduced sporulation (3) (normal respiration)	0.722	136	119	[52]
bleomycin sensitive (4)	0.721	58	55	[38]
sensitivity at 5 generations in synthetic complete - lys medium	0.715	23	22	[4]
decreased rapamycin resistance	0.707	272	256	[53]
whi	0.706	19	19	[41]
sensitivity at 5 generations in 1.5 M sorbitol	0.704	13	11	[4]
decreased wortmannin resistance	0.703	89	85	[53]
sensitivity at 20 generations in 1 M NaCl	0.703	63	59	[4]
K1 killer toxin resistant (40%)	0.698	19	18	[42]
morphology: round (3) (diploid)	0.696	105	99	[4]
uge	0.694	28	26	[41]
sensitivity at 5 generations in synthetic complete - trp medium	0.694	48	45	[4]
sensitivity at 5 generations in 1 M NaCl	0.693	60	56	[4]
rapamycin resist (2)	0.692	26	26	[54]
reduced iron uptake	0.688	5	5	[47]
rate of growth loss of growth in 0.85 M NaCl	0.682	212	189	[55]
sensitivity at 5 generations in medium of pH 8	0.677	102	93	[4]
sensitivity at 15 generations in medium of pH 8	0.676	128	115	[4]
morphology: small (3)(diploid)	0.672	79	69	[4]
sensitivity at 15 generations in 10 uM nystatin	0.672	28	27	[4]
morphology: large (3)(diploid)	0.669	88	80	[4]
reduced glycogen storage (2)	0.666	44	41	[56]
sensitivity at 5 generations in 10 uM nystatin	0.666	124	108	[4]
increased rapamycin resistance	0.662	114	100	[53]
morphology: unusual shmoo (haploid)	0.661	29	25	[34]
morphology: polarized bud growth (haploid)	0.657	5	5	[34]
wortmannin resistant (5)	0.656	25	23	[57]
sensitivity at 5 generations in synthetic complete - thr medium	0.647	31	29	[4]
enhanced glycogen storage (2)	0.645	61	55	[56]
proteasome inhibitor resistant	0.642	7	6	[49]
reduced spores per ascus	0.641	37	34	[52]
rate of growth sensitivity in 0.85 M NaCl	0.629	209	191	[55]
morphology: football (3) (diploid)	0.628	59	53	[4]
germination deficient	0.627	158	147	[51]
sporulation promoting	0.622	102	98	[51]

6AU sensitive (3)	0.618	28	26	[58]
increased wortmannin resistance	0.617	80	75	[53]
morphology: elongated (haploid)	0.603	110	101	[34]
rapamycin sensitive (4)	0.597	20	20	[54]
efficiency of growth sensitivity in 0.85 M NaCl	0.597	65	58	[55]
decreased rapamycin resistance	0.597	8	7	[53]
slow growth in YPD (16x below WT)	0.585	23	22	[4]
MPA sensitive (3)	0.563	24	22	[58]
morphology: round (haploid)	0.552	13	11	[34]
efficiency of growth resistance in 0.85 M NaCl	0.541	44	40	[55]
sensitivity at 5 generations in synthetic complete medium	0.531	88	78	[4]
morphology: large (haploid)	0.527	23	21	[34]
adaptation time loss of growth in 0.85 M NaCl	0.526	103	91	[55]
adaptation time sensitivity in 0.85 M NaCl	0.521	284	259	[55]
decreased sensitivity to the anticancer drug, cisplatin	0.512	22	19	[59]
morphology: chain (diploid)	0.485	5	5	[4]
morphology: small (haploid)	0.480	94	89	[34]
rate of growth resistance in 0.85 M NaCl	0.479	59	49	[55]
morphology: clumped (haploid)	0.479	32	28	[34]
adaptation time resistance in 0.85 M NaCl	0.465	69	60	[55]
efficiency of growth loss of growth in 0.85 M NaCl	0.464	23	21	[55]
morphology: pointed (haploid)	0.453	99	88	[34]

[a]Numbers in parentheses indicate threshold applied to generate seed set, e.g., (3) indicates '+++' or '---', as appropriate.

Each phenotype was evaluated for the area under the ROC curve (a measure of predictability), as described in the methods. Specifically, I used hits from these screens as seed sets for predicting the associated phenotypes from the yeast network, performing leave-one-out cross-validation, just as for the prediction of essential genes. The network's predictive power for a representative assortment of these phenotypes is displayed in **Figure 3.2**. In order to evaluate the overall trends in these data, I calculated the area under each of the 100 ROC curves (AUC) as a measure of prediction strength—an AUC value of 0.5 indicates random performance, while an AUC value of 1.0 indicates perfect predictions. I found that a majority of phenotypes are reasonably predictable (**Figure 3.3**), with 70% of the phenotypes predictable at AUC > 0.65. In contrast, none of 100 random gene sets of the same sizes as the actual phenotypic seed sets exhibited AUC > 0.65. The AUC of the highest scoring random set was 0.64, which indicates that phenotypes with AUC > 0.65 were significant to at least $p < 0.01$. In order to contrast the AUC of real phenotypes derived from the literature with what would be expected under a random distribution, I generated equivalent sets of random genes by drawing from the complete set genes labeled by SGD [15] as verified or uncharacterized (as of January 29, 2007). The random sets were size matched to the real phenotype sets.

A wide range of phenotypes are highly predictable, including: shortened telomeres [39], increased secretion of CPY protein [33] (an indicator of disruption of sorting in the secretory system), and chitin accumulation [32]. Many categories of phenotype are at least moderately predictable, including both very specific phenotypes, such as iron uptake [47] and caspofungin sensitivity [31], and broader phenotypes like gross cellular morphology (small cells [46], round cells [4], etc.). Surprisingly, there is little dependence of predictability on the size of the seed set (**Figure 3.4**), and I observed

strong predictability for both large and small seed sets (e.g., bleomycin resistance [38], n = 4 genes, AUC = 0.87 versus nonviability/essential [4, 29], n = 1027 genes, AUC = 0.85).

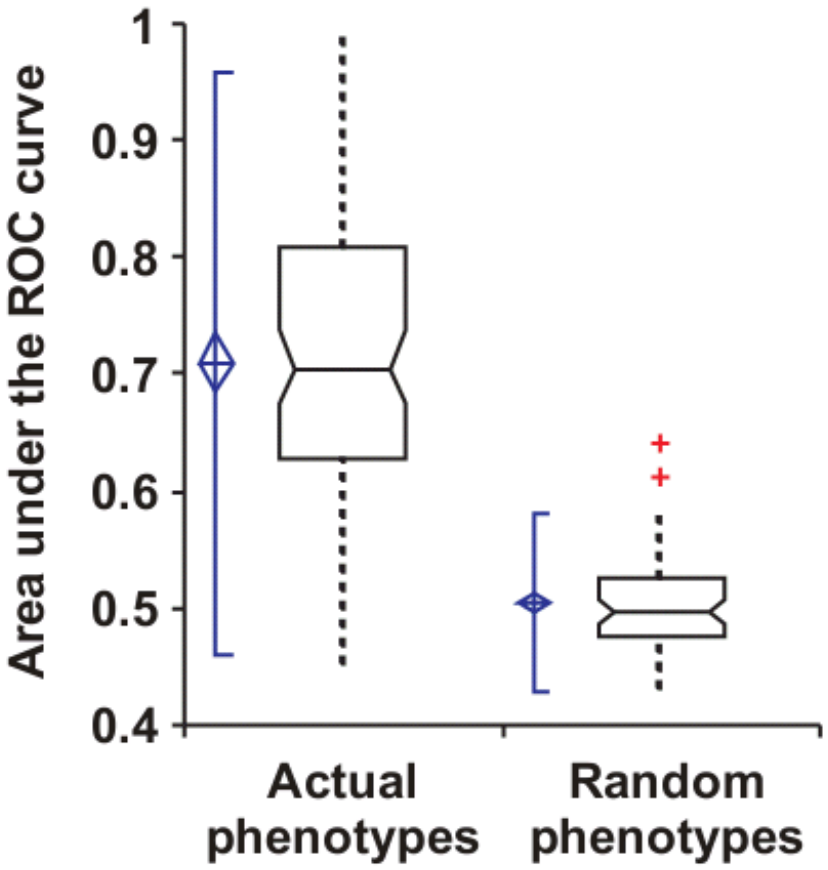


FIGURE 3.3 LOSS-OF-FUNCTION PHENOTYPES ARE PREDICTED SIGNIFICANTLY BETTER THAN RANDOM EXPECTATION. Here, predictability is measured as the area under a ROC curve (AUC), measuring the AUC for each of 100 yeast phenotypes observed in genome-wide screens and plotting the resulting AUC distributions. Real phenotypes are significantly more predictable than size-matched random gene sets. At the left of each box-and-whisker plot, the center of the blue diamond indicates the AUC mean, the top and bottom of the diamond indicate the 95% confidence interval, and the accompanying solid vertical line indicates +/- 2 standard deviations. The bottom, middle, and top horizontal lines of the box-and-whisker plots represent the 1st quartile, the median, and the 3rd quartile of AUCs, respectively; whiskers indicate 1.5 times the interquartile range. Red plus signs represent individual outliers. Figure used by permission [1].

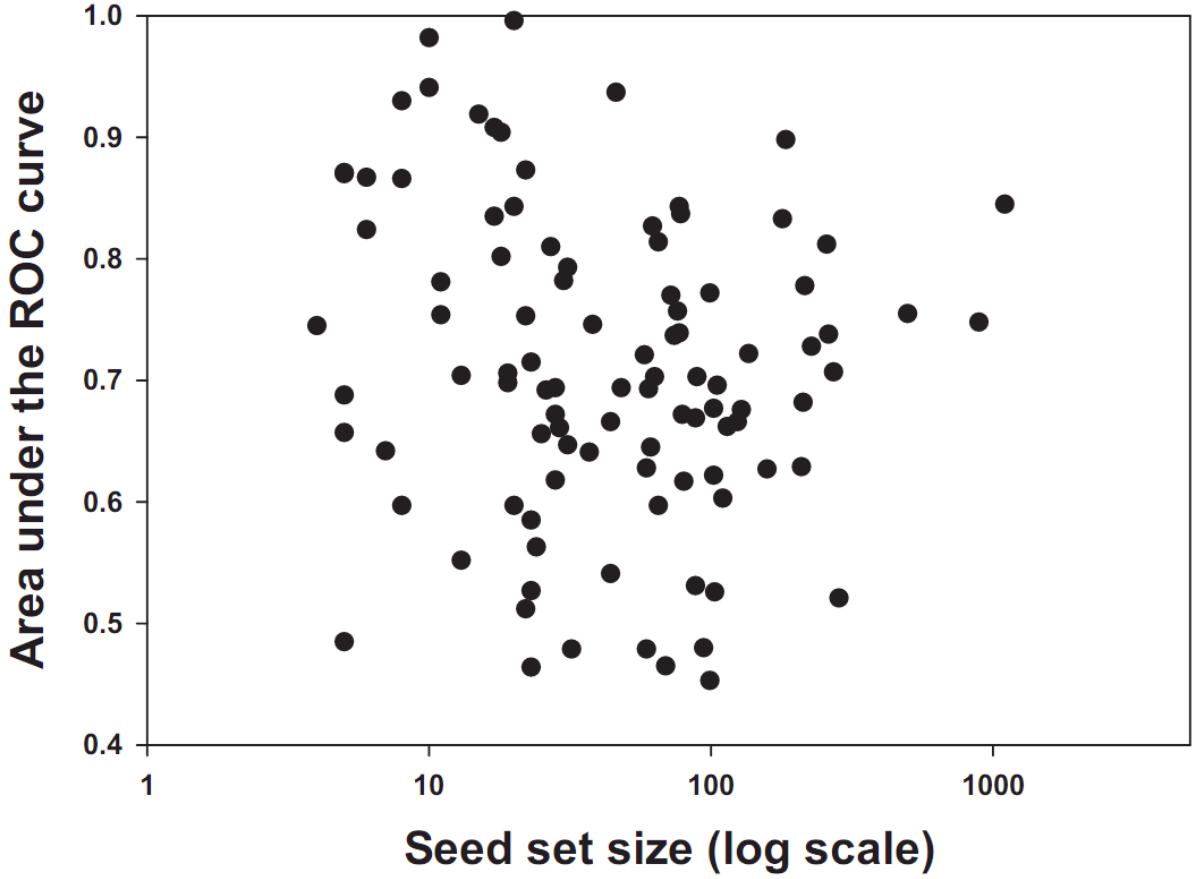


FIGURE 3.4 A PLOT OF SEED SET SIZE VERSUS PREDICTABILITY OF THE PHENOTYPE SHOWS NO SIGNIFICANT CORRELATION. Thus, there does not appear to be an intrinsic limitation for applying network-guided reverse genetics even when seed set size is small. Each filled circle indicates the prediction strength (AUC, as calculated in Figure 3.3) of one of the 100 loss-of-function phenotypes relative to the number of genes in that seed set. Figure used by permission [1].

Integration of functional genomics and proteomics data is important for phenotype prediction

As physically interacting proteins often share related genetic interaction partners (e.g., [60, 61]) and even human disease associations [25, 62, 63], it seemed likely that physical protein interactions might account for a large fraction of the signal I observe. In particular, Lage *et al* [62] has used guilt-by-association among protein complexes to predict disease genes within human genetic linkage groups. Balancing this trend, phenotypes of annotated genes are in part predictable directly from the annotations [64]. Thus, I asked if the integration of functional genomics and proteomics data in the functional network brought additional predictive power over physical interactions alone. To compare the predictive accuracy of the functional network to protein interaction networks, I repeated the GBA analysis with several protein interaction networks [17-19, 21]. I used any weightings reported with the interactions and weighted all interactions equally in the absence of a reported interaction probability. I measured the median AUC across the same 100 phenotypes discussed above for the functional yeast gene network and for each of several versions of the yeast protein physical interaction network [17-19, 21]. I compared these values to the median fraction of each seed gene set covered by the respective networks. The values of AUC and fraction covered therefore serve as measures of precision and recall for each network. As **Figure 3.5** demonstrates, I observe that all networks predict loss-of-function phenotypes to some extent, but find the functional network to predict phenotypes at a significantly higher precision and recall.

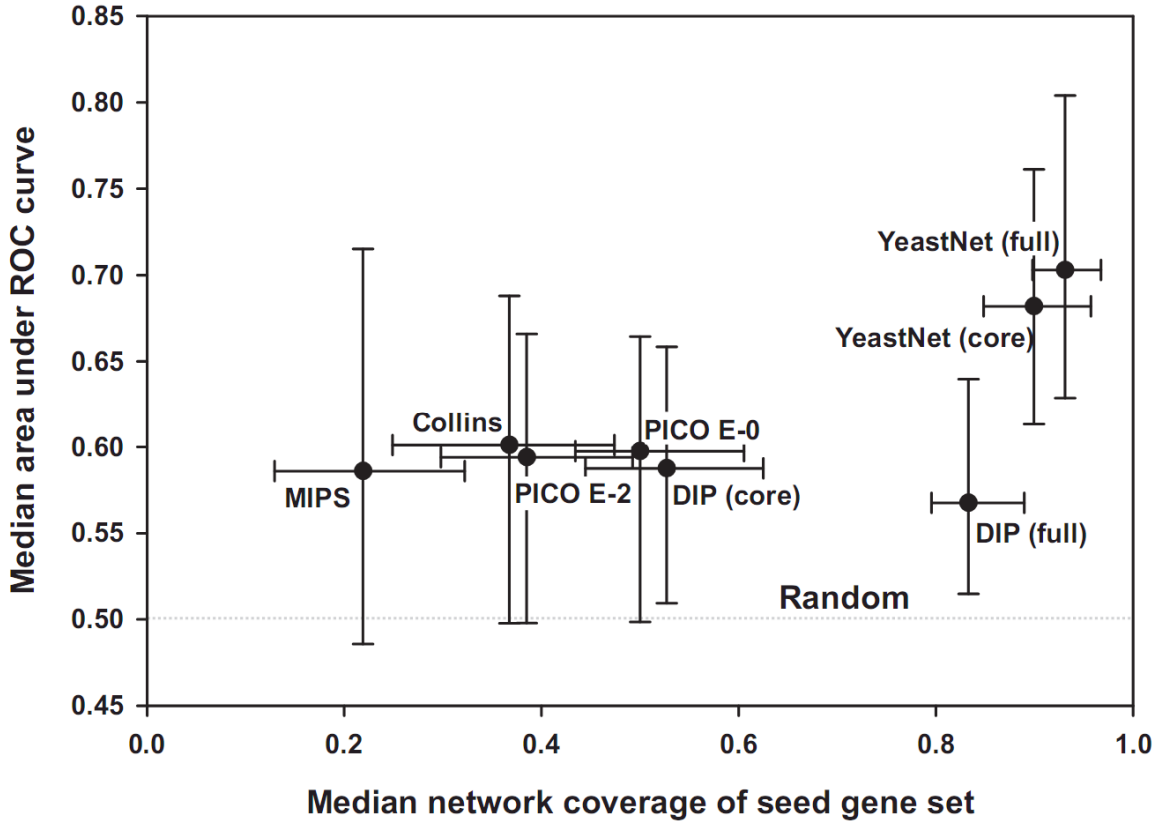


FIGURE 3.5 FUNCTIONAL NETWORKS HAVE GREATER PREDICTIVE POWER FOR PHENOTYPE THAN PHYSICAL PROTEIN NETWORKS. Median values of predictive power (AUC) across 100 loss-of-function phenotypes are plotted versus the median fraction of each seed gene set covered by a network (coverage; measured as the fraction of seed genes with at least one linkage in the network). Five networks are compared: the functional yeast network (YeastNet v. 2 [11]) and four versions of the network of yeast physical protein interactions (DIP [18], PICO [19], MIPS physical complexes [17], and Collins et al. [21]). DIP, PICO, and YeastNet are each evaluated at their two reported confidence thresholds. The YeastNet functional gene network shows considerably higher predictive power than for the networks composed only of physical interactions; the full YeastNet shows higher predictive power than a more confident core set of the top 47,000 linkages, indicating that the lower confidence linkages nonetheless add predictive power. Error bars indicate the 1st and 3rd quartiles. Figure used by permission [1].

I attribute this enhanced performance to the increased comprehensiveness of the functional gene network, both in terms of additional types of gene associations as well as more extensive coverage of the overall set of yeast genes. The functional network accomplishes this by incorporating other sources of functional interaction (e.g., mRNA co-expression) in addition to physical interactions from both small scale (e.g., the DIP and MIPS databases) and genome scale (e.g., mass spectrometry of affinity-purified protein complexes and yeast two hybrid) experiments. **Figure 3.6** illustrates two sub-networks that contribute to the prediction of two difference phenotypes, one of which depends heavily on the physical interaction data **(A)**, but the other does not **(B)**. This suggests that many phenotypes are the result of more than the disruption of protein complexes or interactions. The integration of many types of data, including protein interactions, allows the functional network to predict phenotypes regardless of the underlying mechanism that leads to the phenotype. This is a clear example that integration of multiple types of functional data improves the predictive accuracy and gene coverage of the functional network over the underlying datasets.

Further, as shown in **Figure 3.7**, the sequential addition of progressively lower confidence functional linkages increases both predictive accuracy and coverage. Low confidence linkages do not undercut the predictive power of high confidence linkages because they are weighted in proportion to the strength of the evidence that supports them. These observations highlight the importance of integrating diverse data types within a consistent statistical frame work and suggest that the proteins encoded by genes associated with the same phenotype often may not physically interact. As additional functional data are collected by the scientific community, they can be quickly integrated

within the network framework to generate more accurate functional networks, and, as shown in **Figure 3.8**, more accurate phenotype predictions.

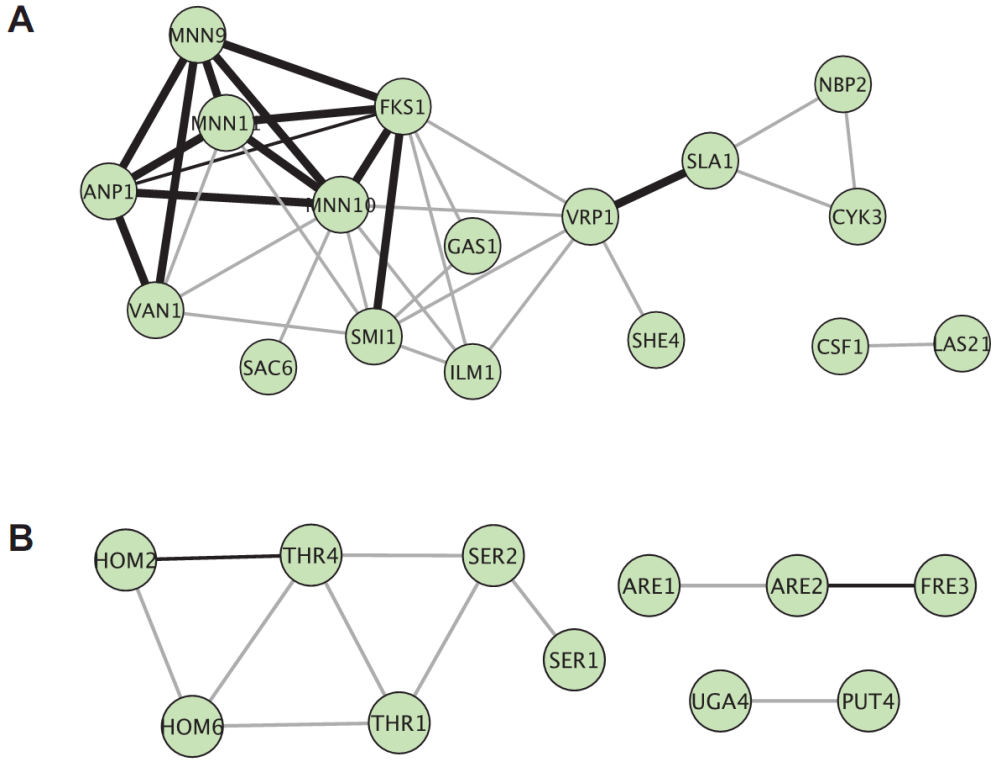


FIGURE 3.6 PREDICTIVE POWER OF FUNCTIONAL NETWORK RELIES ON PHYSICAL AND FUNCTIONAL INFORMATION. (A) and (B) show example seed gene sets (green circles) and their network connections, indicating functional linkages in grey lines, physical interactions in thin black lines, and both functional and physical interactions in thick black lines. (A) shows genes whose deletion increases cellular chitin levels [32] (AUC = 0.87), whose prediction relies upon a mix of physical and functional interactions. (B) shows genes whose deletion confers sensitivity at 5 generations in synthetic complete medium lacking threonine [4] (AUC = 0.65), whose prediction derives predominantly from functional linkages. Network drawn with Cytoscape [12]. Figure used by permission [1].

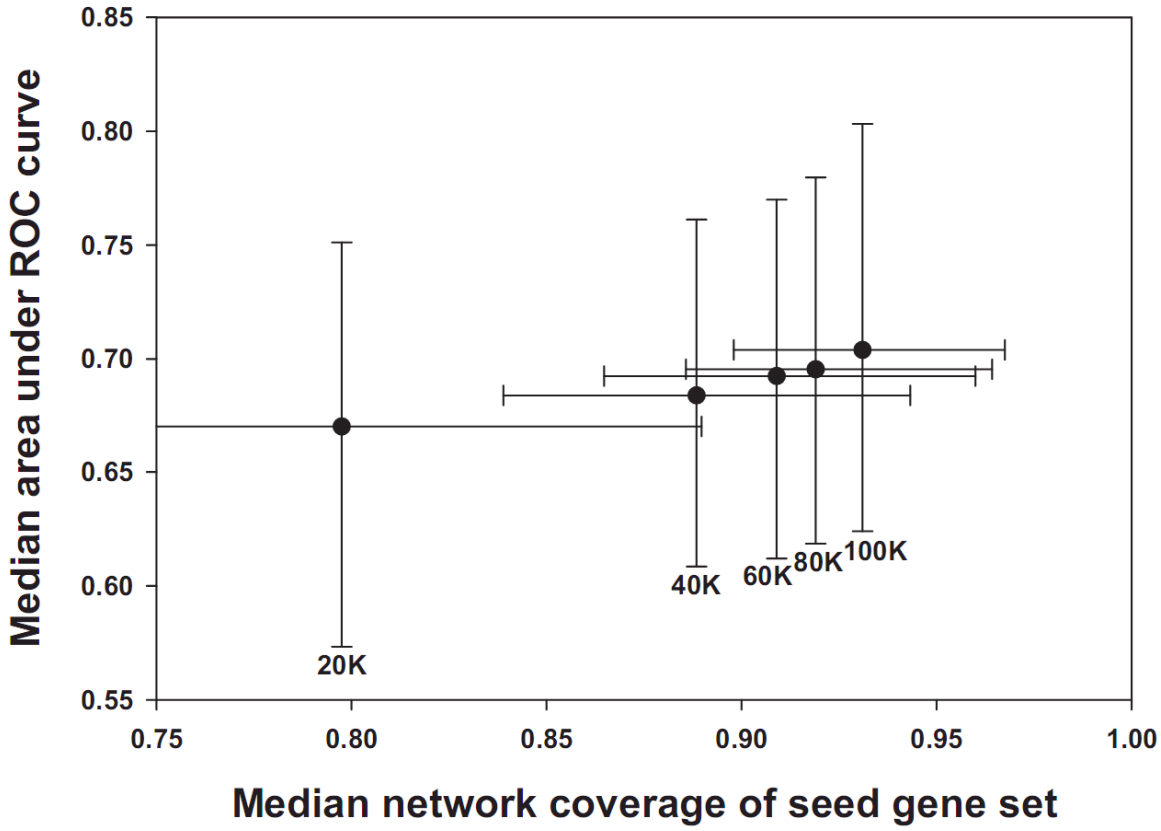


FIGURE 3.7 LOWER PROBABILITY LINKAGES CONTINUE TO IMPROVE PREDICTIVE ACCURACY, ALBEIT WITH DIMINISHING RETURNS, shown in a plot of the predictive accuracy (median AUC across the 100 phenotypes, calculated as in Figure 3.3) versus median network coverage of the 100 phenotype sets, as calculated for the top-ranked 20000 (20K), 40000 (40K), etc. linkages in YeastNet v.2. This trend derives from the fact that all links in this network have at least a 60% probability of linking genes in the same pathway. The probabilistic nature of the network means that low confidence linkages are unlikely to undercut high confidence linkages during phenotype prediction because the links are weighted according to the strength of the evidence supporting them. Error bars indicate the 1st and 3rd quartiles. Figure used by permission [1].

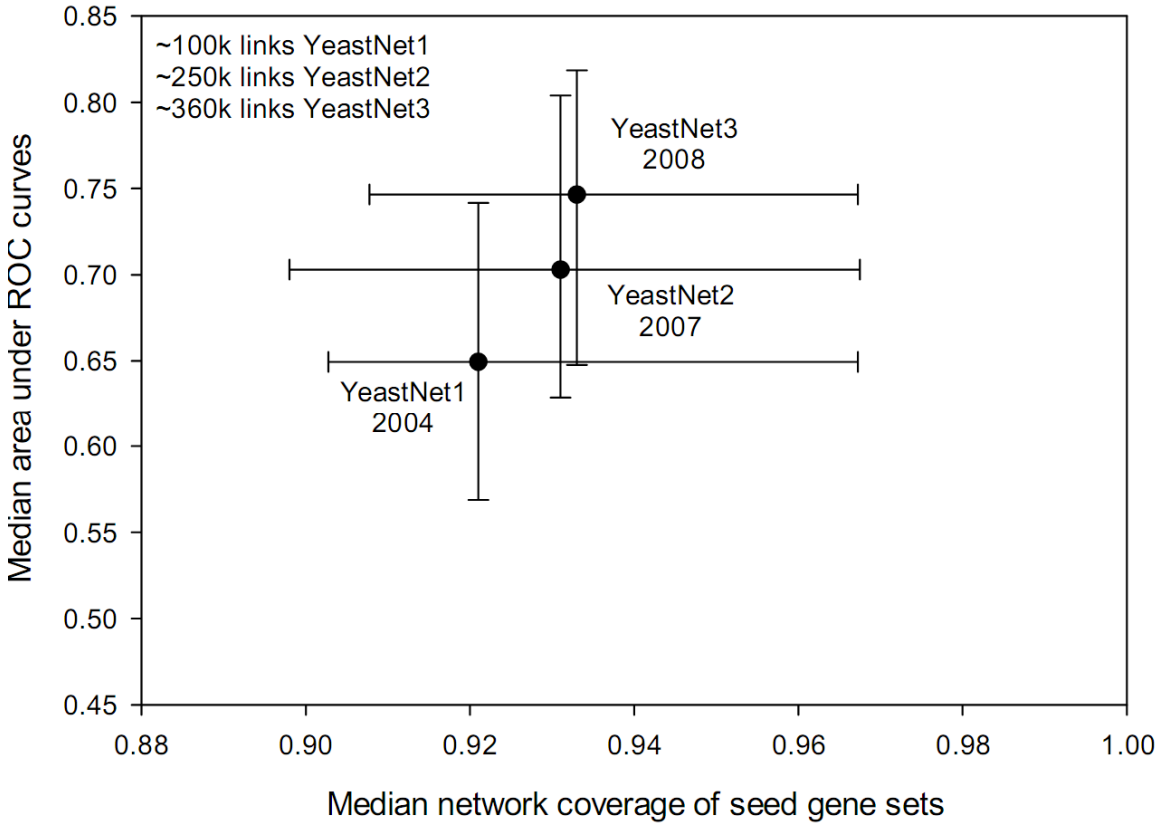


FIGURE 3.8 ITERATIONS OF THE FUNCTIONAL NETWORK IMPROVE PHENOTYPE PREDICTION. Additional data integrated into the functional network improves phenotype prediction over time, with new links improving both predictive accuracy and coverage of genes with known phenotypes. YeastNet1 [9], YeastNet2 [11], and YeastNet3 [65] use the same basic approach to integrate data, but incorporate different data sets and use slightly different methods to insure quality control.

Prediction of quantitative cell morphology phenotypes

Having established the predictive power of the network for largely qualitative phenotypes, I chose to I apply the same approach to the quantitative, morphological phenotypes. Given that the phenotypes analyzed thus far are often based on subjective criteria (*i.e.*, judged to be elongated or not), it is important to ask if such predictions can be made for quantitative phenotypes. So, I tested the predictive power of the GBA approach on quantitative cell morphology data reported by the *Saccharomyces cerevisiae* morphology database (SCMD) [16], which were systematically measured for the set of haploid MATa yeast deletion strains [66]. 281 quantitative features of cell shape, cellular, and subcellular morphology were measured for each strain, including such parameters as the ratio of long cell axis to short cell axis, the angle between a mother cell and bud, and the relative distribution of actin with regards to the bud position. Each feature was measured for many cells from a given strain, and the mean value reported. For 220 of the features, the coefficient of variance (CV) was also reported, describing the variability in that feature across single cells in that strain. Considering the mean value of each feature and the CV as separate traits (the former will be referred to as morphology phenotypes and the latter as CV phenotypes) means that a total of 501 cell shape measurements or CVs were reported for 4,718 strains.

As not all measurable cell shape features are likely to be under selection (for example, they might simply vary stochastically yet neutrally), I do not expect all such phenotypes to correspond to functional pathways and therefore be predictable. Nonetheless, one might expect that a number of these would have functional correlates

and therefore be predictable. In order to test this notion, I therefore evaluated each of the 501 features for predictability using the functional gene network.

To generate seed gene sets from these data, for each of the 281 quantitative features I selected as phenotypic seed sets the 40 genes with the highest measured mean value of that feature and the 40 genes with the lowest measured mean value of that feature, in all generating 562 morphology phenotype seed gene sets (281 features x 2 seed sets each). I then evaluated each of these seed sets for predictability using ROC analysis. As for the 100 genome-wide phenotypic screens, I observed many strongly predictable cell morphology phenotypes, such as those illustrated in **Figure 3.9**. For example, one of the most strongly predictable cell morphology phenotypes is for the genes whose disruption most increases cell ellipticity during nuclear migration to the bud neck (AUC = 0.87). Another strongly predictable phenotype is for deletion strains showing the highest increase in the actin polarization of unbudded cells (AUC = 0.80).

The AUC distribution of 562 quantitative phenotypes from SCMD is compared to the background random distribution in the next figure **(Figure 3.10)**. Although many morphological phenotypes overlap with the random distribution, a significant portion of the phenotypes are more predictable than explainable by chance**.** For the SCMD phenotype sets, the model for random expectation was generated by drawing 1000 sets of 40 genes from the set of genes SMCD analyzed and calculating the AUC for each random set. Note that predictability does not depend strongly on the size of the seed sets; I see similar predictive power with seed sets of 10 - 80 genes (data not shown). These results confirm that even specific quantitative aspects of yeast cell shape often have

functional correlates, and therefore the sets of genes whose disruption most affects such features are predictable.

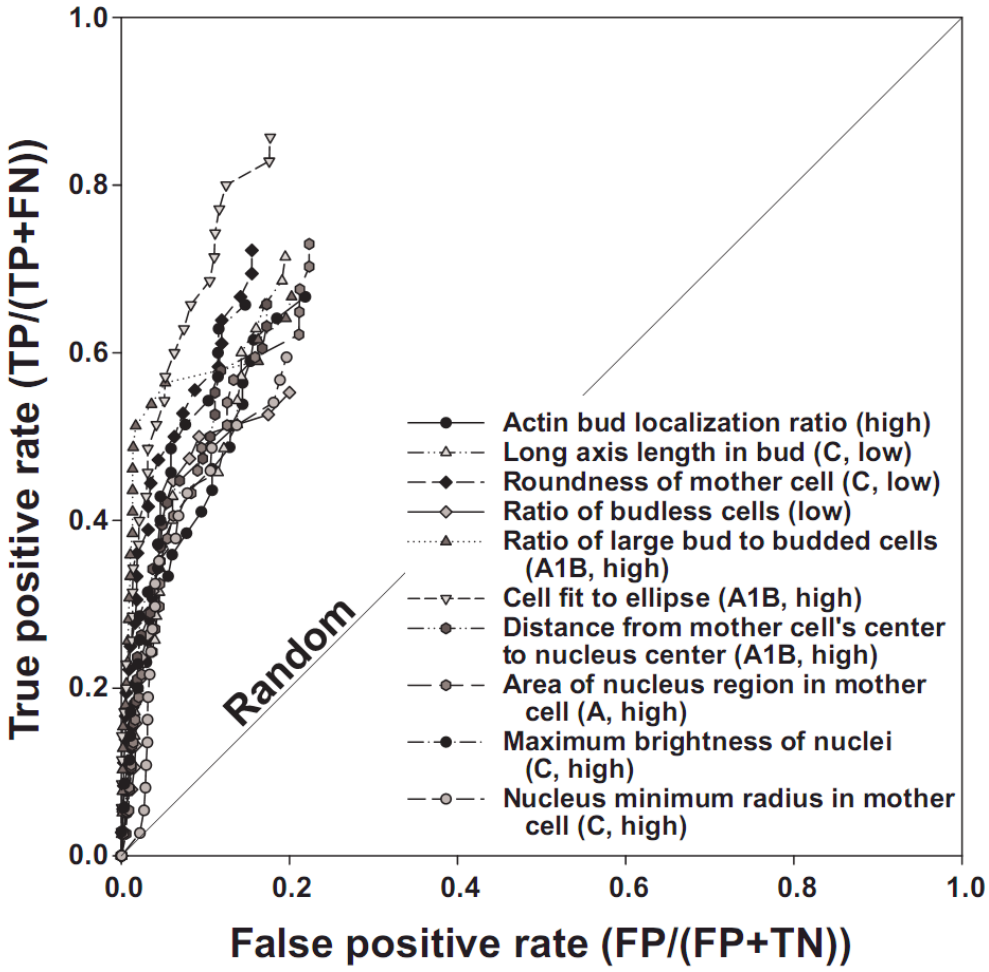


FIGURE 3.9 NETWORK-BASED PREDICTIONS OF QUANTITATIVE CELL MORPHOLOGY PHENOTYPES. A WIDE VARIETY OF PHENOTYPES BASED UPON QUANTITATIVE YEAST CELL SHAPE AND INTRACELLULAR FEATURES ARE PREDICTABLE [66], as shown for the 10 phenotypes in this ROC analysis (selected from SCMD phenotypes with AUC > 0.68). For each of the features, the 40 genes whose deletion mutants show either the 40 highest or 40 lowest values for that quantitative feature (indicated by "high" or "low", respectively) were selected as the seed gene set. Predictability was evaluated using ROC analysis as in Figure 2, plotting the true positive prediction rate versus false positive rate, using leave-one-out cross-validation. For clarity, the line connecting the final point of each graph to the top right corner has been omitted. Labels of features are adapted for clarity from the Saccharomyces cerevisiae Morphology Database [16]; the SCMD labels A, A1B, and C, represent unbudded cells, budded cell with one nucleus in mother cell, and large-budded post-mitotic cells with nuclei in both mother and daughter cell, respectively. Ratio measurements refer to proportions across a population of cells. Figure used by permission [1].

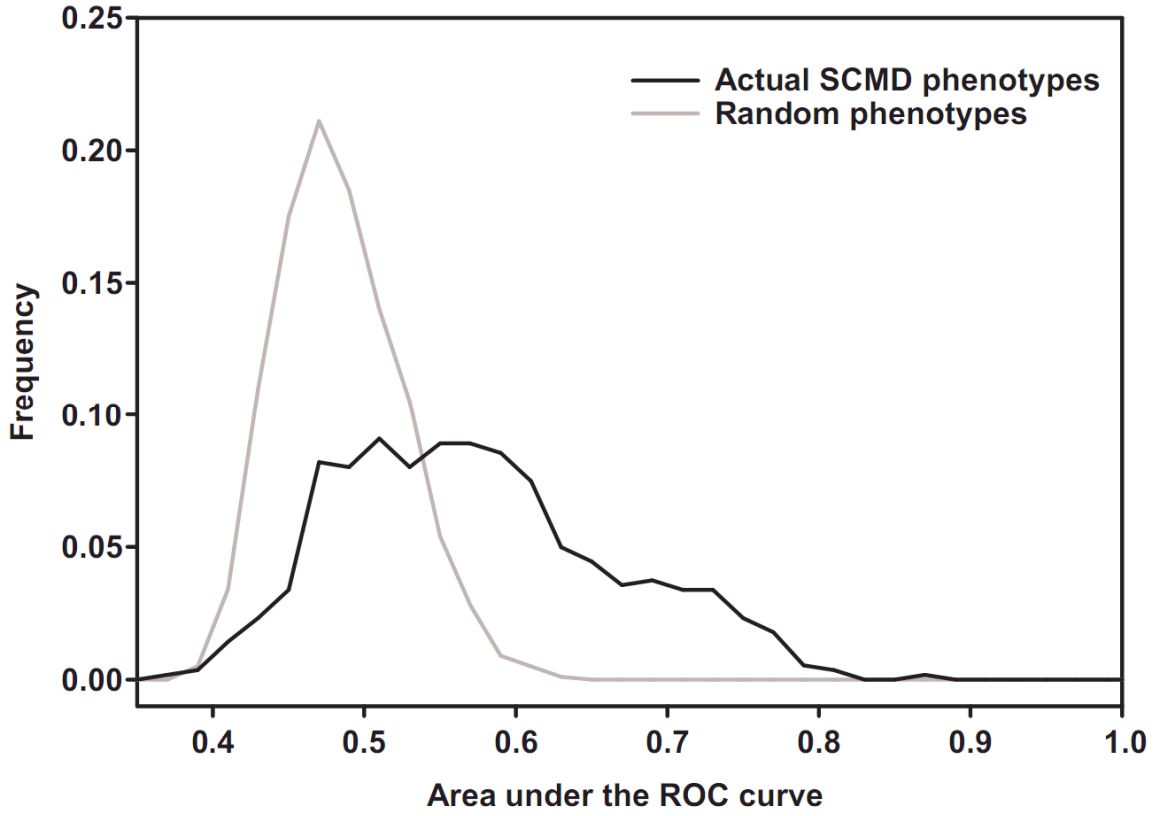


FIGURE 3.10 PREDICTIONS OF QUANTITATIVE CELL MORPHOLOGY PHENOTYPES ARE SIGNIFICANTLY BETTER THAN RANDOM. A histogram plotting the distribution of the AUC values for 562 quantitative morphological phenotypes shows a significantly higher proportion of high AUC values than for 1,000 size-matched random gene sets. Figure used by permission [1].

Genes increasing cell-to-cell variation are less functionally coherent than those decreasing variation

With two types of quantitative traits in SCMD, the traits themselves and their variance, I decided to see if subsets of the phenotypes were differentially predictable. Specifically, I asked if the coefficient of variance of a yeast morphology phenotype across single cells in a population was itself a predictable phenotype. Strikingly, I observed good predictability for sets of genes whose disruption most *increased* the CV of a given morphological feature (e.g., the 40 genes whose deletion caused the highest increase in bud neck width CV; AUC = 0.70), but near random prediction for sets of genes whose disruption most *decreased* the CV in a given morphological feature (e.g., the 40 genes whose deletion most reduced bud neck width CV; AUC = 0.54) (**Figure 3.11**). The high CV phenotypes are significantly more predictable than the low CV phenotypes ($p < 0.0001$, Wilcoxon signed-ranks test). Across the 220 high CV phenotypes, I observed 116 to show significantly greater AUC values than size-matched random sets (at the 95% confidence level as judged by Z-score > 1.95), while only 26 of the set of 220 low CV phenotypes were better than random at this level.

Upon further analysis, it became clear that while quantitative, morphological phenotypes were predictable at both the high and low ends of the measured distribution, only high, but not low, coefficient-of-variance phenotypes are predictable. As successful prediction of a loss-of-function phenotype implies functional coherence of the genes—essentially reflecting clustering of the genes in the functional network—this result indicates that the genes whose disruption most strongly reduced the CV in a given morphological feature do not in general form a functionally coherent set. By contrast, genes whose disruption most increased morphological phenotypic variability were

predictable, and thus functionally coherent. I further observed that the same genes tended to be present in the phenotypic sets from many different CV phenotypes—*i.e.*, there are particular genes whose deletion increases the coefficient of variance of a large number of otherwise unrelated morphological properties.

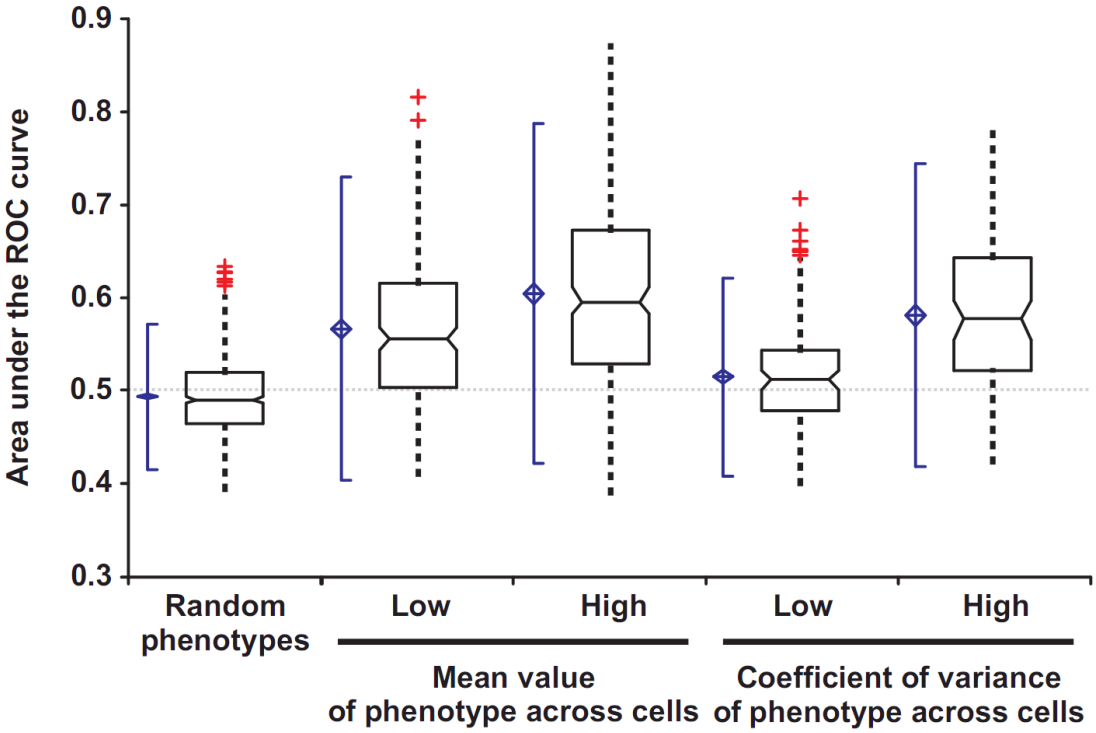


FIGURE 3.11 GENES WHOSE DISRUPTION DECREASES POPULATION CO-EFFICIENT OF VARIANCE (CV) ARE ESSENTIALLY RANDOM. Separate analyses of phenotypes associated with morphological features and phenotypes associated with cell-to-cell variability in the morphological features reveals asymmetry in predictability. Sets of genes whose disruption causes the 40 largest or smallest mean values of a morphological feature (middle plots) are significantly more predictable than random gene sets (left side). By contrast, while the sets of genes whose disruption most increase the CV tend to be predictable (high AUC), those that most decrease the CV are not (low AUC). Box-and-whisker plots are drawn as in Figure 3.3. Figure used by permission [1].

To further explore this observation, for each of the 4,718 yeast genes in the SCMD data set, I calculated the median percentile rank across each of the 220 SCMD CV phenotypes. Thus, the gene whose deletion strain has the highest median percentile rank (the telomere length regulation gene EST1; median percentile rank of 0.98) exhibits the highest cell-to-cell variability across nearly all of the set of 220 CV phenotypes. By contrast, the gene whose deletion strain has the lowest median percentile rank (YAL004W, a small open reading frame that overlaps the coding sequence for the HSP70 family chaperone SSA1; median percentile rank 0.17) consistently exhibits the lowest cell-to-cell variability for the tested phenotypes. Thus, these rankings capture the generic tendency for a gene to increase or decrease cell-to-cell variability across many measured morphology parameters. I tested the top-ranked 40 genes and the bottom-ranked 40 genes for their network-based predictability.

As with my earlier observations, the top-ranked 40 genes (those with highest median percentile rank) show reasonable predictability (AUC = 0.71), while the bottom-ranked 40 genes show random predictability (AUC = 0.49). Thus, either on a phenotype-by-phenotype basis, or across all 220 phenotypes, genes whose disruption most increased morphological phenotypic variability tended to be more predictable and functionally coherent than those that reduced phenotypic variability. By comparing the distribution of the median percentile rank of all 220 coefficient of variance phenotypes for each of the 4,718 knockout strain to 127 wild-type replicates, I were able to identify a cluster of genes involved in DNA repair that consistently lead to higher coefficients of variance for many phenotypes (**Figure 3.12**). The top-ranked set of 40 genes show strong enrichment for specific Gene Ontology terms, with 17 of the 40 genes encoding nuclear proteins ($p < 10^{-6}$; measured using FunSpec [67]); 10 of these are DNA binding proteins ($p < 10^{-4}$),

including genes of DNA recombination and repair ($p < 10^{-6}$). Many of these genes are involved in maintaining genomic stability, including the repair/recombination proteins RAD27, RAD50, RAD51, RAD52, CTF4, HEX3, RTT109, and THP1, the histone HTZ1, and the telomere maintenance protein EST1. Thus, while deletions of these genes may possibly increase phenotypic variation, the most plausible biological explanation for the increased cell-to-cell variation between these strains is that they are no longer clonal due to genomic damage and rearrangement.

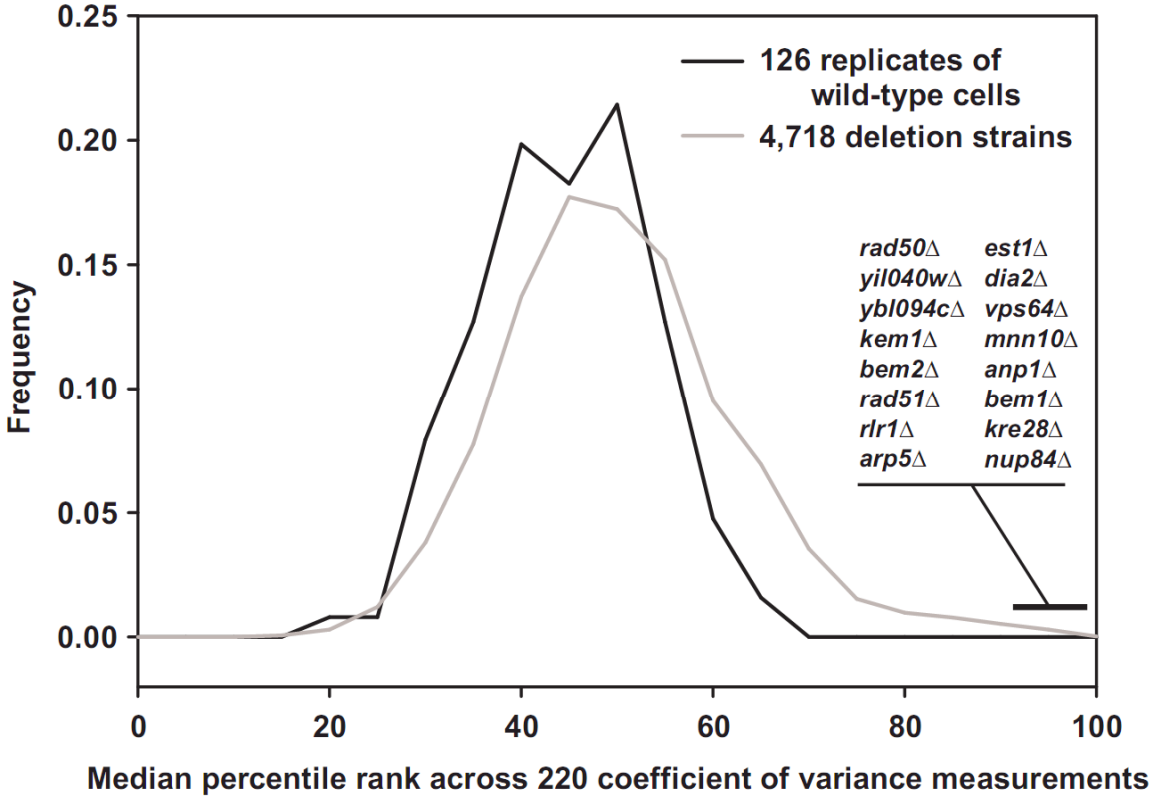


FIGURE 3.12 GENES KNOCKOUTS THAT INCREASE VARIANCE ACROSS MANY MORPHOLOGICAL TRAITS TYPICALLY AFFECT GENOMIC STABILITY. A histogram comparison of the median phenotypic CVs observed for deletion strains versus replicate analyses of wild-type cells shows that deletion strains with the most reduced CVs are essentially wild-type-like in character, while those with the most increased CVs show significantly more cell-to-cell variability than wild-type cells. These latter knockout strains carry deletions for genes predominantly involved in maintaining genomic integrity. This trend is therefore likely to have arisen from non-clonal genetic variation in these strains, recapitulating the classic mutator phenotype. Data from [16]. Figure used by permission [1].

The functional network predicts yeast orthologs of human disease genes

The network's effectiveness at predicting both qualitative and quantitative yeast phenotypes suggests the possibility of application to other organisms, such as for predicting human disease genes. I evaluated the applicability of the functional network GBA approach to predicting human disease by performing a similar AUC analysis on yeast orthologs of human diseases. I used human diseases from OMIM [68], while treating disease variants of a single disease as one category. I mapped human disease genes and yeast genes in the functional network to their human-yeast ortholog group using InParanoid. Using the method described above, I then calculated the AUC (as a measure of predictability) of the yeast orthologs of human genes for diseases that involved at least 4 human-yeast ortholog groups of which at least 4 yeast orthologs existed in YeastNet. Perhaps surprisingly, phenotype predictions from the functional network are robust across species boundaries. Yeast orthologs of human disease genes can be predicted by the functional network as demonstrated in **Figure 3.13** for 28 OMIM diseases involving four or more yeast orthologs. Not only are many of the yeast orthologs of these disease genes predictable, the median predictive accuracy of these phenotypes is even slightly higher than the genome-wide yeast phenotypes (**Figure 3.3**), a probable reflection of the fact that genes conserved between yeast and humans generally compose core cellular machinery, well-captured by the gene network. As an illustration of this, the highest scoring human disease (AUC = 0.998), leukoencephaly with vanishing white matter, results from mutations in any of the subunits of the translation initiation factor EIF2B [69]. Likewise, I observed strong predictability for hemolytic anemia (AUC = 0.89), which involves 11 ortholog groups, involved in

glycolysis and glutathione metabolism, which are linked primarily by co-expression and co-citation, with only a few physical interaction-based linkages.

Although this test was limited to diseases involving biological processes shared between human and yeast, these results support the notion that an integrated human functional network would guide the discovery of new disease genes. As I observe strong disease predictions both from protein complexes (as in leukoencephalopathy) and pathways (as in hemolytic anemia), it appears likely that a functional human gene network might offer strong predictions for genes associated with diverse human diseases, even in the absence of genetic linkage data.

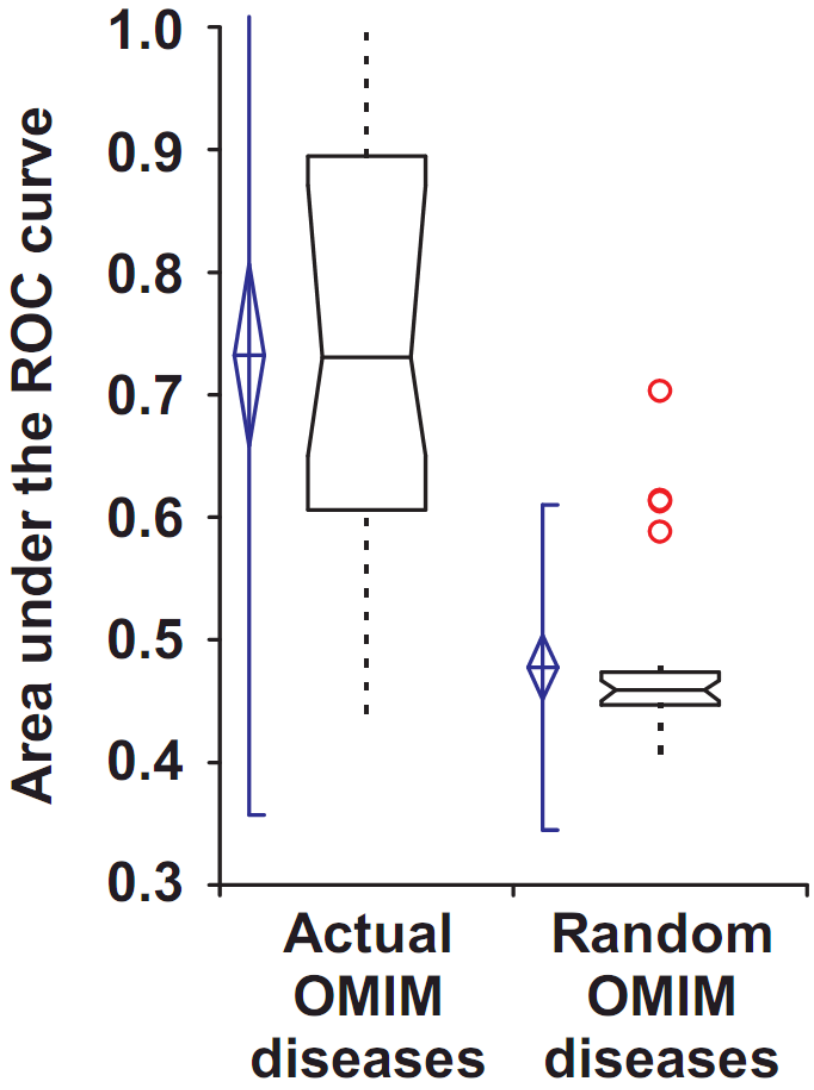


FIGURE 3.13 YEAST GENES WHOSE HUMAN ORTHOLOGS ARE LINKED TO THE SAME DISEASES ARE PREDICTED SIGNIFICANTLY BETTER THAN RANDOM EXPECTATION. Predictability is measured as the area under a ROC curve (AUC) as in Figure 3, measuring the AUC for each of 28 human diseases reported in the OMIM disease database [23] that have four or more yeast orthologs annotated in the yeast function network and plotting the resulting AUC distributions. Real disease gene sets are significantly more predictable than size-matched random gene sets drawn from the set of yeast-human orthologs. Box plots are drawn as in Figure 3.3. Figure used by permission [1].

Experimental Results

Extending a genetic screen by network-guided reverse genetics

To this point, my analysis of the predictive power of the network had used a computational approach to demonstrate its power to retrieve known results. However, for organisms in which reverse genetics is feasible, the observation that phenotypes can be predicted from network connectivity opens the possibility of extending genetic screens in a directed fashion. That is, when in possession of a set of genes known to give rise to a phenotype of interest, rather than randomly screening to identify additional genes, one could instead exploit the predictability of phenotypes by directly screening genes that are most strongly connected to the known set in the network. In this manner, experiments could be focused on the genes most likely to give rise to the phenotype. So, I sought to experimentally demonstrate the predictive power of the network using a seed set of genes, while simultaneously extending a high-throughput screen, in a fashion similar to the extension of the screen for novel shmoo localized proteins as discussed in the previous chapter. However, in this case, rather than re-testing to reduce false negatives, I wanted to screen additional genes that had not been previously tested in an initial screen. I decided to extend a previously published screen [4] of non-essential genes that result in a simple morphological defect, cell elongation, by testing essential genes. Among nonessential genes, 145 genes (3.3%) have been identified that give rise to elongated morphologies in homozygous diploid deletion strains, of which 77 genes (1.7%) show a strong phenotype [4]. I selected these 77 genes as a seed set and found the phenotype to be reasonably predictable from the network using ROC analysis (AUC = 0.74). Using the GBA method, I predicted the top 35 essential genes, which were not tested in the screen [4, 66], and were able to assay 33 of these strains for the elongate phenotype using

a tetracycline-downregulatable library. For a negative control, 17 strains from the same library not linked to known elongate genes were also assayed. I examined conditional loss-of-function strains for elongated cell morphologies, performing light microscopy of yeast strains carrying tetracycline-downregulatable alleles of each candidate gene [13]. Sixteen (~48%) of the 33 tested were elongated, as shown for several examples in **Figure 3.13**. Only one negative control was elongated, which had been previously identified by Mnaimneh *et al.* [13]. The results represent an 8-fold improvement over the negative control set and a >15-fold improvement over genome-wide screening, confirming a set of predictions experimentally while also validating the general strategy of network-guided genetic screening.

My rate of elongation identification is 8-fold higher than the negative control set and greater than 15-fold higher than the genome-wide screen, which demonstrates the utility of network-informed genetic screens and confirms that GBA phenotypic predictions based on the functional network are substantially accurate.

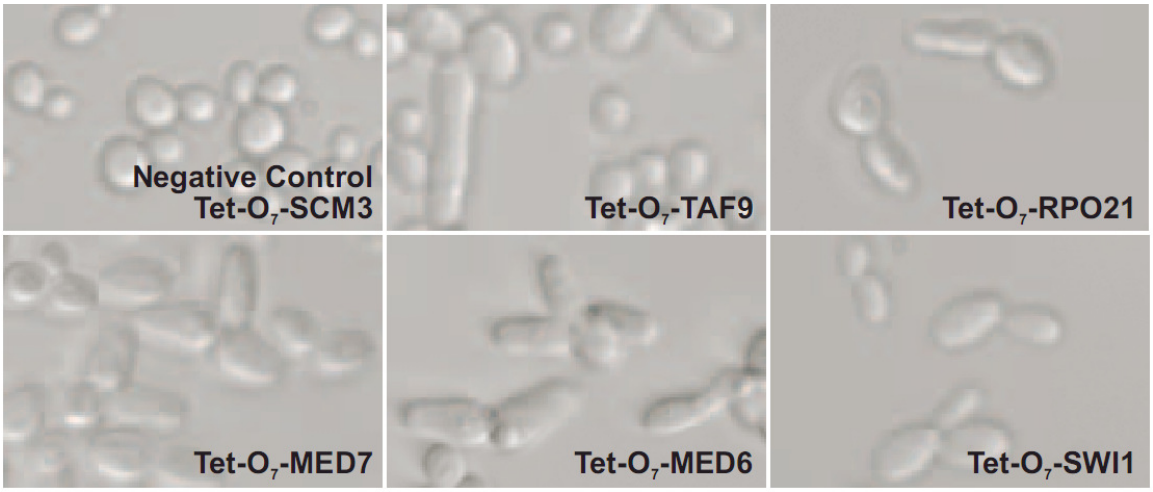


FIGURE 3.13 NETWORK-GUIDED EXTENSION OF A GENETIC SCREEN. GBA was applied to predict essential yeast genes whose disruption resulted in elongated yeast cells, based on the genes' network connectivity to a seed set of 77 nonessential genes already known to cause cell elongation when deleted [4]. Five examples of successful predictions, observed in yeast strains carrying tetracycline-downregulatable conditional alleles [13] of the essential genes TAF9, MED6, MED7, SWI1, and RPO21. By contrast, conditional down-regulation of an unrelated essential gene, SCM3, caused no such cell elongation. See Figure 3.14 for additional details. Figure used by permission [1].

To gain further insight into the genes identified, I examined the network connections among the seed genes and newly identified genes giving rise to the elongated phenotype (**Figure 3.14**). Strikingly, functional analysis of the elongate genes recovered indicates that the elongate phenotype is linked to the disruption of core transcriptional machinery, with the genes associated with elongated yeast cell morphology strongly enriched for core transcriptional functions (for example, they are significantly enriched for the MIPS [70] annotation "mRNA synthesis", $P < 10^{-12}$ [67]). The set of newly identified genes predominantly belong to the RNA polymerase II mediator complex and associated transcriptional machinery. Specifically, the genes recovered in the targeted screen are subunits of the RNA polymerase II mediator complex (MED6, MED7, both previously identified by [23], and MED8), and the transcription initiation complexes TFIID and SAGA (TAF1, TAF5, TAF9, and TAF12), complexes required for RNA polymerase II transcriptional initiation. This illustrates another advantage of network-guided genetic screening: because candidate genes are selected directly from the gene network, functional connections are often already known among the genes, helping to guide later interpretation of the hits. The relationship between an observed phenotype and the corresponding molecular defect is often mysterious: the mechanism is unknown by which defects in transcription initiation lead to elongated cells. Nevertheless, it demonstrates that the phenotypic predictions of the functional networks are accurate even when the mechanisms leading from genotype to phenotype are unclear.

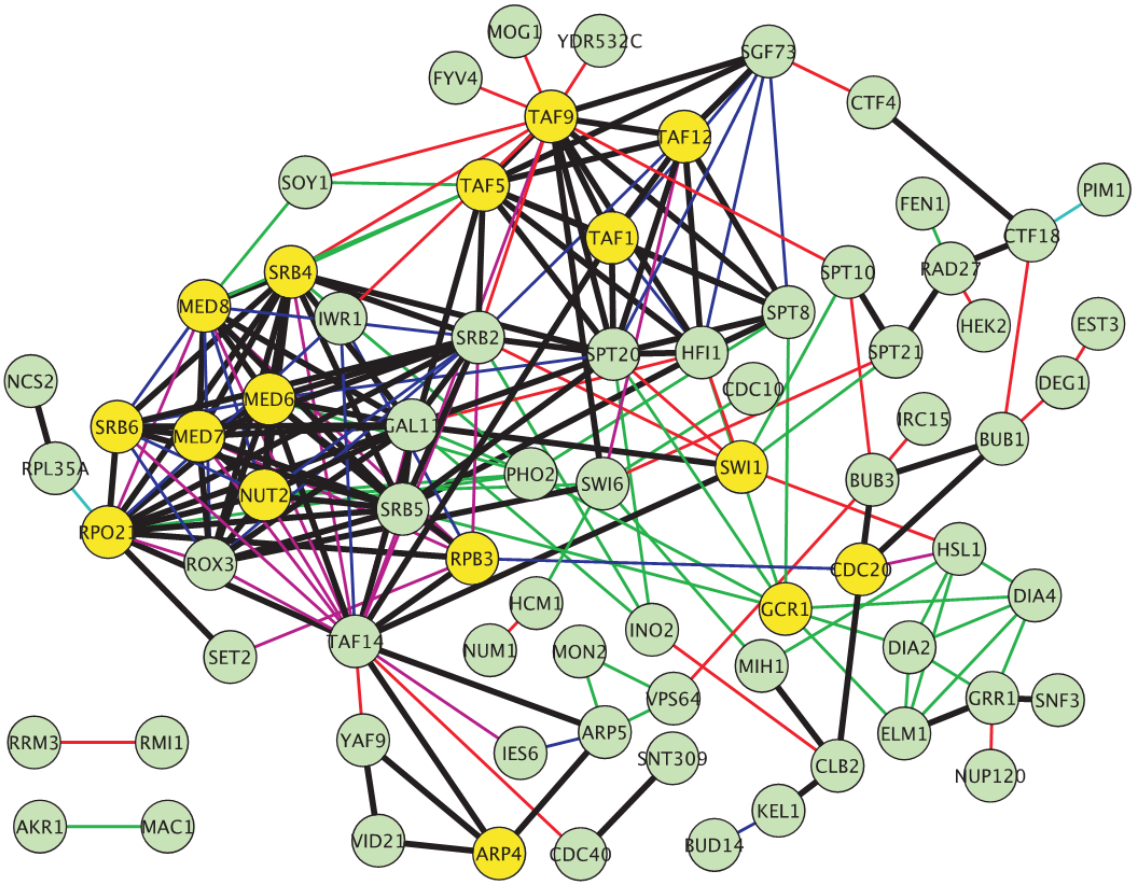


FIGURE 3.14 NETWORK CONNECTIVITY PREDICTS GENES INVOLVED IN CELL ELONGATION. 16 of 33 tested essential genes (yellow circles) showed elongated cell phenotypes (see Figure 3.13) on the basis of their connections to the seed set genes (green circles), with particular enrichment for genes associated with RNA polymerase II transcriptional initiation and the mediator complex. The color of the edge between two genes indicates the source of evidence supporting the functional link: thick black, multiple types of evidence; blue, affinity purification/mass spectrometry; green, literature mining by co-citation; cyan, gene neighbors or tertiary structure; pink, literature curated physical interaction; red, genetic interaction. Network drawn with Cytoscape [12]. Figure used by permission [1].

DISCUSSION

Just as functional networks propagate known functional annotations to un-annotated genes, phenotype prediction *via* GBA is limited to propagating known phenotypes. Therefore, an initial seed set of genes is required, such as might result from a genetic screen for the phenotype of interest, before being able to apply the network in order to identify more such genes. I might also expect genes in the same pathway to often exert inverse effects on a phenotype, acting either as activators or repressors. I will discuss this further in the next chapter, because my next area of research suggests that there may be ways to find "inverse" phenotypes when a number of activators and repressors are involved in a given pathway. Despite possible complicating factors, I demonstrate that GBA can be successfully applied to identify genes giving rise to similar loss-of-function phenotypes. Furthermore, network-guided phenotype prediction can be used to extend a genetic screen in a targeted fashion by providing a ranked list of potential candidates for evaluation. In principle, the screen might be expanded by adding the newly identified genes to the seed set and iterating the prediction and testing.

In particular, large-scale reverse genetic screens using yeast mutant strain collections have become increasingly common [71]. However, as seen in my chapter on shmoo localization, large-scale assays often suffer from high false negative rates. In many cases, the primary source of false negatives may be the limited scope of libraries used in screening (e.g., screening only the nonessential or essential genes). Such partially genome-wide screens can benefit by following up the initial screen with focused screening (or re-screening) of prioritized candidate genes. In order to facilitate such efforts, I have created a web server [69] which allows interactive analysis of a seed gene

set, performing ROC analysis to assess the predictability of the phenotype, then returning a ranked list of candidate genes most likely to share the same loss-of-function phenotype.

Note that I have focused on predicting loss-of-function phenotypes because of the large number of genome-wide screens available; it is not clear that gain-of-function phenotypes will be similarly predictable. However, the recent construction of yeast over-expression libraries [72-74] should soon allow testing of network-based prediction of such phenotypes.

Why are loss-of-function phenotypes predictable?

The results indicate that typical phenotypes represent specific enough defects that they are predictable based upon the genes' functional associations. I observe multiple mechanisms for how loss of different genes leads to disruption of the same phenotypically relevant process, primarily participation in the same protein complex or membership in the same biological pathway. These results are consistent with the partial predictability of human disease from protein complex membership [62, 63] and of the prediction of knockout phenotypes of annotated yeast genes on the basis of pathway annotation [64], which I illustrate with the following contrasting examples from among my predictions. In Figure 3.6A, the proteins ANP1, MNN9, MNN10, MNN11, VAN1 are members of the same alpha-1,6-mannosyltransferase protein complex. Chitin accumulates when the function of the complex is disrupted by the loss of any one of the five members [32]. In contrast, in Figure 3.6B, the three genes THR1, THR4, and HOM6 are involved in the biochemical pathway that converts homoserine to threonine; these genes are linked in the functional network [11] by virtue of the coordinate expression of their bacterial homologs in operons (e.g., as for the *Bacillus subtilis* homologs ThrB,

ThrC, and ThrA), even though there is as yet little evidence that they belong to the same physical complex. The loss of any of the three genes disrupts the threonine synthesis pathway and leads to reduced growth after 5 generations in threonine-depleted media [4]. The functional gene network, which combines both physical and functional interactions, predicts both classes of phenotypes effectively, whether resulting from disruption of physical complexes or pathways.

Nevertheless, some phenotypes are not significantly predictable. Three likely causes exist: First, poor predictability may result from using genome-wide screens with high false positive rates, which would base predictions on incorrectly identified seed sets. I sought to minimize this type of error by adopting stringent thresholds for each phenotype. Second, incomplete screens (e.g., such as by not testing the essential genes), high false negative rates, and the stringent phenotype thresholds that I selected could lead to a large number of positive examples being excluded from the seed sets. Such omitted positive examples scoring higher than seed genes would artificially depress prediction accuracies. Third, unpredictable phenotypes could in principle arise from the disruption of functionally unrelated genes. In order to test this, I compared the GO enrichment for the 25 most predictable phenotypes with the 25 least predictable phenotypes. For each phenotype, I identified the GO term with the most significant enrichment of genes annotated with the term, measured using the hypergeometric distribution. Using a significance threshold of $p < 10^{-7}$, I find that 18 of the 25 highly predictable phenotypes are significantly enriched for at least one GO annotation, versus only 2 of the 25 poorly predictable phenotypes. This suggests that poorly predictable phenotypes largely result from sets of genes with little functional coherence.

AUC is a useful measure of gene functional coherence

By definition, the GBA approach predicts phenotypes associated with functionally coherent sets of genes, presumably reflecting the clustering of the genes in the functional network. Such predictability, which I specifically measure as the AUC, can therefore be regarded as a direct estimate of the functional coherence of the seed gene set. Thus, beyond simply evaluating phenotype prediction, the AUC offers an additional measure of functional coherence that complements other existing measures, such as the enrichment of GO annotations or other biologically meaningful sets of genes (e.g., as calculated by FunSpec [67] or DAVID [75]). For example, the five genes giving rise to the branched cell phenotype are connected by six linkages in the network (AUC = 0.87), but only a single pair shares any GO annotation ($p < 0.001$, for the GO term "transcription from RNA polymerase II promoter"). The network-based AUC measure for functional coherence leverages the massive unbiased data integration of functional networks, extending well beyond known annotations, and allows estimates of functional coherence even among unannotated genes or those spanning multiple systems.

In principle, the AUC approach can therefore measure the functional coherence of genes that annotation-based methods will miss. Beyond unannotated genes, the AUC-based estimate of functional coherence might also work effectively when the genes under study span multiple functional categories—each category may only be partially enriched and therefore otherwise be missed for lack of signal. The functional network, however, considers pairwise linkages, not predetermined categories, so has the potential to identify linked genes across multiple annotation categories.

Recapitulation of the classic mutator phenotype in the yeast knockout collection

I observed a strikingly higher predictability for mutations that increased cell-to-cell phenotypic variation versus those that decreased it. The deletion strains exhibiting higher CVs tended to be consistent across the complete set of CV phenotypes examined, with the deleted genes showing strong enrichment for functions related to DNA repair, recombination, and genomic stability. Note that strains with the lowest CV phenotypes showed neither predictability nor functional enrichment—in fact, the CVs exhibited by these strains were similar to those observed for replicate analyses of wild-type cells (**Figure 3.12**), suggesting that the strains that most decreased cell-to-cell variation were essentially wild-type-like in this regard.

This outcome is consistent with a recapitulation in the yeast deletion strain collection of the classic mutator phenotype. The mutator phenotype was originally observed in DNA repair mutants—such mutants accumulated mutations so rapidly that they showed high variability in colony sizes when grown on Petri dishes, high variability in cell morphologies, high rates of plasmid loss, and increased spontaneous mutagenesis (e.g., as previously observed for RAD27 and RAD52 deletion mutants [76, 77]). The most likely explanation is therefore that strains in the deletion collection harboring deletions in genes related to genomic stability have simply accumulated mutations at a higher rate. A mixed population, no longer clonal, would be expected to exhibit more cell-to-cell variation than other deletion strains, which would accumulate mutations at a lower rate. Thus, I suspect that the phenotypic analysis is correctly revealing the functional signature of a legitimate phenotype inadvertently captured in the process of distributing and passaging the yeast deletion strain collection.

Applying network-based phenotype prediction to humans and other organisms

In principal, the approach I describe could be applied for any organism, using functional network data, if available, or in the absence of such data, using physical interaction data, such as available protein interaction networks for fly [78], worm [58], or human [25, 79-83]. In the absence of an integrated functional gene network or protein interaction network, I expect that networks of mRNA co-expression associations, such as can be derived from DNA microarray data, would provide some utility for phenotype prediction. Such data are a major contributor to functional gene networks (e.g., [9, 84, 85]) and are relatively easily generated from available data for most model organisms.

In particular, application of this approach in humans may allow directed identification of disease genes. Indeed, functional linkages derived largely from known Gene Ontology annotation [86] or protein interactions [62] have shown some utility for prioritizing positional candidate genes from genome-wide linkage screens. However, the results show that across a wide range of yeast phenotypes and human diseases the associated genes (or their yeast orthologs) can be directly identified even in the absence of supporting genetic loci data. In order to apply my approach to human diseases, genes known to be associated with a particular disease, such as found from twin or genome-wide association studies, would form the seed set. Additional candidate genes likely to be associated with that disease could then potentially be identified or prioritized based upon their network connections to the seed set, using the guilt-by-association principle. Potential disease genes could then be tested in disease model systems or screened genetically in a focused manner. Such a directed approach would leverage the tremendous existing body of knowledge about protein interactions and functional pathways.

CONCLUSIONS

In summary, I have demonstrated that yeast gene loss-of-function phenotypes are broadly predictable from connectivity in a functional gene network, with examples presented spanning a wide range of cell growth, cell morphology, metabolite transport, chemical sensitivity, and molecular phenotypes. I demonstrate that this predictability can be used to extend genetic screens in a directed fashion, and that this approach might therefore be important in organisms for which genetics is difficult. Furthermore, based on a computational analysis of gene linkage among yeast-human orthologs involved in disease, I suggest that a similar approach in humans might enable the directed discovery of disease genes. In the following chapter, I will discuss another strategy for predicting phenotype across species that was inspired by this observation and others.

This work is published in Genome Biology [1] from which this chapter is reworked and expanded upon.

REFERENCES

1. McGary KL, Lee I, Marcotte EM: Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes. *Genome Biol* 2007, 8:R258.

2. Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005, 6:95-108.

3. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E *et al*: Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 2007, 448:470-473.

4. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B *et al*: Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 2002, 418:387-391.

5. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al*: Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* 1999, 285:901-906.

6. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M *et al*: Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* 2003, 421:231-237.

7. Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J: Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature* 2000, 408:325-330.

8. Downward J: Use of RNA interference libraries to investigate oncogenic signalling in mammalian cells. *Oncogene* 2004, 23:8376-8383.

9. Lee I, Date SV, Adai AT, Marcotte EM: A probabilistic functional network of yeast genes. *Science* 2004, 306:1555-1558.

10. Lee I, Li Z, Marcotte EM: An Improved, Bias-Reduced Probabilistic Functional Gene Network of the Baker's Yeast, Sacchromyces cerevisiae. *PLOS One* 2007.

11. Lee I, Li Z, Marcotte EM: An Improved, Bias-Reduced Probabilistic Functional Gene Network of Baker's Yeast, Saccharomyces cerevisiae. *PLoS ONE* 2007, 2:e988.

12. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003, 13:2498-2504.

13. Mnaimneh S, Davierwala AP, Haynes J, Moffat J, Peng WT, Zhang W, Yang X, Pootoolal J, Chua G, Lopez A *et al*: Exploration of essential gene functions via titratable promoter alleles. *Cell* 2004, 118:31-44.

14. Ni L, Snyder M: A genomic study of the bipolar bud site selection pattern in Saccharomyces cerevisiae. *Mol Biol Cell* 2001, 12:2147-2170.

15. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M *et al*: SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 1998, 26:73-79.

16. Saito TL, Ohtani M, Sawai H, Sano F, Saka A, Watanabe D, Yukawa M, Ohya Y, Morishita S: SCMD: Saccharomyces cerevisiae Morphological Database. *Nucleic Acids Res* 2004, 32:D319-322.

17. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V: MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res* 2006, 34:D436-441.

18. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 2002, 30:303-305.

19. Hart GT, Lee I, Marcotte EM: A high-accuracy map of yeast protein complexes reveals modular basis of gene essentiality. *BMC Bioinformatics* 2007, 8:236.

20. Deane CM, Salwinski L, Xenarios I, Eisenberg D: Protein interactions: Two methods for assessment of the reliability of high-throughput observations. *Mol Cell Proteomics* 2002:M100037-MCP100200.

21. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ: Towards a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* 2007.

22. Hu Z, Killion PJ, Iyer VR: Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet* 2007, 39:683-687.

23. Mnaimneh S, Davierwala AP, Haynes J, Moffat J, Peng WT, Zhang W, Yang XQ, Pootoolal J, Chua G, Lopez A *et al*: Exploration of essential gene functions via titratable promoter alleles. *Cell* 2004, 118:31-44.

24. Remm M, Storm CE, Sonnhammer EL: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001, 314:1041-1052.

25. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B *et al*: Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* 2006, 38:285-293.

26. Hart GT, Ramani AK, Marcotte EM: How complete are current yeast and human protein-interaction networks? *Genome Biol* 2006, 7:120.

27. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: Lethality and centrality in protein networks. *Nature* 2001, 411:41-42.

28. Dezso Z, Oltvai ZN, Barabasi AL: Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae. *Genome Res* 2003, 13:2450-2454.

29. Willer M, Regnacq M, Reid PJ, Tyson JR, Cui W, Wilkinson BM, Stirling CJ: Disruption and functional analysis of six ORFs on chromosome XII of saccharomyces cerevisiae: YLR124w, YLR125w, YLR126c, YLR127c, YLR128w and YLR129w. *Yeast* 2000, 16:1429-1435.

30. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G *et al*: Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res* 2002, 30:69-72.

31. Markovich S, Yekutiel A, Shalit I, Shadkchan Y, Osherov N: Genomic approach to identification of mutations affecting caspofungin susceptibility in Saccharomyces cerevisiae. *Antimicrob Agents Chemother* 2004, 48:3871-3876.

32. Lesage G, Shapiro J, Specht CA, Sdicu AM, Menard P, Hussein S, Tong AH, Boone C, Bussey H: An interactional network of genes involved in chitin synthesis in Saccharomyces cerevisiae. *BMC Genet* 2005, 6:8.

33. Bonangelino CJ, Chavez EM, Bonifacino JS: Genomic screen for vacuolar protein sorting genes in Saccharomyces cerevisiae. *Mol Biol Cell* 2002, 13:2486-2501.

34. Narayanaswamy R, Niu W, Scouras AD, Hart GT, Davies J, Ellington AD, Iyer VR, Marcotte EM: Systematic profiling of cellular phenotypes with spotted cell microarrays reveals mating-pheromone response genes. *Genome Biol* 2006, 7:R6.

35. Birrell GW, Giaever G, Chu AM, Davis RW, Brown JM: A genome-wide screen in Saccharomyces cerevisiae for genes affecting UV radiation sensitivity. *Proc Natl Acad Sci U S A* 2001, 98:12608-12613.

36. Huang ME, Rio AG, Nicolas A, Kolodner RD: A genomewide screen in Saccharomyces cerevisiae for genes that suppress the accumulation of mutations. *Proc Natl Acad Sci U S A* 2003, 100:11529-11534.

37. Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C, Giaever G: Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* 2005, 169:1915-1925.

38. Aouida M, Page N, Leduc A, Peter M, Ramotar D: A genome-wide screen in Saccharomyces cerevisiae reveals altered transport as a mechanism of resistance to the anticancer drug bleomycin. *Cancer Res* 2004, 64:1102-1109.

39. Askree SH, Yehuda T, Smolikov S, Gurevich R, Hawk J, Coker C, Krauskopf A, Kupiec M, McEachern MJ: A genome-wide screen for Saccharomyces cerevisiae deletion mutants that affect telomere length. *Proc Natl Acad Sci U S A* 2004, 101:8658-8663.

40. Chang M, Bellaoui M, Boone C, Brown GW: A genome-wide screen for methyl methanesulfonate-sensitive mutants reveals genes required for S phase progression in the presence of DNA damage. *Proc Natl Acad Sci U S A* 2002, 99:16934-16939.

41. Zhang J, Schneider C, Ottmers L, Rodriguez R, Day A, Markwardt J, Schneider BL: Genomic scale mutant hunt identifies cell size homeostasis genes in S. cerevisiae. *Curr Biol* 2002, 12:1992-2001.

42. Page N, Gerard-Vincent M, Menard P, Beaulieu M, Azuma M, Dijkgraaf GJ, Li H, Marcoux J, Nguyen T, Dowse T *et al*: A Saccharomyces cerevisiae genome-wide mutant screen for altered sensitivity to K1 killer toxin. *Genetics* 2003, 163:875-894.

43. Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ *et al*: Systematic screen for human disease genes in yeast. *Nat Genet* 2002, 31:400-404.

44. Griffith JL, Coleman LE, Raymond AS, Goodson SG, Pittard WS, Tsui C, Devine SE: Functional genomics reveals relationships between the retrovirus-like Ty1 element and its host Saccharomyces cerevisiae. *Genetics* 2003, 164:867-879.

45. Bennett CB, Lewis LK, Karthikeyan G, Lobachev KS, Jin YH, Sterling JF, Snipe JR, Resnick MA: Genes required for ionizing radiation resistance in yeast. *Nat Genet* 2001, 29:426-434.

46. Jorgensen P, Nishikawa JL, Breitkreutz BJ, Tyers M: Systematic identification of pathways that couple cell growth and division in yeast. *Science* 2002, 297:395-400.

47. Lesuisse E, Knight SA, Courel M, Santos R, Camadro JM, Dancis A: Genome-wide screen for genes with effects on distinct iron uptake activities in Saccharomyces cerevisiae. *Genetics* 2005, 169:107-122.

48. Blackburn AS, Avery SV: Genome-wide screening of Saccharomyces cerevisiae to identify genes required for antibiotic insusceptibility of eukaryotes. *Antimicrob Agents Chemother* 2003, 47:676-681.

49. Fleming JA, Lightcap ES, Sadis S, Thoroddsen V, Bulawa CE, Blackman RK: Complementary whole-genome technologies reveal the cellular response to proteasome inhibition by PS-341. *Proc Natl Acad Sci U S A* 2002, 99:1461-1466.

50. Desmoucelles C, Pinson B, Saint-Marc C, Daignan-Fornier B: Screening the yeast "disruptome" for mutants affecting resistance to the immunosuppressive drug, mycophenolic acid. *J Biol Chem* 2002, 277:27036-27044.

51. Deutschbauer AM, Williams RM, Chu AM, Davis RW: Parallel phenotypic analysis of sporulation and postgermination growth in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 2002, 99:15530-15535.

52. Enyenihi AH, Saunders WS: Large-scale functional genomic analysis of sporulation and meiosis in Saccharomyces cerevisiae. *Genetics* 2003, 163:47-54.

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

54. Chan TF, Carvalho J, Riles L, Zheng XF: A chemical genomics approach toward understanding the global functions of the target of rapamycin protein (TOR). *Proc Natl Acad Sci U S A* 2000, 97:13227-13232.

55. Warringer J, Ericson E, Fernandez L, Nerman O, Blomberg A: High-resolution yeast phenomics resolves different physiological features in the saline response. *Proc Natl Acad Sci U S A* 2003, 100:15724-15729.

56. Wilson WA, Wang Z, Roach PJ: Systematic identification of the genes affecting glycogen storage in the yeast Saccharomyces cerevisiae: implication of the vacuole as a determinant of glycogen level. *Mol Cell Proteomics* 2002, 1:232-242.

57. Zewail A, Xie MW, Xing Y, Lin L, Zhang PF, Zou W, Saxe JP, Huang J: Novel functions of the phosphatidylinositol metabolic pathway discovered by a chemical genomics screen with wortmannin. *Proc Natl Acad Sci U S A* 2003, 100:3345-3350.

58. Riles L, Shaw RJ, Johnston M, Reines D: Large-scale screening of yeast mutants for sensitivity to the IMP dehydrogenase inhibitor 6-azauracil. *Yeast* 2004, 21:241-248.

59. Huang RY, Eddy M, Vujcic M, Kowalski D: Genome-wide screen identifies genes whose inactivation confer resistance to cisplatin in Saccharomyces cerevisiae. *Cancer Res* 2005, 65:5890-5897.

60. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF *et al*: Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 2005, 123:507-519.

61. Kelley R, Ideker T: Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 2005, 23:561-566.

62. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N *et al*: A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007, 25:309-316.

63. Oti M, Snel B, Huynen MA, Brunner HG: Predicting disease genes using protein-protein interactions. *J Med Genet* 2006, 43:691-698.

64. King OD, Lee JC, Dudley AM, Janse DM, Church GM, Roth FP: Predicting phenotype from patterns of annotation. *Bioinformatics* 2003, 19 Suppl 1:i183-189.

65. Lee I: Personal Communication. 2008.

66. Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, Saito TL, Saka A, Fukuda T, Ishihara S, Oka S *et al*: High-dimensional and large-scale phenotyping of yeast mutants. *Proc Natl Acad Sci U S A* 2005, 102:19015-19020.

67. Robinson MD, Grigull J, Mohammad N, Hughes TR: FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics* 2002, 3:35.

68. Online Mendelian Inheritance in Man, OMIM (TM) [http://www.ncbi.nlm.nih.gov/omim]

69. Online Mendelian Inheritance in Man, OMIM (TM) MIM Number:212065. *Johns Hopkins University, Baltimore, MD WWW: http://wwwncbinlmnihgov/omim/.*

70. Mewes HW, Frishman D, Mayer KF, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V: MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res* 2006, 34:D169-172.

71. Scherens B, Goffeau A: The uses of genome-wide yeast mutant collections. *Genome Biol* 2004, 5:229.

72. Kim H, Melen K, Osterberg M, von Heijne G: A global topology map of the Saccharomyces cerevisiae membrane proteome. *Proc Natl Acad Sci U S A* 2006, 103:11142-11147.

73. Sopko R, Huang D, Preston N, Chua G, Papp B, Kafadar K, Snyder M, Oliver SG, Cyert M, Hughes TR *et al*: Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* 2006, 21:319-330.

74. Gelperin DM, White MA, Wilkinson ML, Kon Y, Kung LA, Wise KJ, Lopez-Hoyo N, Jiang L, Piccirillo S, Yu H *et al*: Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. *Genes Dev* 2005, 19:2816-2826.

75. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003, 4:P3.

76. Reagan MS, Pittenger C, Siede W, Friedberg EC: Characterization of a mutant strain of Saccharomyces cerevisiae with a deletion of the RAD27 gene, a structural homolog of the RAD2 nucleotide excision repair gene. *J Bacteriol* 1995, 177:364-371.

77. Hastings PJ, Quah SK, von Borstel RC: Spontaneous mutation by mutagenic repair of spontaneous lesions in DNA. *Nature* 1976, 264:719-722.

78. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al*: A protein interaction map of Drosophila melanogaster. *Science* 2003, 302:1727-1736.

79. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N *et al*: Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005, 437:1173-1178.

80. Lehner B, Fraser AG: A first-draft human protein-interaction map. *Genome Biol* 2004, 5:R63.

81. Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM: Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol* 2005, 6:R40.

82. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S *et al*: A human protein-protein

interaction network: a resource for annotating the proteome. *Cell* 2005, 122:957-968.

83. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L *et al*: Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005, 33:D428-432.

84. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol* 2005, 23:951-959.

85. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D: A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc Natl Acad Sci U S A* 2003, 100:8348-8353.

86. Franke L, Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006, 78:1011-1025.

Chapter 4: Predicting and testing human disease genes in model organisms by finding equivalent phenotypes between species.

INTRODUCTION

Identifying genes responsible for human disease is often challenging, at least in part because phenotypic assays are frequently not possible. In addition, genome-wide, blind genetic studies for mutation/disease correlation often require much more data than can be easily collected and often run the risk of not passing a significance threshold due to corrections for multiple testing. Therefore, it is desirable to find high confidence candidate genes that are strongly evidenced to play a causative role in the disease of interest, prior to screening.

My previous two projects demonstrated that the yeast functional network effectively predicts genes for targeted screens of both intracellular protein localization and organismal phenotype. Furthermore, I found that the network could predict yeast orthologs of human diseases. The logical extension of this research program is to show that the functional network can provide a rational, quantitative approach for prioritizing candidate disease genes across species and to evaluate which organisms are appropriate models for a given disease. As illustrated in the previous chapter by the yeast cell elongation phenotype and the human disease leukoencephaly, once a phenotype or disease is traced to the disruption of a gene, the phenotypic outcome of the disruption of related genes can be predicted even without understanding the exact mechanism that gives rise to the phenotype.

However, it is not always possible to predict the exact phenotypic outcome of disrupting an ortholog present in one species from the gene disruption phenotype of another species. For example, 3 of 4 genes that are implicated in leukoencephaly in humans have yeast orthologs that are essential and the other is sensitive to growth in synthetic complete minus tryptophan media. Even more starkly, mutating the *RB1* gene in humans gives rise to retinoblastoma [1], a cancer of the retina, yet disrupting the RB1 ortholog (and a second redundant gene) in the nematode *C. elegans* gives rise to ectopic vulvae [2]. Mutant phenotypes are thus an emergent property of the system; disruptions of equivalent genes with conserved molecular functions, but in different systems contexts, can lead to different outcomes. It becomes clear that to test the prediction of human disease genes in model organisms it is necessary to start by finding phenotypes that are in some way equivalent across species. However, diverse genetic perturbations can give rise to the same phenotypic outcome (degeneracy), while mutation of a single gene can lead to multiple phenotypic outcomes (pleiotropy). Genes and phenotypes thus have a many-to-many relationship, and mapping equivalent phenotypes between organisms is non-obvious. Nonetheless, once equivalent phenotypes are identified, screens for additional mutations that lead to a phenotype in the model organism may provide high confidence candidate genes for a phenotype/disease in the reference organism.

But, can we tell if a particular disease model is equivalent enough to the human case to be useful? Can this property be quantified, allowing ranking of models according to utility? Are there non-obvious models for human disease, perhaps hidden by differences in emergent appearance? Importantly, a model does not have to exactly reproduce symptoms of a disease to be useful. Thousands of genome-wide mutational

analyses have now been performed, associating genes to phenotypes in model organisms, e.g., yeast, worms, and mice, at a far higher rate than for humans (**Figure 4.1**).

Identifying and expanding models of human disease

I demonstrate a three step process for identifying cross-species models of human diseases. First, I identify equivalent phenotypes based on the shared involvement of orthologous genes involved in phenotypes from two species. Second, I test genes in the model organism for the disease equivalent phenotype and choose genes that are naturally predicted by the overlapping phenotypes or by being closely linked network neighbors as suggested in my previous chapter. Third, I test the candidate genes for their phenotypic outcome in the original organism, or appropriate surrogate.

Identifying equivalent phenotypes (phenologs)

As a framework for considering equivalent phenotypes, my research introduces the notion of *orthologous phenotypes*, defined as phenotypes related by the orthology of the associated genes in two organisms, and corresponding to the phenotype-level equivalent to gene orthologs. Orthologous phenotypes derive from sets of genes in two organisms such that the genes in each organism are associated with the same phenotype (phenotypes can differ between the organisms), and the associated gene sets overlap significantly (*i.e.*, are enriched for the same orthologous genes) (**Figure 4.2**). Orthologous phenotypes are evolutionarily conserved outputs of conserved systems of genes, which can manifest as different traits or structures in different organisms due to organism-specific context effects. The human retinoblastoma eye cancer and the *C. elegans* synthetic multivulval phenotype are orthologous, with failures of orthologous

genes performing equal molecular functions in different contexts causing different phenotypic outcomes. Orthologous phenotypes thus bridge the molecular definitions of homologous and orthologous genes [3] with classic definitions of homologous structures from Owen [4] and Darwin [5], deriving from considerations both of gene heredity and of the traits/structures affected by perturbing the genes. I will refer to orthologous phenotypes as *phenologs*.

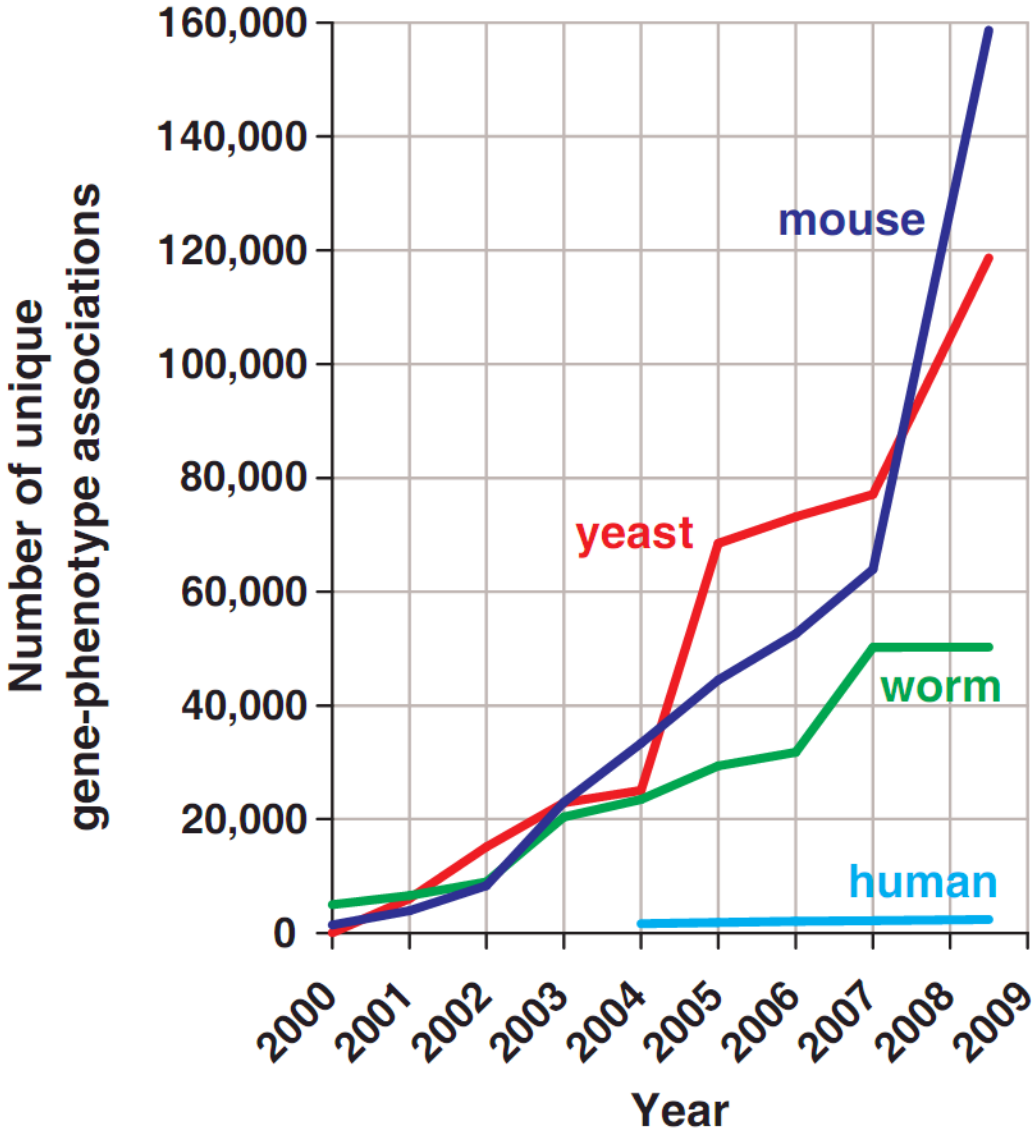


FIGURE 4.1 THE RATE OF ASSOCIATING GENES TO ORGANISM-LEVEL PHENOTYPES IN MODEL ORGANISMS GREATLY EXCEEDS THAT IN HUMANS (data from [6-10]). Thus, appropriate mapping of model organism phenotypes to human diseases could significantly accelerate discovery of human disease gene associations. Orthologous phenotypes (phenologs) offer one such approach. Figure adapted from work in review [42]. Thanks to Greg Weiss for figure.

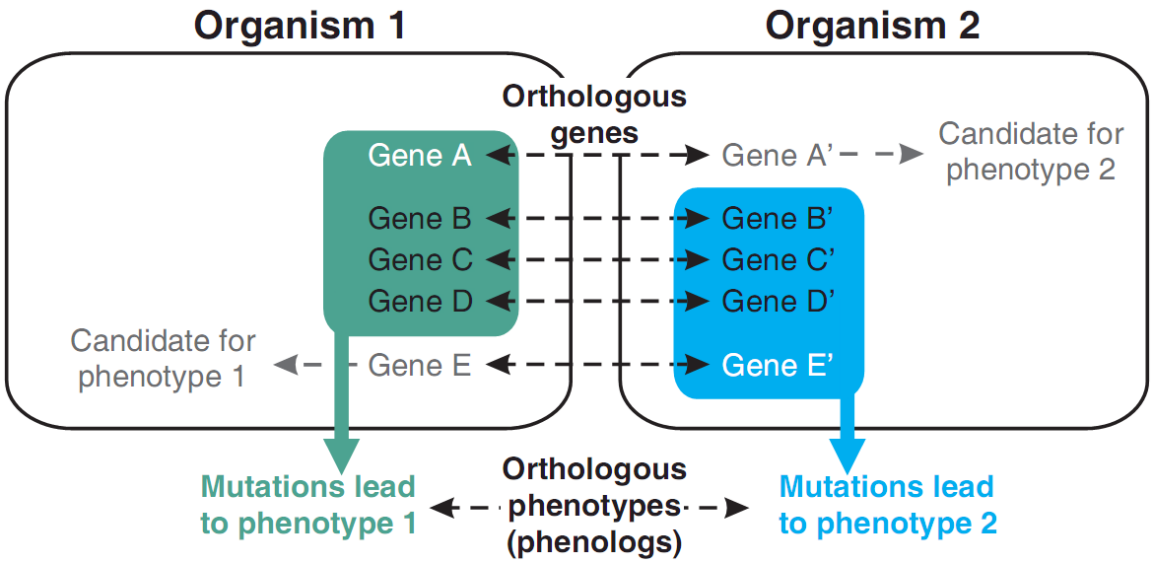


FIGURE 4.2 PHENOLOGS CAN BE IDENTIFIED BASED ON SIGNIFICANTLY OVERLAPPING SETS OF ORTHOLOGOUS GENES (A is orthologous to A', B to B', etc), such that each gene in a given set (green box or cyan box) gives rise to the same phenotype in that organism. The phenotypes may differ in appearance between organisms due to differing organismal contexts. As gene-phenotype associations are often incompletely mapped, genes currently linked to only one of the orthologous phenotypes become candidate genes for the other phenotype (e.g., the ortholog of gene D in organism 2 is a new candidate for phenotype 2). Figure adapted from work in review [42].

Phenologs can be identified by assembling known gene-phenotype associations for two organisms, considering only genes that are orthologous between the two organisms, then testing each inter-organism phenotype pair for significant gene overlap based upon three observations: (1) the total number of orthologs in organism 1 that give rise to phenotype 1; (2) the total number of orthologs in organism 2 that give rise to phenotype 2; and (3) the number of orthologs shared between these two sets. The significance of a phenolog can be calculated from the hypergeometric probability of observing at least that many shared orthologs by chance. **Figure 4.3** shows an example: the set of human genes (with worm orthologs) associated with X-linked breast/ovarian cancer significantly overlaps genes whose mutations lead to a high frequency of male progeny in *C. elegans.* Male *C. elegans* are determined by a single X chromosome, hermaphrodites by 2 copies; thus, X chromosome non-disjunction leads to higher frequencies of males [11]. Human breast/ovarian cancers can derive from a similar mechanism, e.g. as for sporadic basal-like breast cancers [12], supporting the notion that this phenolog is identifying a useful disease model.

Finding genes in model organism for orthologous phenotype

The breast cancer/male progeny example above demonstrates that candidate disease genes in the reference organism can be identified immediately if they are already known to result in the orthologous phenotype upon disruption. Human orthologs of the 13 additional genes associated with the worm trait are reasonable candidate genes for involvement in breast/ovarian cancers. Nine of these genes were not yet included in the databases I employed, but could be confirmed in the primary literature to be linked to breast cancer (e.g., as for the breast cancer biomarker KIF15 [13]); 4 genes (GCC2, PIGA, WDHD1, SEH1L) remain as breast cancer candidate genes **(Figure 4.3)**. The

worm phenotype thus predicts and suggests additional genes relevant to human breast cancer. Furthermore, additional candidates can be identified by targeted screening using guilt by association in the functional network of the target organism, using currently known genes as the seed set, where known genes can either those involved in the overlap (identified by k in **Figure 4.2**), or the entire set of genes involved in the model organism phenotype (k and n_2 in **Figure 4.2**). Later in this chapter, I will discuss the positive results of the attempt to experimentally validate this approach.

Testing candidate genes

Once candidate genes are identified they can be evaluated for their involvement in the disease phenotype in a number of ways depending on the specific phenotype. For some phenotypes, candidate genes can be adequately tested by knockout or knockdown in cell or tissue culture. In other cases, the genomic region surrounding candidate genes can be targeted for sequencing to identify mutations in affected individuals. In this work, we utilize a third approach, where candidate genes from simple organisms are tested in a vertebrate model system that is already known to mimic human disease quite well.

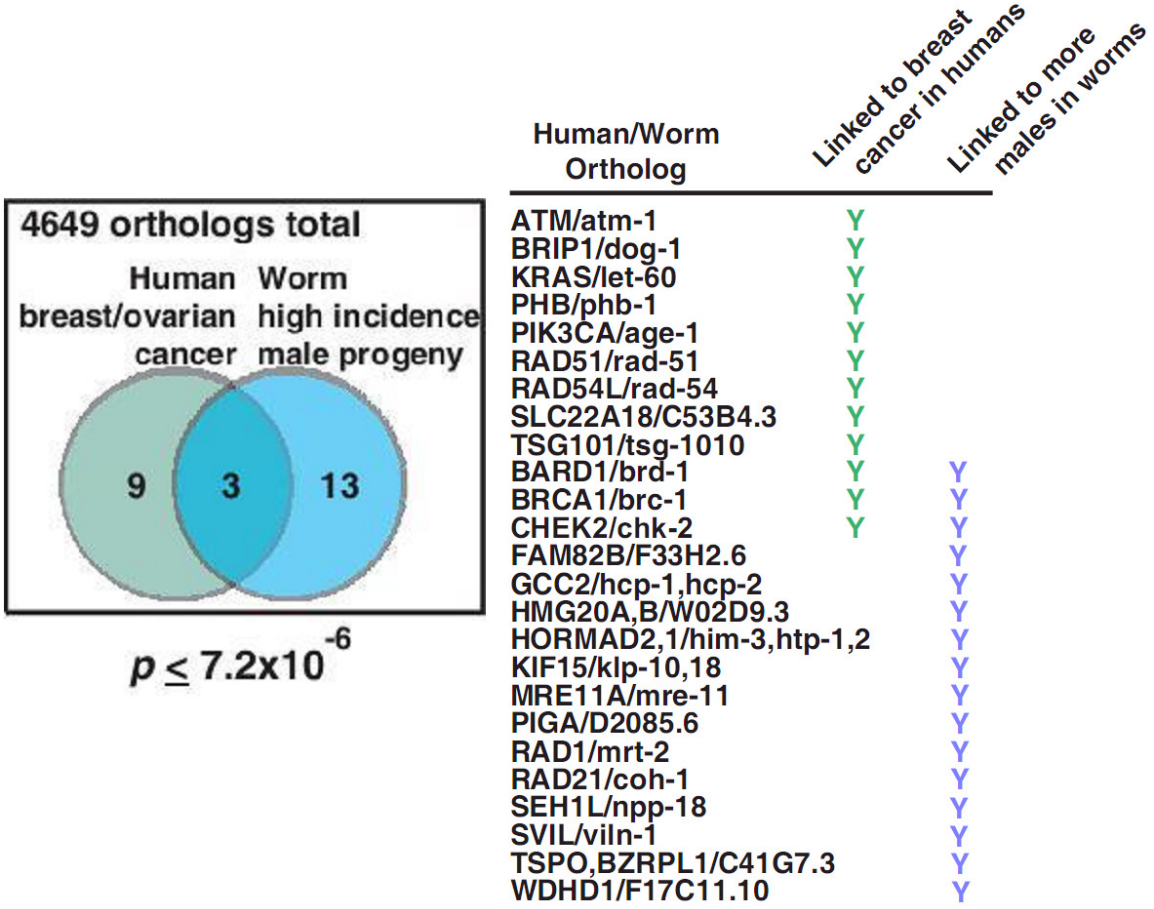


FIGURE 4.3 AN EXAMPLE OF A PHENOLOG MAPPING HIGH INCIDENCE OF MALE C. ELEGANS PROGENY TO HUMAN BREAST/OVARIAN CANCERS (details in text). Figure adapted from work in review [42].

METHODS

Collection of phenotypes

I collected gene-phenotype associations from the literature for four species (worm, yeast, mouse, and human). For human phenotypes, I used employed human diseases from the OMIM database [14], using the compressed OMIM disease categories previously described in McGary *et al*. [15], such that multiple variants of a disease were grouped together. (For example "Corneal dystrophy, hereditary polymorphous posterior" and "Corneal dystrophy, lattice type I," reduce to a single category of corneal dystrophies). Mouse gene-phenotype associations were downloaded from MGI [16] (MGI_PhenoGenoMP.rpt; downloaded on April 21, 2008). Gene-phenotype associations involving more than one locus or that could not be linked to an Entrez Gene were removed. MGI identifiers were converted to Entrez GeneIDs using MGI_Coordinate.rpt (downloaded April 25, 2008). MGI mouse phenotype descriptions were from VOC_MammalianPhenotype.rpt, downloaded May 7, 2008. All MGI data were downloaded from ftp://ftp.informatics.jax.org/pub/reports/index.html. The MGI associations were supplemented with a small number of broadly defined mouse phenotypes obtained from http://hugheslab.med.utoronto.ca/supplementary-data/mouseFunc_I/MGI_phenotype.txt, but which are ultimately derived from MGI data. Worm gene-phenotype associations were assembled from the literature-reported RNAi studies assembled in Lee *et al.* [17] supplemented by the addition of phenotype data downloaded from WormBase 188 [7] (ftp://ftp.wormbase.org/pub/wormbase/acedb/WS188/). Worm gene-phenotype association data come from phenotype_association.WS188.wb, phenotype descriptions from phenotype_ontology.WS188.obo, and gene information from geneIDs.WS188.gz. Files

were downloaded on March 26, 2008. WormBase phenotypes were filtered for positive associations only. All allelic variants and RNAi data were reduced to gene-phenotype pairs. Gene IDs (e.g. WBGene00044645) were translated to sequence names (e.g. Y51H7BR.8) using geneIDs.WS188.gz. Of approximately 22K gene-phenotype pairs, 384 could not be linked to a sequence name. These derived primarily from uncloned genes and were thus omitted from further analysis. Yeast gene-phenotype associations were obtained from McGary *et al*. [15] (a literature compilation plus SGD [8]), supplemented with associations from a recent set of genome-wide screens of drug sensitivity [18] (homozygous and heterozygous screens, het.z_tdist_pval_nm.goodbatch.pub hom.z_tdist_pval_nm.pub downloaded from http://chemogenomics.stanford.edu/supplements/global/download/data/). All gene-phenotype associations from the drug screens were filtered using the authors' recommended cutoff of $p<1x10^{-5}$. For the purposes of calculating phenologs, I considered only a subset of the gene-phenotype associations plotted in Figure 1A, analyzing only those implicating single genes (*i.e.*, not genetic interactions or traits requiring simultaneous mutation of multiple loci), and only those phenotypes in which a defect was observed (*i.e.*, omitting genes associated with the phenotype "normal", "wild-type", "no effect", or other such cases.)

Identification of non-redundant phenotype sets

In order to minimize the number of redundant comparisons performed, all phenotype-associated gene sets within a single organism were tested for significant overlap and non-redundant sets were selected for subsequent analyses. Within each organism, phenotypes were identified that reciprocally covered ≥80% of each other's genes; for each such pair of phenotypes, only the phenotype with the greater number of

genes was retained. (For example, in mouse, genes associated with defects in the small petrosal ganglion and small nodose ganglion overlap considerably. The former has 9 associated genes, of which a subset of 8 is also associated with the latter phenotype; only the former was retained.)

Calculating Orthologs

Orthologs between species were calculated using the following translated genomes: Human, ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/protein/protein.fa.gz, downloaded on Feb. 7, 2008; Mouse, ftp://ftp.ncbi.nih.gov/genomes/M_musculus/protein/protein.fa.gz, downloaded Oct. 13, 2007; Worm, ftp://ftp.wormbase.org/pub/wormbase/data_freezes/WS170/sequences/wormpep170.tar.gz, downloaded on Feb. 19, 2007; Yeast, ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/orf_protein/ orf_trans.fasta.gz, downloaded Feb. 19, 2007.

For human and mouse proteomes, I analyzed only sequences with protein refseq identifiers (NP_ only). For humans, 43 genes without Gene IDs were removed (mostly hypothetical proteins). For mouse, three proteins without current records were removed.

In order to identify orthologous genes in different species, orthologs were calculated using INPARANOID v. 1.35 [19] with default parameters, using blastall 2.2.15 also with default parameters. All genes assigned as orthologs (strictly speaking, ortholog groups or orthogroups, due to inclusion of in-paralogs) by INPARANOID were kept, regardless of their INPARANOID score. Using orthogroups, rather than bidirectional best hits, captures the many-to-many relationships that exist for gene duplicates that exist in more than one copy in one or both species. In order to prevent

isoform variations from resulting in skewed blast results, mouse and human sequences with the same Entrez GeneID but separate RefSeqIDs were treated separately in INPARANOID. Following INPARANOID analysis, orthologs sharing GeneIDs were combined so that gene variants would be considered together in subsequent analyses.

Calculation of phenologs

For each pair of species, I first converted gene-phenotype associations to ortholog-phenotype associations using the orthologs calculated by INPARANOID. In cases were paralogous genes within an organism result in the same phenotype, multiple gene-phenotype associations thus collapse to a single ortholog-phenotype association, which eliminates artificial inflation of the significance of ortholog overlap. Second, I compared the set of orthologs associated with a given phenotype within one species (species 1) to the set of orthologs associated with a given phenotype in the second species (species 2), repeating this analysis for all pairwise comparisons of phenotypes from species 1 and species 2. For each pair of phenotypes in which the ortholog sets overlapped (shared members), I calculated the probability of the overlap due to chance using the cumulative hypergeometric distribution, where N is the total number of orthologs shared between the two species; n and m are the number of orthologs linked to the species 1 and species 2 phenotypes, respectively; and k is the number of common orthologs, *i.e.*, those linked to both phenotypes:

$$p(\#\,\text{shared orthologs} \geq k \mid m,n,N) = \sum_{i=k}^{\min(m,n)} p(i \mid m,n,N)$$

where

$$p(i \mid m,n,N) = \frac{\binom{n}{i}\binom{N-n}{m-i}}{\binom{N}{m}}$$

The hypergeometric probability does not correct for multiple comparisons, so I estimated the false discovery rate with an empirical permutation test. I performed 1,000 random permutations of the ortholog-phenotype associations, for each permutation repeating the all versus all phenotype comparison using ortholog set sizes identical to those associated with the actual phenotypes. Significant phenologs were identified at a false discovery rate of 0.05 by ranking real & permuted phenologs on the basis of the associated hypergeometric probabilities and selecting a threshold of probability where the proportion of permuted phenologs above the cutoff accounted for 5% of the phenologs.

Tests of sub-network modularity

I measured the degree of network interconnectivity among orthologs involved in overlapping phenotypes from yeast and worms using a modification to a recently developed measure of the network clustering of a set of genes [15, 17]. Given a query set of genes, their interconnectivity in a functional gene network (a gene network with edge weights corresponding to the log likelihood of the linked genes functioning in the same biological process [17, 20]) is calculated as the area under a receiver-operator characteristic curve (AUC) for predicting back members of the query gene set when rank-ordering all genes in the network by each gene's sum of edge weights to the query gene set (corresponding to the *naïve* Bayes probability of participating in the same process as genes in the query set), performing the test using cross-validation (each query gene is omitted in turn from the query set for purposes of its evaluation). AUC ranges from 0 to

1. A high AUC indicates that query genes are more tightly connected in the network to each other than to other genes, while an intermediate AUC (near 0.5) corresponds to no better than random recovery of query genes, indicating negligible interconnectivity of the query gene set in the network. (AUC values in the range of 0 to near 0.5 indicate worse than random expectation, e.g., systematically lower connectivity of the query set).

To analyze phenolog gene sets, I modified the method by converting the gene-centric functional networks [17, 20] into networks of orthologs based upon INPARANOID ortholog assignments. I retained only network edges connecting orthologs present in both yeast and worm. In the case that multiple genes are assigned to a single ortholog, multiple network edges could exist between a pair of orthologs; I retained only the edge with the greatest weight (confidence). The resulting yeast and worm networks thus each contain ortholog-ortholog functional associations, rather than gene-gene associations. Using these two networks, I calculated AUC as in [15, 17]: for a given ortholog query set (e.g. the set of orthologs in the intersection of a phenolog), I rank ordered all orthologs shared between yeast and worm by the sum of the edges connecting them to the query set, then calculated AUC for recovery of the query ortholog set using cross-validation.

I calculated network AUC for genes (orthologs) within and outside of phenolog intersections (**Figure 4.7**), considering all significant (5% FDR) yeast-worm phenologs with at least 4 genes in both the phenolog intersection ortholog set and the ortholog set outside the intersection. In order to correct for possible query gene size effects, I sub-sampled the larger of the two sets. For example, if the intersection of a worm phenotype and a yeast phenotype has 30 orthologs and the yeast phenotype has 15 additional

orthologs, I calculated the AUC of the 15 additional orthologs, then randomly sampled 15 genes at a time from the intersection set, calculating the AUC of each subset of 15 genes, taking the median value of 100 such samplings as the AUC for the intersection set.

Treatment of animals

Animal care met the principles and guidelines of the Institute for Laboratory Animal Research ‘‘Guide for Care and Use of Laboratory Animals’’ and the University of Texas at Austin Institutional Animal Care and Use Committee.

***Xenopus laevis* embryo manipulations**

Female *Xenopus laevis* were ovulated overnight after injecting human chorionic gonadotropin, and eggs were squeezed out for fertilization *in* vitro. At the two cell stage, the jelly layer of embryos was removed by swirling in 3% cysteine (pH 7.9) in 1/3x MMR medium and washed in 1/3xMMR five times. For microinjections, embryos were placed in 2% Ficoll in 1/3xMMR, and injected using forceps and an Oxford universal micromanipulator, then reared in 2% Ficoll in 1/3xMMR to stage 9, then washed and reared in 1/3xMMR.

Whole-mount *in situ* hybridization was performed using a modified method omitting acetylation steps from the standard method [21]. For all experiments, morpholino antisense oligonucleotides (MOs) were injected at 20-60ng/blastomere. To target ciliated epidermis, injections were made into the two ventral cells at the 4 cell stage. Two dorsal cells were injected to analyze neural tube closure. The posterior cardinal vein and intersomitic veins were targeted by injecting into the two ventral cells equatorially at the 4 cell stage. For whole mount *in situ* hybridization for Erg and XMsr,

embryos were fixed in MEMFA medium at stage 34 to 36. The hemorrhage phenotype was photographed at stage 45 after anesthetizing with Benzocaine.

Images of embryos were obtained with a Leica MZ16FA stereomicroscope using ImageProPlus software.

All methods involving *Xenopus* were performed by Tae Joo Park.

Confocal imaging

For epidermal cilia staining, embryos were fixed in MEMFA at stage 25 to 27 and washed with Ptw solution (PBS+0.1% Tween-20). The embryos were incubated with mouse anti-α-tubulin IgG in Ptw for 30 min. After washing with Ptw, the embryos were incubated with Alexa Fluor 555 goat anti-mouse IgG in Ptw for 30 min followed by washing in Ptw. Actin filaments were co-stained using Alexa 488 conjugated Phalloidin. In all cases, embryos were mounted in Ptw and 3D projections of cilia were made by collecting overlapping sections with a Zeiss LSM5 PASCAL confocal microscope. 3D projections, image processing, and image analysis were performed with LSM5, Image ProPlus, and Adobe Photoshop software. All confocal methods were performed by Tae Joo Park.

Morpholino oligonucleotides and cDNA clones

xSox12, Erg, and XMsr cDNAs were obtained from Open BioSystem (xSox12: IMAGE:6636177, Erg: IMAGE:5512670, XMsr: IMAGE:8321886). Centrin-GFP was obtained from Dr. Chris Kintner at the Salk institute. Translation blocking antisense morpholinos for IFT140, RFX2, and xSox12 were designed based on the sequences from

the NCBI database (IFT140: x17243.1, RFX2: BC108517.1, xSox12: BC068647.1). MOs were obtained from Gene Tools. All MO sequences are listed below:

IFT140-MO: 5'-TTCCTAAGGCACTCCAGTCACCCAT-3'

RFX2-MO: 5'-AATTCTGCATACTGGTTTCTCCGTC-3'

xSox12-MO: 5'-TCACCCTGTATGGTATCCATTTAAG-3'

xSox12-MM: 5'-TCAGCCTCTATGCTATGCATTCAAG-3'

Morpholinos were designed by Tae Joo Park.

RESULTS

Computational Results

To systematically discover phenologs, I collected from the literature a set of 1,923 human disease-gene associations [6], 74,250 transgenic mouse phenotype-gene associations [9], 27,065 *C. elegans* gene-phenotype associations [7], and 86,383 yeast gene-phenotype associations [8, 10, 15, 18], spanning ~300 human diseases and >6,000 model organism phenotypes. With these data and the sets of orthologous gene relationships between each pair of organisms [19], I quantitatively examined each inter-organism phenotype pair, measuring the significance of each (**Figure 4.4**). I corrected for testing multiple hypotheses by repeating all analyses 1,000 times with randomly permuted gene-phenotype associations to calculate a false discovery rate (FDR) based upon the observed null distribution of scores (**Figures 4.5**). With this correction, I observe of 154 significant phenologs (5% FDR) between human diseases and yeast mutational phenotypes, 3,755 between human and mouse, 147 between mouse and worm, 105 between mouse and yeast, and 206 between yeast and worm, and 9 between human and worm (the low number stems from limited mutational data in both species) (**Figure 4.6**).

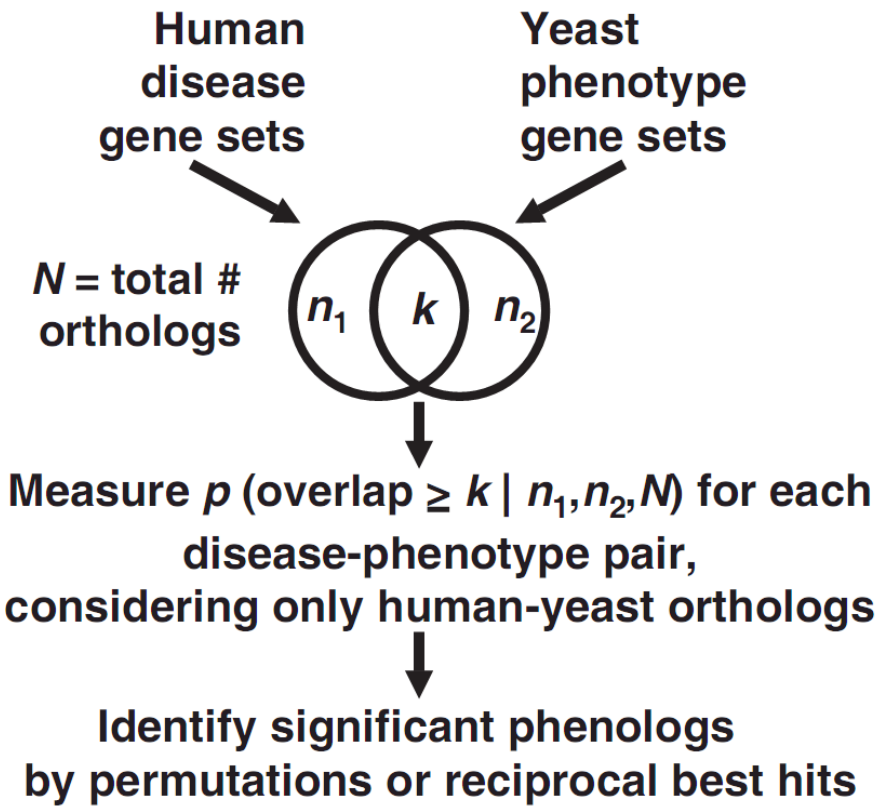


FIGURE 4.4 SYSTEMATIC IDENTIFICATION OF PHENOLOGS. For a pair of organisms, sets of genes known to be associated with mutational phenotypes are assembled, considering only orthologous genes between the two organisms. Pairs of mutational phenotypes—one phenotype from each organism, each associated with a set of genes—are then compared to determine the extent of overlap of the associated gene sets, calculating the significance of overlap by the hypergeometric probability. Figure adapted from work in review [42].

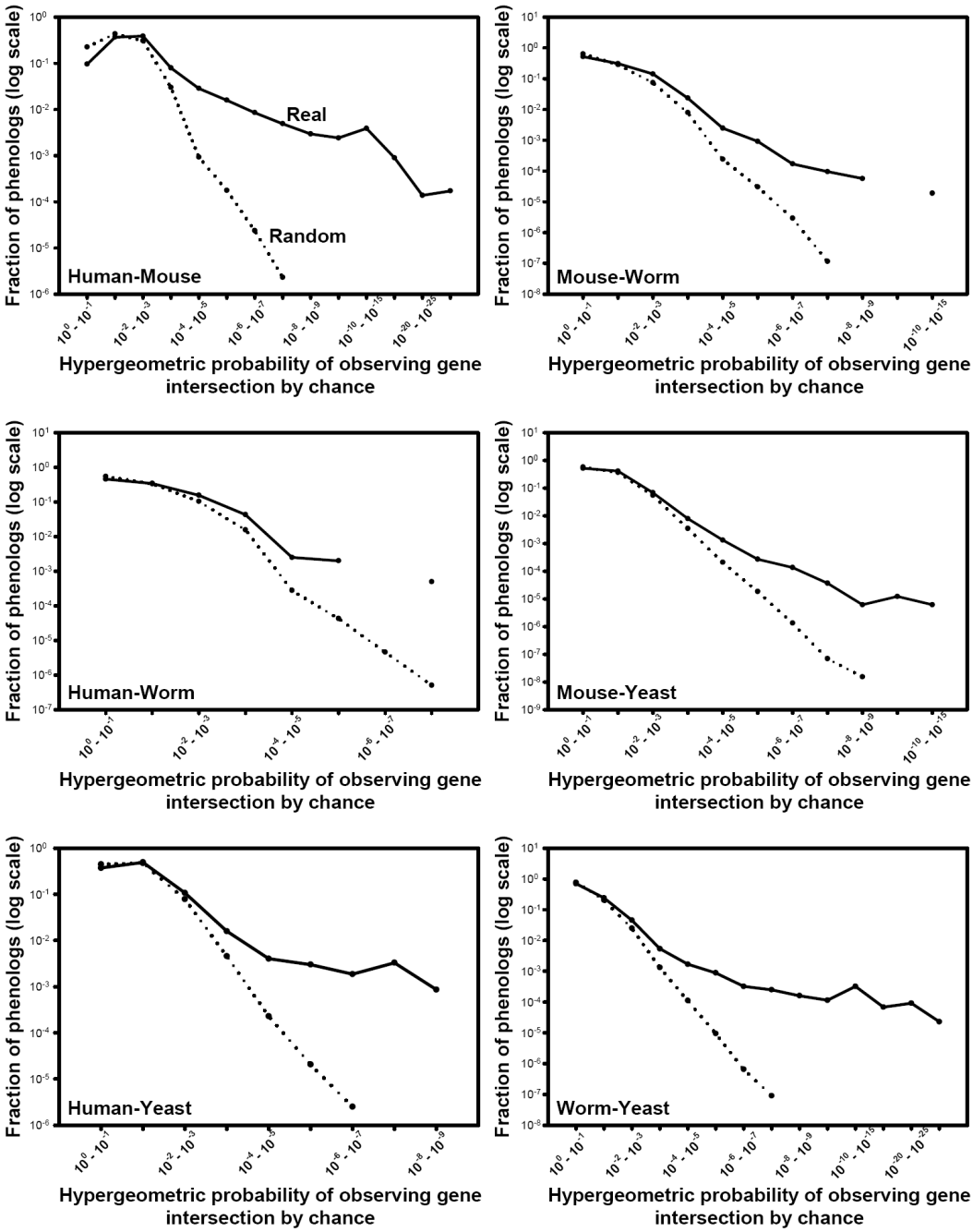


FIGURE 4.5 MANY MORE ORTHOLOGOUS PHENOTYPES ARE OBSERVED THAN EXPECTED BY RANDOM CHANCE as revealed by comparison of the distribution of observed probabilities with those derived from the same analysis following permutation of gene-phenotype associations, as shown in all pairwise comparisons of the mutational phenotypes from mouse, human, yeast, or worm. Figure from work in review [42].

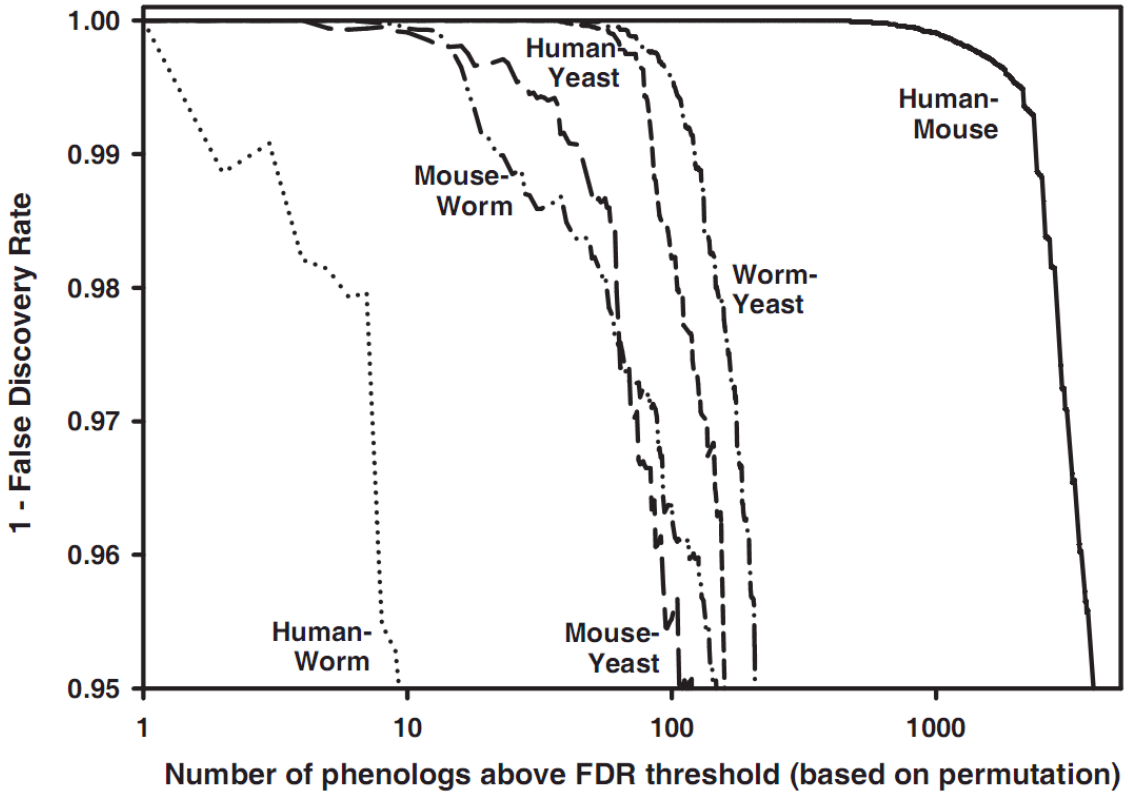


FIGURE 4.6 COUNT OF PHENOLOGS ABOVE A FALSE DISCOVERY RATE THRESHOLD for all pairwise comparisons of the mutational phenotypes from mouse, human, yeast, or worm. Figure adapted from work in review [42].

Phenologs identifies obviously equivalent phenotypes

Many intuitively obvious phenologs are identified in this manner, which serve as positive controls: nonviable *C. elegans* (RNAi) are found to be phenologous to inviable yeast (gene deletion), given that of 705 worm genes (with yeast orthologs) associated with nonviability, and 653 yeast genes (with worm orthologs) associated with nonviability, 369 orthologs are shared between these sets ($p \leq 10^{-33}$). Embryonic lethality before somite formation in mice is found to be phenologous to nonviable *C. elegans* following RNAi ($p \leq 10^{-5}$). Mouse pre-/peri-natal lethality or embryogenesis defects are phenologous with sterility in *C. elegans* following RNAi ($p \leq 10^{-6}$). Many lethality, sterility, and embryonic developmental phenotypes are related across organisms.

Importantly, many more specific phenologs are revealed, especially for the comparison of mouse and human phenotypes; these recapitulate many known mouse models of disease, serving as additional positive controls. **Table 4.1** lists specific examples. For example, one of the most significant phenologs identified between human disease and mouse mutational phenotypes is that linking Bardet-Biedl syndrome with four mouse traits, each of which relates to the disruption of ciliary function (abnormal brain ventricle/choroid plexus morphology, small hippocampus, enlarged third ventricle, absent sperm flagella; all $p \leq 10^{-11}$), consistent with the apparent molecular defects in Bardet-Biedl syndrome [22]. The argument is thus that mouse ciliary defects provide a powerful model for studying human Bardet-Biedl syndrome, consistent with its recently recognized utility in this regard. Similarly, human cataracts are observed to be phenologous to mouse cataracts ($p \leq 10^{-24}$), human obesity is phenologous to mouse obesity ($p \leq 10^{-14}$), human deafness to mouse deafness ($p \leq 10^{-29}$), human retinitis to mouse retinal degeneration ($p \leq 10^{-26}$), and human goiter to mouse enlarged thyroid

glands ($p \leq 10^{-8}$). Thus, the calculation of phenologs correctly identifies many known mouse models of human diseases and therefore has the potential to identify new models.

Techniques developed for identifying homologous genes can be applied to phenologs

Much of the powerful conceptual framework established for gene sequence homology and orthology may be applicable to phenologs. For example, equivalent phenotypes might be defined on the basis of homologous, rather than orthologous, gene sequences, in this manner examining the divergence of phenotypic outcome of homologous systems. Similarly, many of the algorithmic approaches used to identify orthologous genes might also be applied to the identification of phenologs. I explored this notion for one effective and easily automated approach to identify orthologous sequences, the reciprocal best hit (RBH) strategy. The RBH criterion holds that genes *X* and *Y* are orthologs if gene *X* is the most similar sequence to gene *Y* when searched genome-wide, provided the reciprocal search is also true. I adapted the RBH criterion to the identification of phenologs in order to identify the most equivalent phenotypes between two organisms from among those assayed, by asking if the phenotypes have the most significant (lowest *p*-value) gene overlaps with each other when searched against all phenotypes in their respective organisms. Such analysis gives a second criterion for identifying phenologs, useful for legitimate phenologs with poor *p*-values due to limited phenotypic data sets. Examples of such RBH phenologs are indicated in **Table 4.1**.

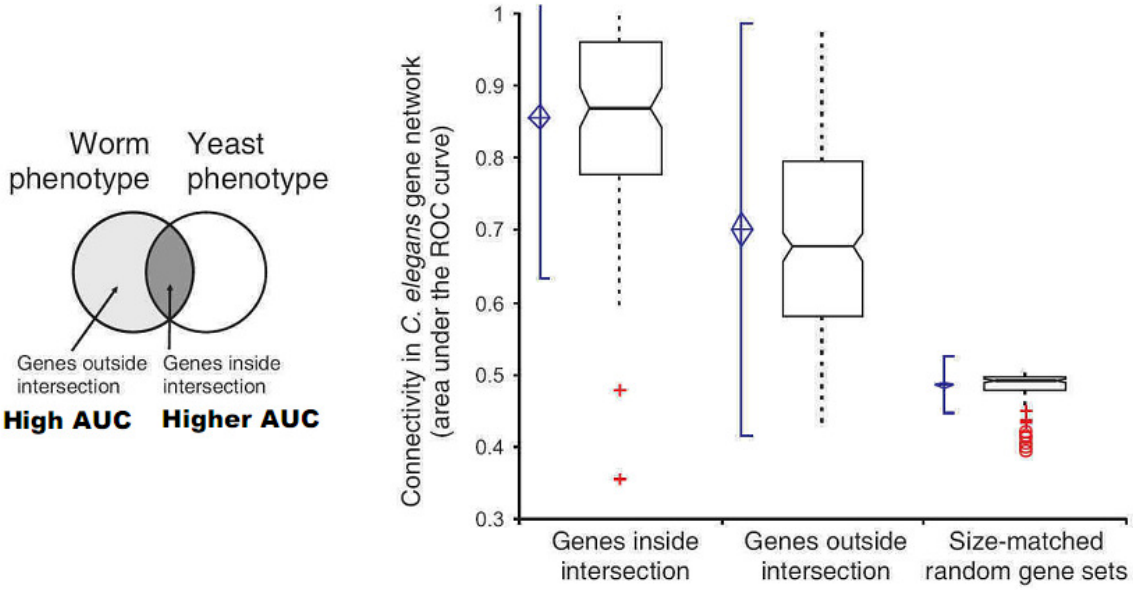
Phenologs identify dense subnetworks in functional network

Phenologs imply that although phenotypes diverge, the orthology of the underlying gene networks and probably their immediate functional output is conserved. I might therefore expect genes involved in a given phenolog to represent a coherent

biological module, and thus to be highly interconnected in gene networks. Moreover, one might expect that the genes already confirmed to show the signature phenotypes in both organisms (e.g., the intersection labeled by *k* in **Figure 4.4**) would be even more highly interconnected than the genes associated with the signature phenotype in only one organism; these latter genes might or might not belong to this sub-network, as multiple mechanisms might give rise to the phenotype. Evidence in current gene networks of more linkages among the genes in each such intersection would support this notion of phenologs recapitulating modular subnetworks. I therefore systematically tested all significant phenologs involving yeast and worm genes for the genes' connectivity in available functional networks [15, 17]. I find the network connectivity of genes in phenolog intersections to be significantly higher ($p < 0.0001$; Wilcoxon signed-rank) than the phenolog genes outside of the intersections, which nonetheless show significantly higher network connectivity than random size-matched gene sets ($p < 0.0001$) (**Figure 4.7**). This indicates that phenologs do identify evolutionarily conserved subnetworks of genes relevant to particular phenotypes or diseases, while still predicting new candidate genes significantly better than random expectation.

	***Phenotype*$_1$**		***Phenotype*$_2$**	n_1	n_2	k	p-value	PPV
Hs	cataracts	Mm	cataracts	19	47	11	$6x10^{-24}$	1.00
Hs	X-linked conductive deafness	Mm	circling	47	50	12	$2x10^{-20}$	1.00
Hs	**Bardet-Biedl syndrome**	Mm	**absent sperm flagella**	11	5	4	$8x10^{-13}$	1.00
Mm	**lymphoma**	Sc	**CANR mutator high**	14	11	6	$1x10^{-11}$	1.00
Hs	**Zellweger syndrome**	Sc	**reduced number of peroxisomes**	8	6	4	$1x10^{-9}$	1.00
Hs	xeroderma pigmentosum	Sc	high UVC irradiation sensitivity	7	9	4	$5x10^{-9}$	1.00
Hs	**susceptible to autism**	Mm	**abnormal social investigation**	5	16	3	$1x10^{-8}$	1.00
Hs	**susceptible to neural tube defects**	Mm	**abnormal circulating amino acid level**	3	32	2	$1x10^{-5}$	1.00
Hs	**porphyria**	Sc	**damnacanthal sensitive**	4	4	2	$2x10^{-5}$	1.00
Mm	**abnormal heart development**	Ce	**male tale morphology abnormal**	52	7	4	$5x10^{-7}$	1.00
Mm	pre-/peri-natal lethality	Ce	sterile	498	344	66	$1x10^{-6}$	0.99
Mm	**abnormal angiogenesis**	Sc	**lovastatin sensitive**	8	67	5	$1x10^{-6}$	0.99
Mm	**Spleen hypoplasia**	Sc	**uge (enlarged cells)**	5	16	3	$3x10^{-6}$	0.99
Mm	gastrointestinal hemorrhage	Ce	abnormal body wall muscle cell polarization	6	3	2	$4x10^{-6}$	0.98
Hs	breast/ovarian cancer	Ce	high incidence male progeny	12	16	3	$7x10^{-6}$	0.98
Hs	**achromatopsia**	Ce	**chemotaxis defective**	3	9	2	$1x10^{-5}$	0.98
Hs	**congenital disorder of glycosylation**	Sc	**CID 604586 sensitive**	10	25	3	$2x10^{-4}$	0.98
Hs	hemolytic anemia	Sc	hydroxyurea sensitive	11	23	3	$2x10^{-4}$	0.98
Mm	**abnormal olfactory neuron morphology**	Ce	**dauer constitutive**	7	4	2	$1x10^{-5}$	0.97
Hs	**glycogen storage disease**	Sc	**glycogen storage reduced**	3	20	2	$2x10^{-4}$	0.97
Hs	**amyotrophic lateral sclerosis**	Sc	**increased resistance to wortmannin**	2	34	2	$2x10^{-4}$	0.97
Mm	**abnormal placenta**	Sc	**sorbitol sensitive**	8	14	3	$1x10^{-5}$	0.96
Mm	**abnormal endocardium morphology**	Sc	**cantharidin sensitive**	2	11	2	$2x10^{-5}$	0.95
Hs	**somatic basal cell carcinoma**	Ce	**egg size abnormal**	1	3	1	$6x10^{-4}$	0.77
Hs	**hypothyroidism**	Ce	**blistered cuticle**	3	2	1	$1x10^{-3}$	0.75

TABLE 4.1 EXAMPLES FROM THE >6,000 SIGNIFICANT PHENOLOGS DETECTED among human (Hs) diseases and mouse (Mm), yeast (Sc), and worm (Ce) mutant phenotypes. n1 indicates the number of orthologs in organism 1 with phenotype1, n2 the number in organism 2 with phenotype2, and k the number in both sets. The significance of each phenolog is assessed by the hypergeometric probability (p-value), the positive predictive value (PPV) when considering multiple testing (1 – false discovery rate), and the reciprocal best hit criterion (bold text). Table from work in review [42].


Worm phenotype
Yeast phenotype
Genes outside intersection
Genes inside intersection
High AUC
Higher AUC
Connectivity in C. elegans gene network (area under the ROC curve)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Genes inside intersection
Genes outside intersection
Size-matched random gene sets

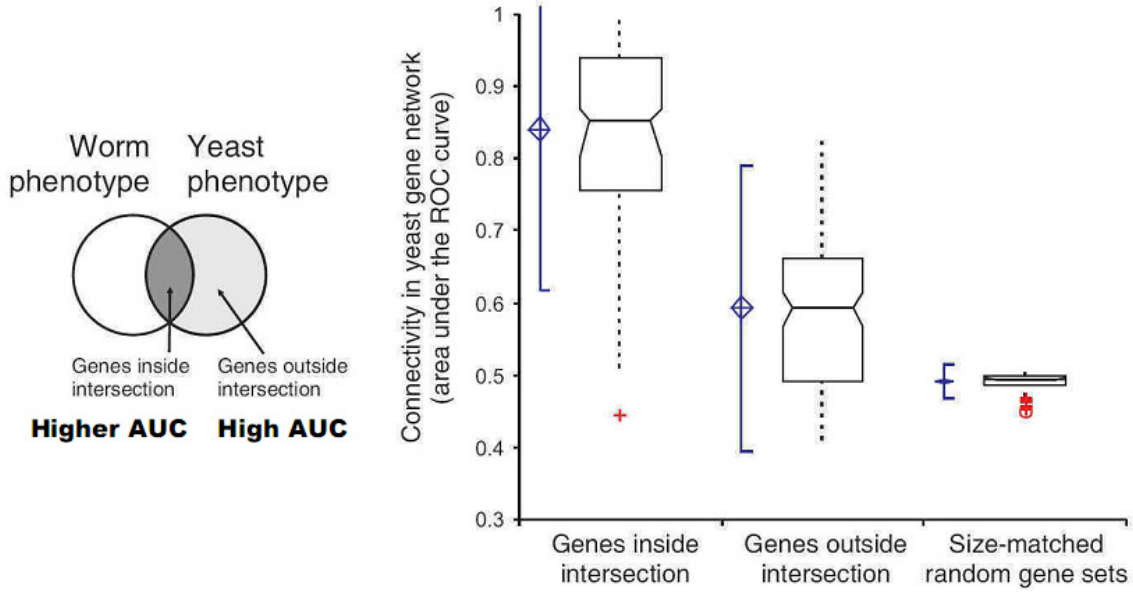

Worm phenotype
Yeast phenotype
Genes inside intersection
Genes outside intersection
Higher AUC
High AUC
Connectivity in yeast gene network (area under the ROC curve)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Genes inside intersection
Genes outside intersection
Size-matched random gene sets

FIGURE 4.7 GENES INVOLVED IN PHENOLOGS SHOW ENHANCED INTERCONNECTIVITY IN GENE NETWORKS, shown here for worm (top) and yeast (bottom) gene networks [17, 20]. All significant yeast-worm phenologs with at least 4 orthologs in both the 'intersection' and 'non-intersection' sets (see Methods) were tested for network connectivity, measured as the area under a receiver-operator characteristic (ROC) plot as described in [15], with values ranging from 0.5 (random network connectivity) to 1 (high network connectivity). Genes from phenolog intersections show significantly higher network connectivity than genes associated with a phenolog, but outside of the intersection, which in turn show significantly higher connectivity than size-matched random gene sets. Thus, phenologs capture subnetworks or network modules informative about a given phenotype pair, and carry predictive value for additional genes relevant to the phenotypes. At the left of each box-and-whisker plot, the center of the blue diamond indicates the mean AUC across phenologs, the top and bottom of the diamond indicate the 95% confidence interval, and the accompanying solid vertical line indicates ± 2 standard deviations. The bottom, middle, and top horizontal lines of the box-and-whisker plots represent the first quartile, the median, and the third quartile of AUCs, respectively; whiskers indicate 1.5 times the interquartile range. Red plus signs represent individual outliers. Figure adapted from work in review [42].

Experimental Results

Experimental confirmation of a yeast model for vertebrate angiogenesis

The power of the phenolog framework lies in discovery of non-obvious disease models. I observed just such a serendipitous phenolog between abnormal angiogenesis in mutant mice and reduced growth rate of yeast deletion strains when grown in the hypercholesterolemia drug lovastatin (8 mouse genes, 67 yeast, 5 shared, $p \leq 10^{-6}$; **Figure 4.8**). This observation, consistent with the action of lovastatin in reducing tumor-induced angiogenesis (e.g., [23]), suggests that budding yeast, which entirely lack blood vessels, could potentially model aspects of mammalian vasculature formation, and help to define genes affecting this process. In particular, the five shared genes between these processes are, in yeast, the mitogen activated protein (MAP) kinases SLT2, PBS2, and HOG1, the calcineurin B protein CNB1, and the uncharacterized protein VPS70; the four characterized proteins regulate osmosensing and aspects of cell wall organization and biogenesis. Strikingly, mutations of their mouse orthologs (MAPK7, MAP2K1, MAPK14, PPP3R1, and the prostate-specific membrane antigen PSMA, respectively) all show strong angiogenesis defects—e.g., MAPK7 deletion causes defective blood vessel and cardiac development [24]; ablation in adult mice leads to leaky blood vessels [25]. Similarly, PSMA regulates angiogenesis by modulating integrin signal transduction [26]. Thus, this conserved subnetwork of genes was alternately repurposed to regulate osmosensing and cell wall biogenesis in yeast cells and proper formation and maintenance of blood vessels in mice.

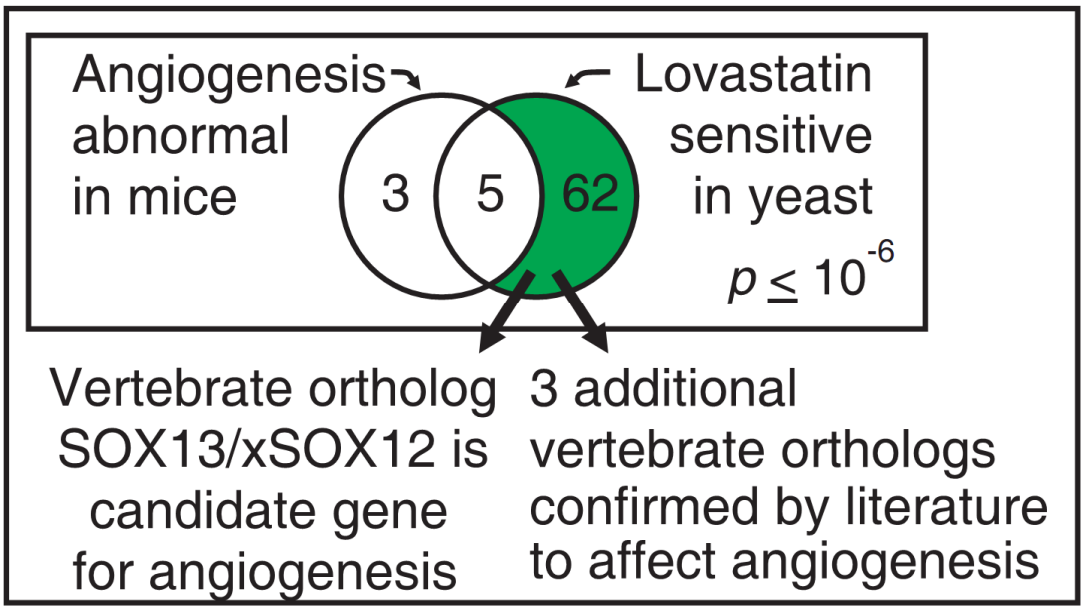


FIGURE 4.8 EXAMPLE OF A NON-OBVIOUS DISEASE MODEL REVEALED BY PHENOLOGS: YEAST MUTANTS SENSITIVE TO THE HYPERCHOLESTEROLEMIA DRUG LOVASTATIN PREDICT MAMMALIAN ANGIOGENESIS DEFECTS. The set of 8 genes (considering only mouse/yeast orthologs) associated with mouse angiogenesis defects and the set of 67 genes associated with lovastatin hypersensitivity in yeast significantly overlap, suggesting that the yeast gene set may predict angiogenesis genes. This prediction was verified in Xenopus embryos for the case of the transcription factor xSOX12. Figure adapted from work in review [42].

Orthology of phenotypes predicts that additional human orthologs of genes associated with a phenologous model organism trait are more likely to be associated with the human disease. I therefore examined the yeast angiogenesis model for other yeast genes (with mammalian orthologs) whose deletion induced sensitivity to lovastatin. Of the 62 candidates, three of the corresponding mouse genes were confirmed by literature to function in angiogenesis, but had yet to be annotated as such. These genes included the known target of lovastatin, HMG-CoA reductase, whose role in angiogenesis has been previously observed [27], the sirtuin SIRT1, whose disruption in zebrafish and mice caused defective blood vessel formation and blunted ischemia-induced neovascularization [28], and the casein kinase Csnk2a1, inhibitors of which inhibit mouse retinal neovascularization [29]. Additional genes were involved in other aspects of cardiovascular development, such as the gene mitoferrin, being expressed most highly in hematopoietic organs, fetal liver, bone marrow, and spleen, and mutations in which block terminal erythroid maturation, leading to profound anemia [30]. Similarly, SMAP1 positively regulates erythrocyte differentiation [31]. Thus, mammalian orthologs of the 62 yeast lovastatin-sensitivity genes include additional genes relevant to cardiovascular development, supporting the notion that a yeast model might predict angiogenesis genes.

In order to more directly evaluate predictions of this phenolog, 13 of the 62 genes not already associated with angiogenesis were tested in the frog *Xenopus laevis*. Using whole mount *in situ* hybridization, my collaborators examined mRNA expression of the *Xenopus* orthologs for patterns relevant to angiogenesis. The gene xSOX12 (the *Xenopus* ortholog of mammalian SOX13, a transcription factor known to regulate T lymphocyte differentiation [32] and to be expressed in mouse arterial walls [33]) was prominently expressed in the posterior cardinal vein, intersomitic veins, and developing heart,

consistent with a role affecting developing vasculature (**Figure 4.9**). They knocked down xSOX12 expression using microinjection of morpholino antisense oligonucleotides (MO) and assayed for vasculature defects by *in situ* hybridization to the vasculature reporter genes Erg and XMsr (**Figure 4.10**). The knockdown of xSOX12 leads to a strong defect in angiogenesis, with morpholino injected animals largely lacking intersomitic and posterior cardinal veins. By later stages, hemorrhaging was apparent in morphants due to the defective vasculature (**Figure 4.11**). Thus, xSOX12/SOX13 is a novel regulator of angiogenesis, discovered in the absence of any previous functional data linking it to angiogenesis, on the basis of orthology between mouse angiogenesis defects and yeast lovastatin sensitivity. Notably, these data also demonstrate that differentiation both of blood cells [32] and blood vessels are controlled by the same transcription factor.

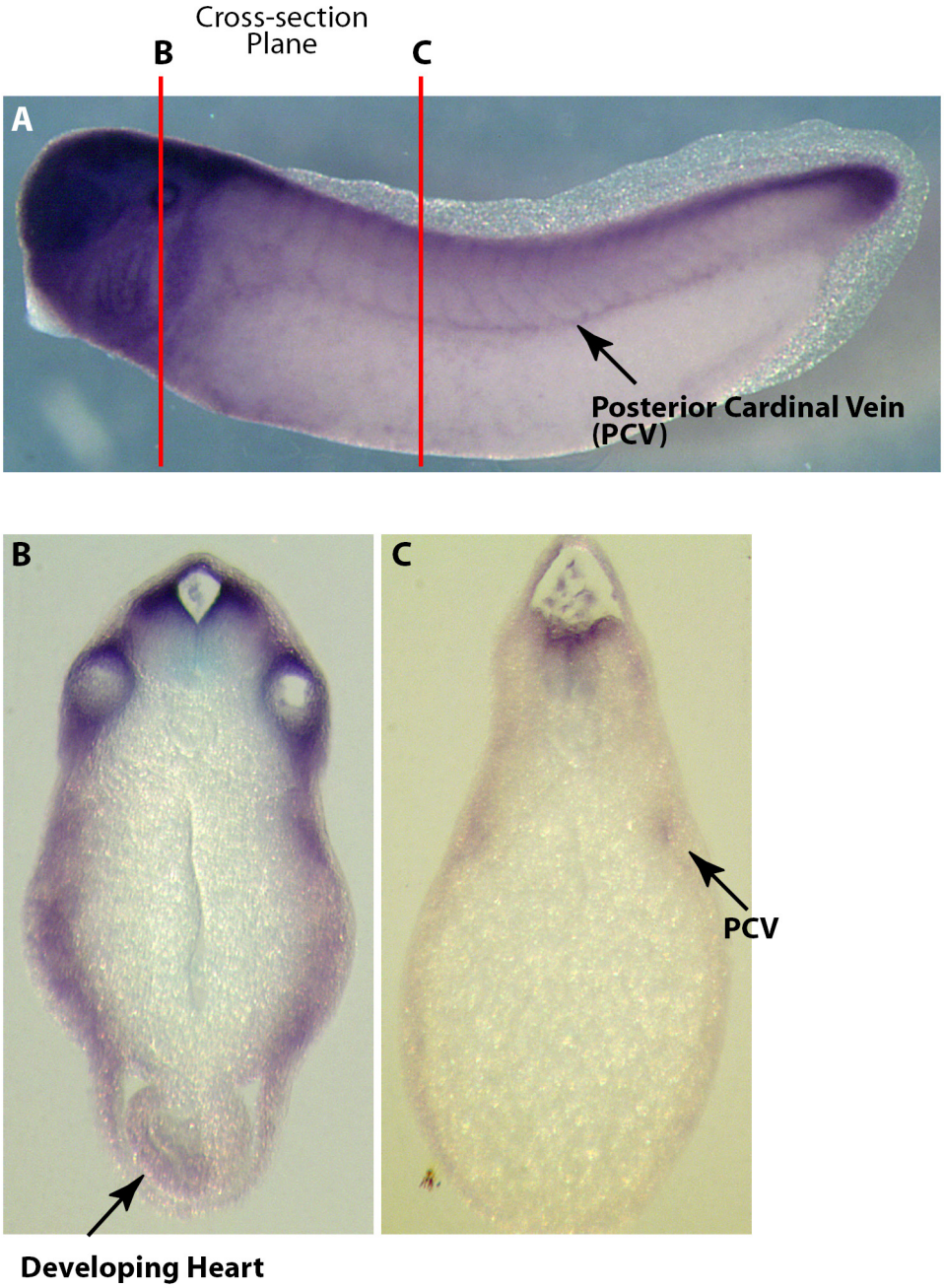


FIGURE 4.9 IN SITU HYBRIDIZATION SHOWS XSOX12 EXPRESSION IN VEINS AND DEVELOPING HEART OF A STAGE 32 XENOPUS EMBRYO. Figure from work in review [42].

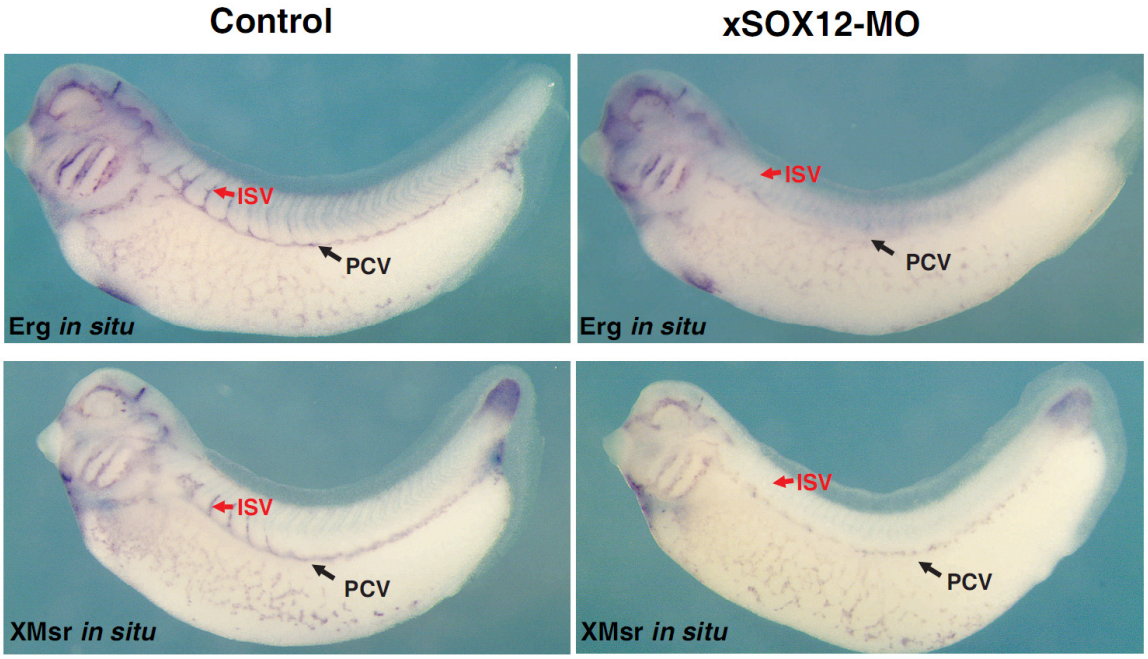


FIGURE 4.10 MORPHOLINO (MO) KNOCKDOWN OF XSOX12 INDUCES DEFECTS IN VASCULATURE, measured using in situ hybridization versus two independent markers of the vasculature, the angiogenesis-regulating transcription factor Erg (defects observed in 31 of 49 animals tested) and the angiotensin receptor homolog XMsr (12 of 19 animals tested). Such defects are rare in untreated control animals and 5 base pair mismatch morpholino (MM) knockdowns (0 of 22 control animals tested with XMsr, 2 of 46 tested with Erg; 5 of 28 MM animals tested with Erg). Figure adapted from work in review [42].

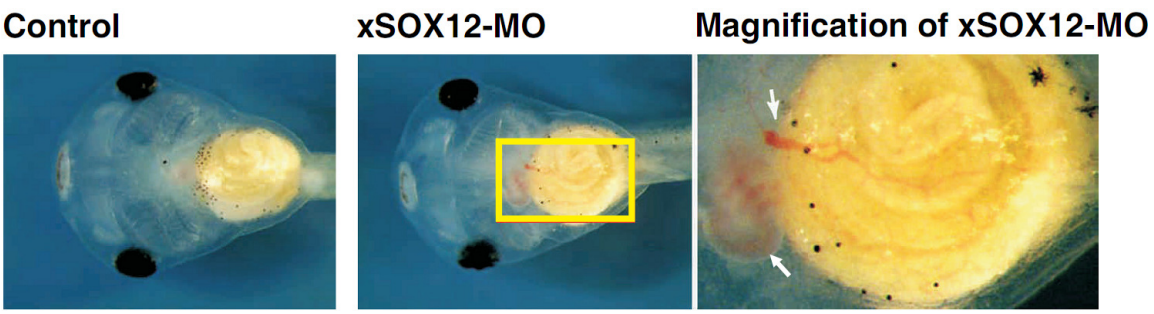


FIGURE 4.11 HEMORRHAGING IS APPARENT IN STAGE 45 XENOPUS EMBRYOS DUE TO DYSFUNCTIONAL VASCULATURE FOLLOWING XSOX12 MORPHOLINO KNOCKDOWN (12 of 50 animals tested; 2 also showed unusually small hearts with defective morphology; right-hand panel magnifies yellow boxed region in middle panel), but is rare in control animals (1 of 45 tested untreated animals, 1 of 22 xSOX12-MM knockdown animals tested). All phenotypes in Figures 4.10 and 4.11 are significantly different from controls by chi-square tests ($p < 0.001$). Figure adapted from work in review [42].

Experimental confirmation of a worm model for neural tube defects

Given a phenolog for a human disease, any approach for associating more genes with the model organism trait, e.g., a genetic screen, will suggest new human disease gene candidates. I used this approach and a phenolog between abnormal *C. elegans* cilia morphology and mouse neural tube defects—consistent with a known role for cilia in neural tube formation [34]—to identify new genes affecting vertebrate neural tube closure (**Figure 4.12**). Defects in neural tube closure are among the most common and debilitating human birth defects, afflicting nearly 1 in 1,000 live births world-wide [35], yet they have a complex genetic basis and knowledge of the underlying genes is still incomplete. We first tested a direct prediction of the phenolog to confirm that knockdown of the vertebrate intraflagellar transport gene IFT140 causes defective ciliogenesis and failure of neural tube closure in developing *Xenopus* embryos (**Figure 4.13**). We then applied the emerging technique of network-guided genetics [17] to prioritize the transcription factor *daf-19,* a master regulator of worm ciliogenesis, as the gene most likely to show a similar effect (based on known genetic interactions to the cilia morphology defect genes). We knocked down the *Xenopus* ortholog of this gene, RFX2, and observed a defect in the developing neural tube at stage 20 (**Figure 4.13**), confirming RFX2's association with neural tube defects for the first time in a vertebrate. As RFX2 is a transcription factor, it might potentially control many downstream processes; analysis of an early marker of ciliated cell fate specification (TEX15 [36]) confirms that ciliated cells are intact in the RFX2 knockdown animals (**Figures 4.14**). Characterization of the precise defects of IFT140 and RFX2 knockdown in *Xenopus* shows normal deployment of basal bodies but marked reduction of cilia on multi-ciliated epithelial cells if either gene is knocked down (**Figure 4.13**). Given the good mechanistic and genetic agreement

between *Xenopus* and mammalian neural tube closure [37], there is thus a high likelihood that defects in these genes are associated with human neural tube birth defects.

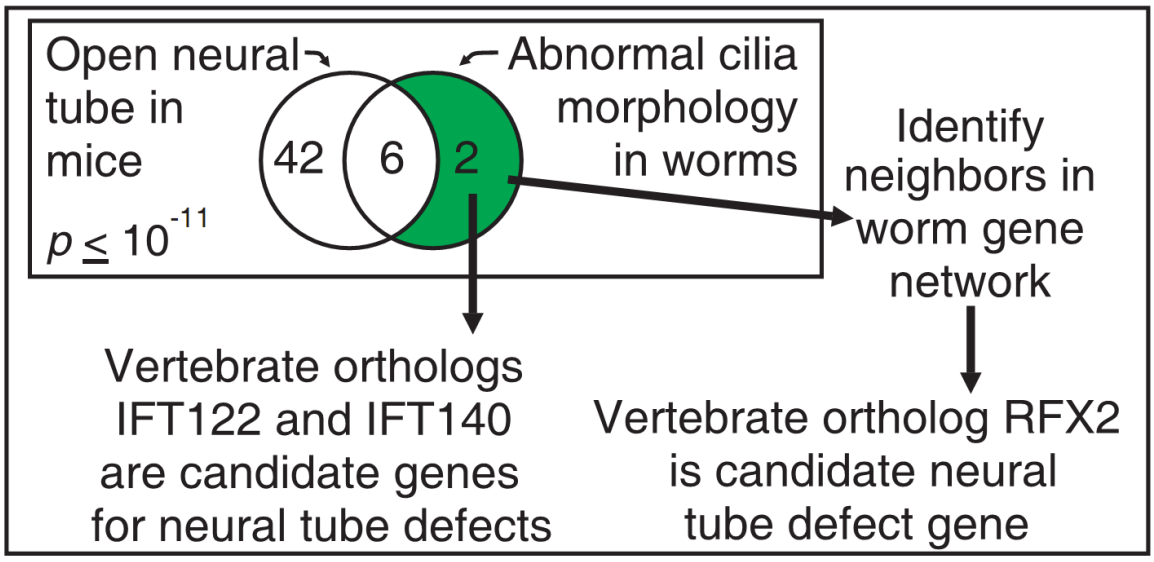


FIGURE 4.12 SCHEMATICALLY REPRESENTATION OF THE VALIDATION OF TWO NEW NEURAL TUBE DEFECT GENES PREDICTED BY PHENOLOGS AND GENE NETWORKS. Figure adapted from work in review [42].

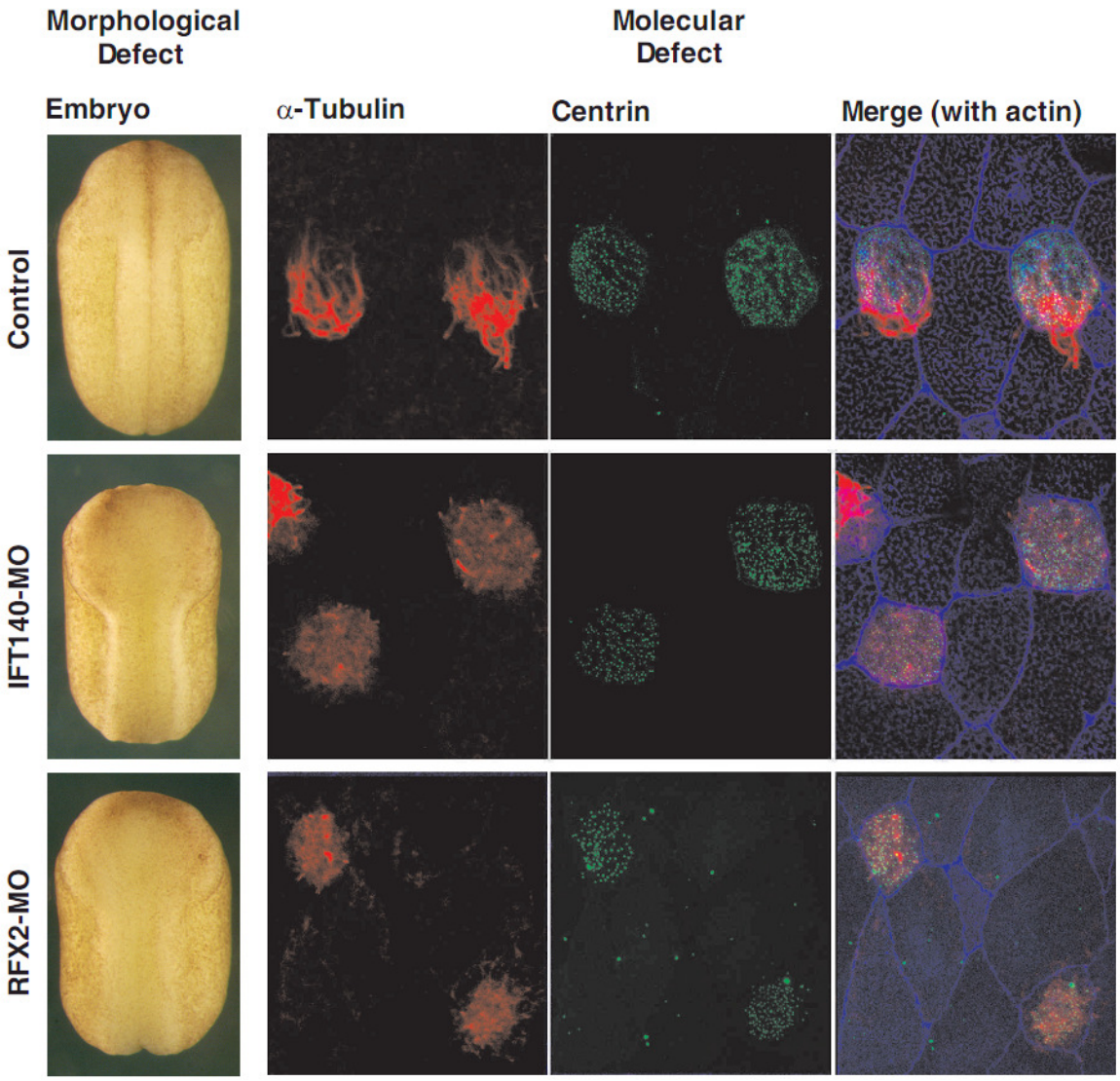


FIGURE 4.13 MORPHOLINO KNOCKDOWNS OF XENOPUS GENES RFX2 AND IFT140 SHOW STRONG NEURAL TUBE DEFECTS (left column), in contrast to control animals. (RFX2-MO, 41 of 43 animals tested; IFT140-MO, 46 of 52 tested; untreated control, 0 of 55 tested.) Immunofluorescence of the Xenopus ciliated epithelium from IFT140 or RFX2 morpholino knockdown animals reveals normal deployment of basal bodies (centrin marker) but abnormal or missing cilia (alpha-tubulin marker) on multi-ciliated epithelial cells. Figure adapted from work in review [42].

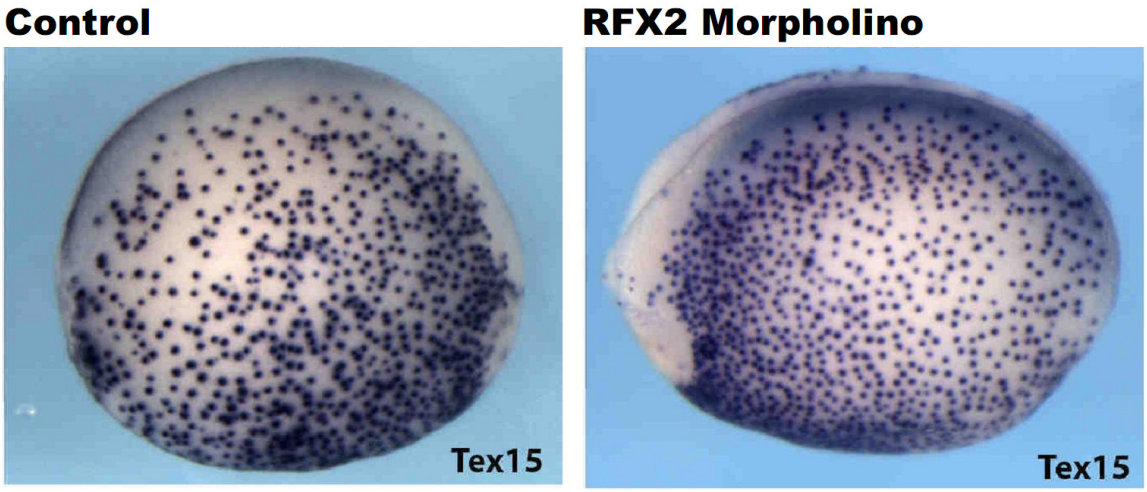


FIGURE 4.14 RFX2-MO KNOCKDOWN ANIMALS SHOWS THAT CILIATED CELLS ARE INTACT, BUT LACK CILIA as shown in a representative in situ hybridization versus TEX15, a marker of ciliated cell fate specification [36]. The numbers of ciliated cells visible per embryo did not differ significantly between control and RFX2-MO embryos (13 control embryos were scored, with 6 showing high numbers of ciliated cells, 4 medium, 3 low; 11 RFX-MO embryos were scored showing 4 high, 6 medium, 1 low; no significant difference by chi-square test.) Figure adapted from work in review [42].

Discussion

Phenologs reflect the innate modularity of gene systems and help illuminate the prolific adaptive reuse of conserved genetic elements because they identify sets of genes that maintain a shared relationship across varied biological contexts. Within this framework, it is possible to address questions like, "Does a genetic module maintains a recognizable identity in single-cell yeast and in the blood vessels of vertebrates?" This approach identifies genetic modules that would otherwise be obscured when their rewiring with other downstream modules leads to divergent phenotypic outcomes in other organisms. The participation of multiple genetic modules in determining a shared phenotypic outcome may help explain why the genes in the intersection of two phenotypes are so tightly interlinked even relative to other genes associated with the same phenotype (**Figure 4.7**). I propose that one possible explanation for this trend is that phenologs are identified on the basis of a shared, conserved genetic module, but that other organism-specific modules (or organism-specific relationships among modules) determine the specific phenotypic outcome (**Figure 4.15**). When randomly sampling from the non-intersecting genes, my algorithm would sample only one module in the intersection, but would sample multiple modules among the rest of the genes involved in the phenotype, which would reduce the relative density of functional links among the genes, as measured by the area under the ROC curve. In the future, with a wider sampling of phenotypic data across taxa, it may eventually be possible to track the functional coherence of sets of genes over time, and piece together how they are rewired for different purposes in various organisms. Ultimately, this wider, comparative view will give greater insight into specific mechanisms, since it will capture not just the systems as they currently exist, but also illustrate the ways in which they can vary.

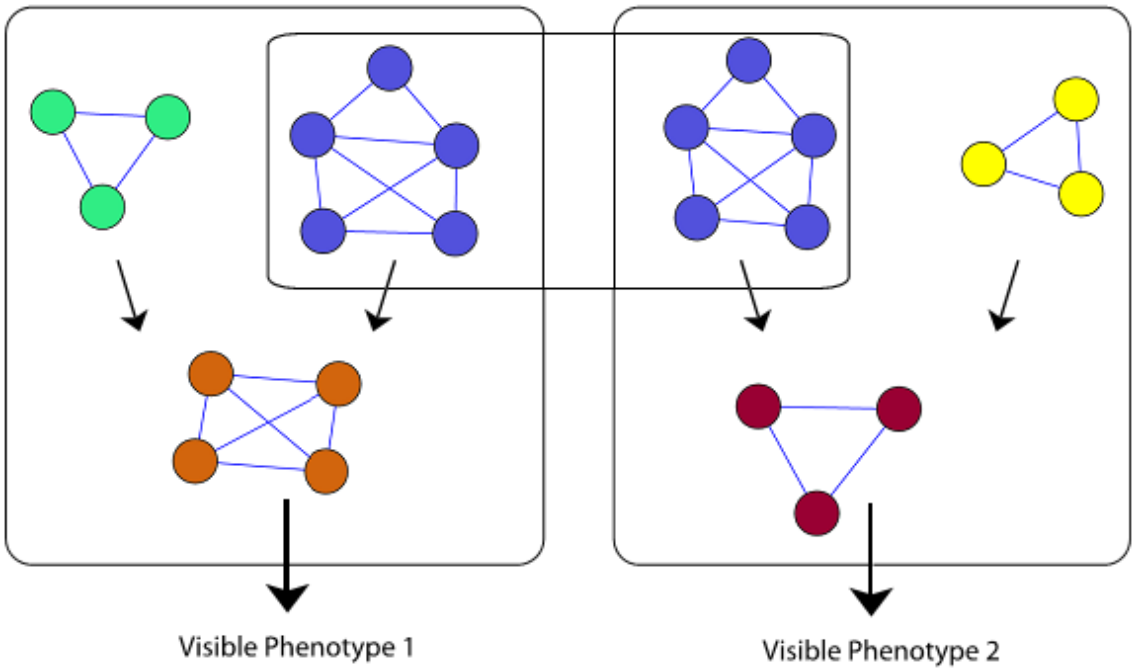


FIGURE 4.15 PROPOSED MODEL TO EXPLAIN GREATER FUNCTIONAL COHERENCE AMONG ORTHOLOGS INVOLVED IN BOTH PHENOTYPES RELATIVE TO ORTHOLOGS INVOLVED IN A SINGLE PHENOTYPE (see Figure 4.7). Phenologs may be identified primarily by the overlap of a single genetic module (overlapping orthologs are represented by blue circles, functional relationships by blue lines.). However, in each organism multiple additional non-conserved modules may be involved in the orthologous phenotype (represented by circles of various other colors), including downstream modules that give rise most directly to the organism specific phenotype. Sampling of the non-overlapping orthologs will reveal fewer functional links since members of different modules will have fewer functional links between them. However, as a set they remain somewhat coherent due to the modules internal links and links that connect the modules together.

CONCLUSION

This research shows that phenologs provide a rich framework for comparing mutational phenotypes with potential for finding non-obvious models of human disease. The phenologs also naturally identify candidate genes that have a clear relationship to human disease. This will facilitate the study of underlying mechanisms of diseases in simple, tractable model organisms with the confidence that the research will apply directly to understanding aspects of complex human diseases. To that end, the phenolog approach provides a quantitative heuristic for estimating the likely utility of a chosen organism as a model for a disease of interest. Additional experimental work will be needed to evaluate how useful this heuristic will be.

The combination of phenologs with the method of network guided genetics, explored in the previous chapter, has already identified a neural tube gene, RFX2, that is a very strong contender for playing a role in human neural tube defects. A more thorough exploration of the synergy between these two approaches may find refinements that provide even more impressive predictions.

Phenologs identify closely associated genes and provide insight into the systems involved in a phenotype, but detailed molecular work is still necessary to determine the exact mechanisms that ultimately lead to the phenotypic effects of the disruption of these systems. To this end, I have provided a web application to interactively search for phenologs, available at http://www.phenologs.org, making both the algorithm and the data readily available to those who study specific diseases or biological systems. At the beginning of the age of sequencing, identifying orthologous genes would have been

nearly impossible without algorithms like Needleman-Wunsch [38], Smith-Waterman [39], or BLAST [40]. However, prior to GenBank [41] and other sequence databases, collecting sequence data reported in the literature would make orthologous genes difficult to track down. Hopefully, this work provides an initial algorithm for identifying orthologous phenotypes and will further motivate the creation and use of a standard repository of gene-phenotype associations for all organisms, so that phenologs and their underlying genetic systems will be more readily identified.

This chapter has been reworked and expanded from a submitted paper [42] that is currently under review.

REFERENCES

1. Dryja TP, Cavenee W, White R, Rapaport JM, Petersen R, Albert DM, Bruns GA: Homozygosity of chromosome 13 in retinoblastoma. N Engl J Med 1984, 310:550-553.

2. Lu X, Horvitz HR: lin-35 and lin-53, two genes that antagonize a C. elegans Ras pathway, encode proteins similar to Rb and its binding protein RbAp48. Cell 1998, 95:981-991.

3. Fitch WM: Distinguishing homologous from analogous proteins. Syst Zool 1970, 19:99-113.

4. Owen R: Lectures on Comparative Anatomy and Physiology of the Invertebrate Animals. London: Longmans, Brown, Green and Longmans; 1843.

5. Darwin C: On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London: John Murray; 1859.

6. Amberger J, Bocchini CA, Scott AF, Hamosh A: McKusick's Online Mendelian Inheritance in Man (OMIM(R)). Nucleic Acids Res 2008.

7. Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen CK et al: WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. Nucleic Acids Res 2005, 33:D383-389.

8. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G et al: Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res 2002, 30:69-72.

9. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE: The mouse genome database (MGD): new features facilitating a model system. Nucleic Acids Res 2007, 35:D630-637.

10. Saito TL, Ohtani M, Sawai H, Sano F, Saka A, Watanabe D, Yukawa M, Ohya Y, Morishita S: SCMD: Saccharomyces cerevisiae Morphological Database. Nucleic Acids Res 2004, 32:D319-322.

11. Hodgkin J, Horvitz HR, Brenner S: Nondisjunction Mutants of the Nematode CAENORHABDITIS ELEGANS. Genetics 1979, 91:67-94.

12. Richardson AL, Wang ZC, De Nicolo A, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, Ganesan S: X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell 2006, 9:121-132.

13. Scanlan MJ, Gout I, Gordon CM, Williamson B, Stockert E, Gure AO, Jager D, Chen YT, Mackay A, O'Hare MJ et al: Humoral immunity to human breast cancer: antigen definition and quantitative analysis of mRNA expression. Cancer Immun 2001, 1:4.

14. OMIM: Online Mendelian Inheritance in Man (OMIM). In.: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

15. McGary KL, Lee I, Marcotte EM: Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes. Genome Biol 2007, 8:R258.

16. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, Baldarelli RM, Baya M, Beal JS, Bello SM et al: The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. Nucleic Acids Res 2005, 33:D471-475.

17. Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM: A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nat Genet 2008, 40:181-188.

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

19. Remm M, Storm CE, Sonnhammer EL: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 2001, 314:1041-1052.

20. Lee I, Li Z, Marcotte EM: An Improved, Bias-Reduced Probabilistic Functional Gene Network of Baker's Yeast, Saccharomyces cerevisiae. PLoS ONE 2007, 2:e988.

21. Harland RM: In situ hybridization: an improved whole-mount method for Xenopus embryos. Methods Cell Biol 1991, 36:685-695.

22. Ross AJ, May-Simera H, Eichers ER, Kai M, Hill J, Jagger DJ, Leitch CC, Chapple JP, Munro PM, Fisher S et al: Disruption of Bardet-Biedl syndrome ciliary proteins perturbs planar cell polarity in vertebrates. Nat Genet 2005, 37:1135-1140.

23. Feleszko W, Balkowiec EZ, Sieberth E, Marczak M, Dabrowska A, Giermasz A, Czajka A, Jakobisiak M: Lovastatin and tumor necrosis factor-alpha exhibit potentiated antitumor effects against Ha-ras-transformed murine tumor via inhibition of tumor-induced angiogenesis. Int J Cancer 1999, 81:560-567.

24. Regan CP, Li W, Boucher DM, Spatz S, Su MS, Kuida K: Erk5 null mice display multiple extraembryonic vascular and embryonic cardiovascular defects. Proc Natl Acad Sci U S A 2002, 99:9248-9253.

25. Hayashi M, Kim SW, Imanaka-Yoshida K, Yoshida T, Abel ED, Eliceiri B, Yang Y, Ulevitch RJ, Lee JD: Targeted deletion of BMK1/ERK5 in adult mice perturbs vascular integrity and leads to endothelial failure. J Clin Invest 2004, 113:1138-1148.

26. Conway RE, Petrovic N, Li Z, Heston W, Wu D, Shapiro LH: Prostate-specific membrane antigen regulates angiogenesis by modulating integrin signal transduction. Mol Cell Biol 2006, 26:5310-5324.

27. Demierre MF, Higgins PD, Gruber SB, Hawk E, Lippman SM: Statins and cancer prevention. Nat Rev Cancer 2005, 5:930-942.

28. Potente M, Ghaeni L, Baldessari D, Mostoslavsky R, Rossig L, Dequiedt F, Haendeler J, Mione M, Dejana E, Alt FW et al: SIRT1 controls endothelial angiogenic functions during vascular growth. Genes Dev 2007, 21:2644-2658.

29. Ljubimov AV, Caballero S, Aoki AM, Pinna LA, Grant MB, Castellon R: Involvement of protein kinase CK2 in angiogenesis and retinal neovascularization. Invest Ophthalmol Vis Sci 2004, 45:4583-4591.

30. Shaw GC, Cope JJ, Li L, Corson K, Hersey C, Ackermann GE, Gwynn B, Lambert AJ, Wingert RA, Traver D et al: Mitoferrin is essential for erythroid iron assimilation. Nature 2006, 440:96-100.

31. Sato Y, Hong HN, Yanai N, Obinata M: Involvement of stromal membrane-associated protein (SMAP-1) in erythropoietic microenvironment. J Biochem 1998, 124:209-216.

32. Melichar HJ, Narayan K, Der SD, Hiraoka Y, Gardiol N, Jeannet G, Held W, Chambers CA, Kang J: Regulation of gammadelta versus alphabeta T lymphocyte differentiation by the transcription factor SOX13. Science 2007, 315:230-233.

33. Roose J, Korver W, Oving E, Wilson A, Wagenaar G, Markman M, Lamers W, Clevers H: High expression of the HMG box factor sox-13 in arterial walls during embryonic development. Nucleic Acids Res 1998, 26:469-476.

34. Wallingford JB: Planar cell polarity, ciliogenesis and neural tube defects. Hum Mol Genet 2006, 15 Spec No 2:R227-234.

35. Botto LD, Moore CA, Khoury MJ, Erickson JD: Neural-tube defects. N Engl J Med 1999, 341:1509-1519.

36. Hayes JM, Kim SK, Abitua PB, Park TJ, Herrington ER, Kitayama A, Grow MW, Ueno N, Wallingford JB: Identification of novel ciliogenesis factors using a new in vivo model for mucociliary epithelial development. Dev Biol 2007, 312:115-130.

37. Wallingford JB: Neural tube closure and neural tube defects: Studies in animal models reveal known knowns and known unknowns. American Journal of Medical Genetics 2005, 135C:59-68.

38. Needleman SB, Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970, 48:443-453.

39. Smith TF, Waterman MS: Identification of common molecular subsequences. J Mol Biol 1981, 147:195-197.

40. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215:403-410.

41. Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Rindone WP, Swindell CD, Tung CS: The GenBank genetic sequence databank. Nucleic Acids Res 1986, 14:1-4.

42. McGary KL, Park TJ, Srimoyee G, Wallingford JB, Marcotte EM: Systematic discovery of non-obvious human disease models through orthologous phenotypes (In review) 2008.

Chapter 5: Putting the pieces together

Widespread genome sequencing has provided a substantial list of the parts that make up living organisms. In some cases, we already know where those pieces fit together, in other cases, there is still much work to be done; particularly to work out how molecular sequences and their defects lead to specific, macroscopic morphologies and phenotypes. If we approach biology like a jigsaw puzzle, we can see that genome sequencing has defined many of the pieces of the puzzle and basic research has already found many of the edge pieces that frame the work that remains to be done. However, much labor will be necessary to place the remaining pieces and to understand how individual processes work together as a whole. It is my hope that the predictive approaches that have been developed here will contribute to associating poorly understood genes with their correct biological context, which will facilitate targeted research that addresses the mechanistic, molecular level details and yet play a role in connect the details to a unified biological understanding that bridges the various biological processes and functions, both within a given organism and, eventually, across the diversity of life.

To that end, I have developed: a tool for understanding cellular level events (chapter 2), a method for associating genes with organismal phenotypes (chapter 3), and, finally, a framework that leverages a comparative approach to learn more about the unity and diversity of the mechanisms that function in many species (chapter 4). Using these tools is analogous to sorting pieces of a puzzle according to their color and texture prior to assembling them. My disappointing attempt to predict the export adaptor of the small subunit of the ribosome suggests that our current tools may not yet be able to predict the specific mechanisms by which the biological puzzle will fit together. However, my

subsequent papers has shown three different domains in which current knowledge can be leveraged to predict genes associated with a number of biological processes. The adoption of these tools by other researchers has the potential to further accelerate the pace of the collection of biological knowledge by prioritizing candidate genes in intuitively understandable, broadly applicable, quantitative ways that illuminate their relevance to individual biological processes of interest.

In particular, I hope that the phenolog approach and supporting web application will provide a very practical tool that helps reinforce the important contributions that can be made by integrating comparative methods in the normal work flow of molecular and cellular biology. By providing a pragmatic motivation for the collection of gene-phenotype data across many taxa, I hope that the collected data will eventually facilitate comparative studies across a range of spatial and temporal scales (e.g. developmental, behavioral, physiological, and metabolic) and grant us a better understanding of how the structure of biological mechanisms permit organisms to adapt to a changing world.

References

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Amberger, J., C. A. Bocchini, et al. (2008). "McKusick's Online Mendelian Inheritance in Man (OMIM(R))." Nucleic Acids Res.

Aouida, M., N. Page, et al. (2004). "A genome-wide screen in Saccharomyces cerevisiae reveals altered transport as a mechanism of resistance to the anticancer drug bleomycin." Cancer Res **64**(3): 1102-9.

Askree, S. H., T. Yehuda, et al. (2004). "A genome-wide screen for Saccharomyces cerevisiae deletion mutants that affect telomere length." Proc Natl Acad Sci U S A **101**(23): 8658-63.

Bennett, C. B., L. K. Lewis, et al. (2001). "Genes required for ionizing radiation resistance in yeast." Nat Genet **29**(4): 426-34.

Benson, D. A., I. Karsch-Mizrachi, et al. (2008). "GenBank." Nucleic Acids Res **36**(Database issue): D25-30.

Berglund, A. C., E. Sjolund, et al. (2008). "InParanoid 6: eukaryotic ortholog clusters with inparalogs." Nucleic Acids Res **36**(Database issue): D263-6.

Bilofsky, H. S., C. Burks, et al. (1986). "The GenBank genetic sequence databank." Nucleic Acids Res **14**(1): 1-4.

Birrell, G. W., G. Giaever, et al. (2001). "A genome-wide screen in Saccharomyces cerevisiae for genes affecting UV radiation sensitivity." Proc Natl Acad Sci U S A **98**(22): 12608-13.

Blackburn, A. S. and S. V. Avery (2003). "Genome-wide screening of Saccharomyces cerevisiae to identify genes required for antibiotic insusceptibility of eukaryotes." Antimicrob Agents Chemother **47**(2): 676-81.

Bonangelino, C. J., E. M. Chavez, et al. (2002). "Genomic screen for vacuolar protein sorting genes in Saccharomyces cerevisiae." Mol Biol Cell **13**(7): 2486-501.

Botto, L. D., C. A. Moore, et al. (1999). "Neural-tube defects." N Engl J Med **341**(20): 1509-19.

Bowers, P. M., M. Pellegrini, et al. (2004). "Prolinks: a database of protein functional linkages derived from coevolution." Genome Biol **5**(5): R35.

Chan, T. F., J. Carvalho, et al. (2000). "A chemical genomics approach toward understanding the global functions of the target of rapamycin protein (TOR)." Proc Natl Acad Sci U S A **97**(24): 13227-32.

Chang, M., M. Bellaoui, et al. (2002). "A genome-wide screen for methyl methanesulfonate-sensitive mutants reveals genes required for S phase progression in the presence of DNA damage." Proc Natl Acad Sci U S A **99**(26): 16934-9.

Chen, N., T. W. Harris, et al. (2005). "WormBase: a comprehensive data resource for Caenorhabditis biology and genomics." Nucleic Acids Res **33**(Database issue): D383-9.

Cherry, J. M., C. Adler, et al. (1998). "SGD: Saccharomyces Genome Database." Nucleic Acids Res **26**(1): 73-9.

Collins, S. R., P. Kemmeren, et al. (2007). "Towards a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae." Mol Cell Proteomics.

Combs, D. J., R. J. Nagel, et al. (2006). "Prp43p is a DEAH-box spliceosome disassembly factor essential for ribosome biogenesis." Mol Cell Biol **26**(2): 523-34.

Conway, R. E., N. Petrovic, et al. (2006). "Prostate-specific membrane antigen regulates angiogenesis by modulating integrin signal transduction." Mol Cell Biol **26**(14): 5310-24.

Darwin, C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London, John Murray.

Deane, C. M., L. Salwinski, et al. (2002). "Protein interactions: Two methods for assessment of the reliability of high-throughput observations." Mol Cell Proteomics: M100037-MCP200.

Demierre, M. F., P. D. Higgins, et al. (2005). "Statins and cancer prevention." Nat Rev Cancer **5**(12): 930-42.

Dennis, G., Jr., B. T. Sherman, et al. (2003). "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol **4**(5): P3.

Desmoucelles, C., B. Pinson, et al. (2002). "Screening the yeast "disruptome" for mutants affecting resistance to the immunosuppressive drug, mycophenolic acid." J Biol Chem **277**(30): 27036-44.

Deutschbauer, A. M., D. F. Jaramillo, et al. (2005). "Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast." Genetics **169**(4): 1915-25.

Deutschbauer, A. M., R. M. Williams, et al. (2002). "Parallel phenotypic analysis of sporulation and postgermination growth in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **99**(24): 15530-5.

Dezso, Z., Z. N. Oltvai, et al. (2003). "Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae." Genome Res **13**(11): 2450-4.

Downward, J. (2004). "Use of RNA interference libraries to investigate oncogenic signalling in mammalian cells." Oncogene **23**(51): 8376-83.

Dryja, T. P., W. Cavenee, et al. (1984). "Homozygosity of chromosome 13 in retinoblastoma." N Engl J Med **310**(9): 550-3.

Dwight, S. S., M. A. Harris, et al. (2002). "Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO)." Nucleic Acids Res **30**(1): 69-72.

Edmonds, D., B. J. Breitkreutz, et al. (2004). "A genome-wide telomere screen in yeast: the long and short of it all." Proc Natl Acad Sci U S A **101**(26): 9515-6.

Eisenberg, D., E. M. Marcotte, et al. (2000). "Protein function in the post-genomic era." Nature **405**(6788): 823-6.

Enyenihi, A. H. and W. S. Saunders (2003). "Large-scale functional genomic analysis of sporulation and meiosis in Saccharomyces cerevisiae." Genetics **163**(1): 47-54.

Eppig, J. T., J. A. Blake, et al. (2007). "The mouse genome database (MGD): new features facilitating a model system." Nucleic Acids Res **35**(Database issue): D630-7.

Eppig, J. T., C. J. Bult, et al. (2005). "The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology." Nucleic Acids Res **33**(Database issue): D471-5.

Feleszko, W., E. Z. Balkowiec, et al. (1999). "Lovastatin and tumor necrosis factor-alpha exhibit potentiated antitumor effects against Ha-ras-transformed murine tumor via inhibition of tumor-induced angiogenesis." Int J Cancer **81**(4): 560-7.

Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." Syst Zool **19**(2): 99-113.

Fleischer, T. C., C. M. Weaver, et al. (2006). "Systematic identification and functional screens of uncharacterized proteins associated with eukaryotic ribosomal complexes." Genes Dev **20**(10): 1294-307.

Fleming, J. A., E. S. Lightcap, et al. (2002). "Complementary whole-genome technologies reveal the cellular response to proteasome inhibition by PS-341." Proc Natl Acad Sci U S A **99**(3): 1461-6.

Franke, L., H. Bakel, et al. (2006). "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes." Am J Hum Genet **78**(6): 1011-25.

Fraser, A. G., R. S. Kamath, et al. (2000). "Functional genomic analysis of C. elegans chromosome I by systematic RNA interference." Nature **408**(6810): 325-30.

Fraser, A. G. and E. M. Marcotte (2004). "Development through the eyes of functional genomics." Curr Opin Genet Dev **14**(4): 336-42.

Gandhi, T. K., J. Zhong, et al. (2006). "Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets." Nat Genet **38**(3): 285-93.

Gelperin, D. M., M. A. White, et al. (2005). "Biochemical and genetic analysis of the yeast proteome with a movable ORF collection." Genes Dev **19**(23): 2816-26.

Ghaemmaghami, S., W. K. Huh, et al. (2003). "Global analysis of protein expression in yeast." Nature **425**(6959): 737-41.

Giaever, G., A. M. Chu, et al. (2002). "Functional profiling of the Saccharomyces cerevisiae genome." Nature **418**(6896): 387-91.

Giot, L., J. S. Bader, et al. (2003). "A protein interaction map of Drosophila melanogaster." Science **302**(5651): 1727-36.

Griffith, J. L., L. E. Coleman, et al. (2003). "Functional genomics reveals relationships between the retrovirus-like Ty1 element and its host Saccharomyces cerevisiae." Genetics **164**(3): 867-79.

Guldener, U., M. Munsterkotter, et al. (2006). "MPact: the MIPS protein interaction resource on yeast." Nucleic Acids Res **34**(Database issue): D436-41.

Harland, R. M. (1991). "In situ hybridization: an improved whole-mount method for Xenopus embryos." Methods Cell Biol **36**: 685-95.

Hart, G. T., I. Lee, et al. (2007). "A high-accuracy map of yeast protein complexes reveals modular basis of gene essentiality." BMC Bioinformatics **8**: 236.

Hart, G. T., A. K. Ramani, et al. (2006). "How complete are current yeast and human protein-interaction networks?" Genome Biol **7**(11): 120.

Hastings, P. J., S. K. Quah, et al. (1976). "Spontaneous mutation by mutagenic repair of spontaneous lesions in DNA." Nature **264**(5588): 719-22.

Hayashi, M., S. W. Kim, et al. (2004). "Targeted deletion of BMK1/ERK5 in adult mice perturbs vascular integrity and leads to endothelial failure." J Clin Invest **113**(8): 1138-48.

Hayes, J. M., S. K. Kim, et al. (2007). "Identification of novel ciliogenesis factors using a new in vivo model for mucociliary epithelial development." Dev Biol **312**(1): 115-30.

Hillenmeyer, M. E., E. Fung, et al. (2008). "The chemical genomic portrait of yeast: uncovering a phenotype for all genes." Science **320**(5874): 362-5.

Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nat Rev Genet **6**(2): 95-108.

Ho, J. H., G. Kallstrom, et al. (2000). "Nmd3p is a Crm1p-dependent adapter protein for nuclear export of the large ribosomal subunit." J Cell Biol **151**(5): 1057-66.

Hodgkin, J., H. R. Horvitz, et al. (1979). "Nondisjunction Mutants of the Nematode CAENORHABDITIS ELEGANS." Genetics **91**(1): 67-94.

Hu, Z., P. J. Killion, et al. (2007). "Genetic reconstruction of a functional transcriptional regulatory network." Nat Genet **39**(5): 683-7.

Huang, M. E., A. G. Rio, et al. (2003). "A genomewide screen in Saccharomyces cerevisiae for genes that suppress the accumulation of mutations." Proc Natl Acad Sci U S A **100**(20): 11529-34.

Huang, R. Y., M. Eddy, et al. (2005). "Genome-wide screen identifies genes whose inactivation confer resistance to cisplatin in Saccharomyces cerevisiae." Cancer Res **65**(13): 5890-7.

Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-91.

Huynen, M., B. Snel, et al. (2000). "Predicting protein function by genomic context: quantitative evaluation and qualitative inferences." Genome Res **10**(8): 1204-10.

Jeong, H., S. P. Mason, et al. (2001). "Lethality and centrality in protein networks." Nature **411**(6833): 41-2.

John Moult, J. T. P. R. J. K. F. (1995). "A large-scale experiment to assess protein structure prediction methods." Proteins: Structure, Function, and Genetics **23**(3): ii-iv.

Jorgensen, P., J. L. Nishikawa, et al. (2002). "Systematic identification of pathways that couple cell growth and division in yeast." Science **297**(5580): 395-400.

Joshi-Tope, G., M. Gillespie, et al. (2005). "Reactome: a knowledgebase of biological pathways." Nucleic Acids Res **33**(Database issue): D428-32.

Kamath, R. S., A. G. Fraser, et al. (2003). "Systematic functional analysis of the Caenorhabditis elegans genome using RNAi." Nature **421**(6920): 231-7.

Karaoz, U., T. M. Murali, et al. (2004). "Whole-genome annotation by using evidence integration in functional-linkage networks." Proc Natl Acad Sci U S A **101**(9): 2888-93.

Kelley, R. and T. Ideker (2005). "Systematic interpretation of genetic interactions using protein networks." Nat Biotechnol **23**(5): 561-6.

Kim, H., K. Melen, et al. (2006). "A global topology map of the Saccharomyces cerevisiae membrane proteome." Proc Natl Acad Sci U S A **103**(30): 11142-7.

King, O. D., J. C. Lee, et al. (2003). "Predicting phenotype from patterns of annotation." Bioinformatics **19 Suppl 1**: i183-9.

Kumar, A., K. H. Cheung, et al. (2000). "TRIPLES: a database of gene function in Saccharomyces cerevisiae." Nucleic Acids Res **28**(1): 81-4.

la Cour, T., L. Kiemer, et al. (2004). "Analysis and prediction of leucine-rich nuclear export signals." Protein Eng Des Sel **17**(6): 527-36.

Lage, K., E. O. Karlberg, et al. (2007). "A human phenome-interactome network of protein complexes implicated in genetic disorders." Nat Biotechnol **25**(3): 309-16.

Lee, I. (2008). "Personal Communication."

Lee, I., S. V. Date, et al. (2004). "A probabilistic functional network of yeast genes." Science **306**(5701): 1555-8.

Lee, I., B. Lehner, et al. (2008). "A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans." Nat Genet **40**(2): 181-8.

Lee, I., Z. Li, et al. (2007). "An Improved, Bias-Reduced Probabilistic Functional Gene Network of Baker's Yeast, Saccharomyces cerevisiae." PLoS ONE **2**(10): e988.

Lee, I., Z. Li, et al. (2007). "An Improved, Bias-Reduced Probabilistic Functional Gene Network of the Baker's Yeast, Sacchromyces cerevisiae." PLOS One.

Lehner, B. and A. G. Fraser (2004). "A first-draft human protein-interaction map." Genome Biol **5**(9): R63.

Lesage, G., J. Shapiro, et al. (2005). "An interactional network of genes involved in chitin synthesis in Saccharomyces cerevisiae." BMC Genet **6**(1): 8.

Lesuisse, E., S. A. Knight, et al. (2005). "Genome-wide screen for genes with effects on distinct iron uptake activities in Saccharomyces cerevisiae." Genetics **169**(1): 107-22.

Li, S., C. M. Armstrong, et al. (2004). "A map of the interactome network of the metazoan C. elegans." Science **303**(5657): 540-3.

Liu, E. T. (2005). "Systems biology, integrative biology, predictive biology." Cell **121**(4): 505-6.

Ljubimov, A. V., S. Caballero, et al. (2004). "Involvement of protein kinase CK2 in angiogenesis and retinal neovascularization." Invest Ophthalmol Vis Sci **45**(12): 4583-91.

Lu, X. and H. R. Horvitz (1998). "lin-35 and lin-53, two genes that antagonize a C. elegans Ras pathway, encode proteins similar to Rb and its binding protein RbAp48." Cell **95**(7): 981-91.

Madden, K. and M. Snyder (1998). "Cell polarity and morphogenesis in budding yeast." Annu Rev Microbiol **52**: 687-744.

Marcotte, E. M., M. Pellegrini, et al. (1999). "A combined algorithm for genome-wide prediction of protein function." Nature **402**(6757): 83-6.

Markovich, S., A. Yekutiel, et al. (2004). "Genomic approach to identification of mutations affecting caspofungin susceptibility in Saccharomyces cerevisiae." Antimicrob Agents Chemother **48**(10): 3871-6.

McGary, K. L., I. Lee, et al. (2007). "Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes." Genome Biol **8**(12): R258.

McGary, K. L., T. J. Park, et al. (2008). "Systematic discovery of non-obvious human disease models through orthologous phenotypes " (In review).

Melichar, H. J., K. Narayan, et al. (2007). "Regulation of gammadelta versus alphabeta T lymphocyte differentiation by the transcription factor SOX13." Science **315**(5809): 230-3.

Mellor, J. C., I. Yanai, et al. (2002). "Predictome: a database of putative functional links between proteins." Nucleic Acids Res **30**(1): 306-9.

Mewes, H. W., D. Frishman, et al. (2006). "MIPS: analysis and annotation of proteins from whole genomes in 2005." Nucleic Acids Res **34**(Database issue): D169-72.

Mnaimneh, S., A. P. Davierwala, et al. (2004). "Exploration of essential gene functions via titratable promoter alleles." Cell **118**(1): 31-44.

Moffatt, M. F., M. Kabesch, et al. (2007). "Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma." Nature **448**(7152): 470-3.

Muller, H. A. (2000). "Genetic control of epithelial cell polarity: lessons from Drosophila." Dev Dyn **218**(1): 52-67.

Myers, C. L., D. Robson, et al. (2005). "Discovery of biological networks from diverse functional genomic data." Genome Biol **6**(13): R114.

Narayanaswamy, R., E. Moradi, et al. (2008). "Systematic definition of protein constituents along the major polarization axis reveals an adaptive re-use of the polarization machinery in pheromone treated budding yeast." Journal of Proteome Research.

Narayanaswamy, R., W. Niu, et al. (2006). "Systematic profiling of cellular phenotypes with spotted cell microarrays reveals mating-pheromone response genes." Genome Biol **7**(1): R6.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-53.

Ni, L. and M. Snyder (2001). "A genomic study of the bipolar bud site selection pattern in Saccharomyces cerevisiae." Mol Biol Cell **12**(7): 2147-70.

Ohya, Y., J. Sese, et al. (2005). "High-dimensional and large-scale phenotyping of yeast mutants." Proc Natl Acad Sci U S A **102**(52): 19015-20.

OMIM Online Mendelian Inheritance in Man (OMIM), McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

Oti, M., B. Snel, et al. (2006). "Predicting disease genes using protein-protein interactions." J Med Genet **43**(8): 691-8.

Owen, R. (1843). Lectures on Comparative Anatomy and Physiology of the Invertebrate Animals. London, Longmans, Brown, Green and Longmans.

Page, N., M. Gerard-Vincent, et al. (2003). "A Saccharomyces cerevisiae genome-wide mutant screen for altered sensitivity to K1 killer toxin." Genetics **163**(3): 875-94.

Pellegrini, M., E. M. Marcotte, et al. (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc Natl Acad Sci U S A **96**(8): 4285-8.

Potente, M., L. Ghaeni, et al. (2007). "SIRT1 controls endothelial angiogenic functions during vascular growth." Genes Dev **21**(20): 2644-58.

Proszynski, T. J., R. Klemm, et al. (2006). "Plasma membrane polarization during mating in yeast cells." J Cell Biol **173**(6): 861-6.

Ramani, A. K., R. C. Bunescu, et al. (2005). "Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome." Genome Biol **6**(5): R40.

Reagan, M. S., C. Pittenger, et al. (1995). "Characterization of a mutant strain of Saccharomyces cerevisiae with a deletion of the RAD27 gene, a structural homolog of the RAD2 nucleotide excision repair gene." J Bacteriol **177**(2): 364-71.

Regan, C. P., W. Li, et al. (2002). "Erk5 null mice display multiple extraembryonic vascular and embryonic cardiovascular defects." Proc Natl Acad Sci U S A **99**(14): 9248-53.

Reimand, J., M. Kull, et al. (2007). "g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments." Nucleic Acids Res **35**(Web Server issue): W193-200.

Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-52.

Rhodes, D. R., S. A. Tomlins, et al. (2005). "Probabilistic model of the human protein-protein interaction network." Nat Biotechnol **23**(8): 951-9.

Richardson, A. L., Z. C. Wang, et al. (2006). "X chromosomal abnormalities in basal-like human breast cancer." Cancer Cell **9**(2): 121-32.

Riles, L., R. J. Shaw, et al. (2004). "Large-scale screening of yeast mutants for sensitivity to the IMP dehydrogenase inhibitor 6-azauracil." Yeast **21**(3): 241-8.

Robinson, M. D., J. Grigull, et al. (2002). "FunSpec: a web-based cluster interpreter for yeast." BMC Bioinformatics **3**(1): 35.

Roose, J., W. Korver, et al. (1998). "High expression of the HMG box factor sox-13 in arterial walls during embryonic development." Nucleic Acids Res **26**(2): 469-76.

Ross, A. J., H. May-Simera, et al. (2005). "Disruption of Bardet-Biedl syndrome ciliary proteins perturbs planar cell polarity in vertebrates." Nat Genet **37**(10): 1135-40.

Rual, J. F., K. Venkatesan, et al. (2005). "Towards a proteome-scale map of the human protein-protein interaction network." Nature **437**(7062): 1173-8.

Saito, T. L., M. Ohtani, et al. (2004). "SCMD: Saccharomyces cerevisiae Morphological Database." Nucleic Acids Res **32**(Database issue): D319-22.

Sato, Y., H. N. Hong, et al. (1998). "Involvement of stromal membrane-associated protein (SMAP-1) in erythropoietic microenvironment." J Biochem **124**(1): 209-16.

Scanlan, M. J., I. Gout, et al. (2001). "Humoral immunity to human breast cancer: antigen definition and quantitative analysis of mRNA expression." Cancer Immun **1**: 4.

Scherens, B. and A. Goffeau (2004). "The uses of genome-wide yeast mutant collections." Genome Biol **5**(7): 229.

Schuldiner, M., S. R. Collins, et al. (2005). "Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile." Cell **123**(3): 507-19.

Schwikowski, B., P. Uetz, et al. (2000). "A network of protein-protein interactions in yeast." Nat Biotechnol **18**(12): 1257-61.

Scott, M. S., S. J. Calafell, et al. (2005). "Refining protein subcellular localization." PLoS Comput Biol **1**(6): e66.

Shannon, P., A. Markiel, et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-504.

Sharan, R., I. Ulitsky, et al. (2007). "Network-based prediction of protein function." Mol Syst Biol **3**: 88.

Shaw, G. C., J. J. Cope, et al. (2006). "Mitoferrin is essential for erythroid iron assimilation." Nature **440**(7080): 96-100.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-7.

Sopko, R., D. Huang, et al. (2006). "Mapping pathways and phenotypes by systematic gene overexpression." Mol Cell **21**(3): 319-30.

Steinmetz, L. M., C. Scharfe, et al. (2002). "Systematic screen for human disease genes in yeast." Nat Genet **31**(4): 400-4.

Stelzl, U., U. Worm, et al. (2005). "A human protein-protein interaction network: a resource for annotating the proteome." Cell **122**(6): 957-68.

Troyanskaya, O. G., K. Dolinski, et al. (2003). "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)." Proc Natl Acad Sci U S A **100**(14): 8348-53.

Vanrobays, E., A. Leplus, et al. (2008). "TOR regulates the subcellular distribution of DIM2, a KH domain protein required for cotranscriptional ribosome assembly and pre-40S ribosome export." Rna **14**(10): 2061-73.

von Dassow, G., E. Meir, et al. (2000). "The segment polarity network is a robust developmental module." Nature **406**(6792): 188-192.

von Mering, C., M. Huynen, et al. (2003). "STRING: a database of predicted functional associations between proteins." Nucleic Acids Res **31**(1): 258-61.

Walker, M. G., W. Volkmuth, et al. (1999). "Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes." Genome Res **9**(12): 1198-203.

Wallingford, J. B. (2005). "Neural tube closure and neural tube defects: Studies in animal models reveal known knowns and known unknowns." American Journal of Medical Genetics **135C**(1): 59-68.

Wallingford, J. B. (2006). "Planar cell polarity, ciliogenesis and neural tube defects." Hum Mol Genet **15 Spec No 2**: R227-34.

Warringer, J., E. Ericson, et al. (2003). "High-resolution yeast phenomics resolves different physiological features in the saline response." Proc Natl Acad Sci U S A **100**(26): 15724-9.

Willer, M., M. Regnacq, et al. (2000). "Disruption and functional analysis of six ORFs on chromosome XII of saccharomyces cerevisiae: YLR124w, YLR125w, YLR126c, YLR127c, YLR128w and YLR129w." Yeast **16**(15): 1429-35.

Wilson, W. A., Z. Wang, et al. (2002). "Systematic identification of the genes affecting glycogen storage in the yeast Saccharomyces cerevisiae: implication of the vacuole as a determinant of glycogen level." Mol Cell Proteomics **1**(3): 232-42.

Winzeler, E. A., D. D. Shoemaker, et al. (1999). "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis." Science **285**(5429): 901-6.

Xenarios, I., L. Salwinski, et al. (2002). "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions." Nucleic Acids Res **30**(1): 303-5.

Xie, M. W., F. Jin, et al. (2005). "Insights into TOR function and rapamycin response: chemical genomic profiling by using a high-density cell array method." Proc Natl Acad Sci U S A **102**(20): 7215-20.

Zewail, A., M. W. Xie, et al. (2003). "Novel functions of the phosphatidylinositol metabolic pathway discovered by a chemical genomics screen with wortmannin." Proc Natl Acad Sci U S A **100**(6): 3345-50.

Zhang, J., C. Schneider, et al. (2002). "Genomic scale mutant hunt identifies cell size homeostasis genes in S. cerevisiae." Curr Biol **12**(23): 1992-2001.

Vita

Kriston McGary was born in Alton, Illinois, September 15, 1977 to Elton and Lynn McGary. He received his bachelor's of science from Bryan College in biology (2000), where he served as undergraduate teacher's assistant for several semesters. After graduation, he worked as fellow of the Oak Ridge Institute of Science and Education (ORISE) at the United States Army Medical Research Institute for Chemical Defense. He began graduate studies at the University of Texas at Austin in 2002.

PUBLICATIONS

Dillman, J. F., 3rd, K. L. McGary, et al. (2003). "Sulfur mustard induces the formation of keratin aggregates in human epidermal keratinocytes." Toxicol Appl Pharmacol **193**(2): 228-36.

Dillman, J. F., 3rd, K. L. McGary, et al. (2004). "An inhibitor of p38 MAP kinase downregulates cytokine release induced by sulfur mustard exposure in human epidermal keratinocytes." Toxicol In Vitro **18**(5): 593-9.

McGary, K. L., I. Lee, et al. (2007). "Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes." Genome Biol **8**(12): R258.

McGary, K. L., T. J. Park, et al. (2008). "Systematic discovery of non-obvious human disease models through orthologous phenotypes " (In review).

Narayanaswamy, R., E. Moradi, et al. (2008). "Systematic definition of protein constituents along the major polarization axis reveals an adaptive re-use of the polarization machinery in pheromone treated budding yeast." Journal of Proteome Research. In press.

Permanent address: 12820 N Lamar Blvd., Austin, TX 78753

This dissertation was typed by the author.