Title
Assembling, analyzing, refining, and cataloging molecular interaction network

Permalink
https://escholarship.org/uc/item/0hf2t2fp

Author
Mak, Huajiang Craig

Publication Date
2008

Peer reviewed|Thesis/dissertation
ASSEMBLING, ANALYZING, REFINING, AND CATALOGING
MOLECULAR INTERACTION NETWORKS

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

in

Biology

by

Huajiang Craig Mak

Committee in charge:

Professor Trey Ideker, Chair
Professor Joseph Ecker
Professor Amy Kiger
Professor Richard Kolodner
Professor William Loomis

2008
The Dissertation of Huajiang Craig Mak is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

___________________________________________________________
___________________________________________________________
___________________________________________________________
___________________________________________________________
___________________________________________________________

Chair

University of California, San Diego

2008
I stepped from Plank to Plank
A slow and cautious way
The Stars about my Head I felt
About my Feet the Sea.

I knew not but the next
Would be my final inch—
This gave me the precarious Gait
Some call Experience

*Emily Dickinson*
TABLE OF CONTENTS

SIGNATURE PAGE ................................................................. iii
EPIGRAPH ............................................................................... iv
TABLE OF CONTENTS ............................................................. v
LIST OF FIGURES ..................................................................... x
LIST OF TABLES ....................................................................... xii
ACKNOWLEDGEMENTS .......................................................... xiii
VITA .......................................................................................... xv
ABSTRACT OF THE DISSERTATION ........................................... xvi

Chapter 1: Introduction

Networks: the output of high-throughput experiments ....................... 1
The Network Lifecycle, Part I: Assemble and Analyze ......................... 2
The Network Lifecycle, Part II: Refine and Catalog ............................. 5
Contributions of this dissertation ..................................................... 8

(continued next page)
Chapter 2: Assembling a DNA damage response network

Abstract .................................................................................................................................11
Introduction ..........................................................................................................................12
Results ..................................................................................................................................13
Identifying transcription factors involved in the MMS response ........................................13
Measuring the MMS-induced transcriptional network .........................................................14
Deletion-buffering analysis .................................................................................................18
Building a network of regulatory pathways .........................................................................20
Experimentally testing the model ........................................................................................23
Methods ...............................................................................................................................25
Strains and Media ................................................................................................................25
Phenotypic Sensitivity Assays ...............................................................................................25
Whole Genome Expression Analysis ....................................................................................26
Genome Wide Transcription Factor Binding Analysis .........................................................26
Array Preparation and Hybridization ..................................................................................27
Data Post-processing and Significance Assessment ..............................................................27
Overlap in Binding Between Conditions ............................................................................28
Significance of expansion or contraction ..............................................................................29
Motif Enrichment in Transcription Factor Binding Data .....................................................30
Deletion Buffering Analysis .................................................................................................31
Assembly of Transcriptional Pathways From Integrated Data .............................................32
Assessing overlaps in binding between TF pairs .................................................................33
Tables ....................................................................................................................................35
Supplemental Figures ............................................................................................................36
Acknowledgements ..............................................................................................................44

(continued next page)
Chapter 3: Analyses of position effects within the yeast transcriptional network

Abstract ...............................................................................................................................45
Introduction .........................................................................................................................46
Results .................................................................................................................................47

Discovery of a large family of Subtelomere Binding Transcription Factors .......................................47
SBTFs do not behave like known telomere-binding complexes or transcription factors but do interact with these proteins ..........................................................52
SBTFs regulate stress and carbon metabolism in three broad clusters .............................54
Six SBTFs target heterochromatin domains that are derepressed during growth on alternative carbon sources .................................................................55
Dynamic subtelomeric binding is correlated with gene expression ..................................58
Discussion ..........................................................................................................................60
Methods ...............................................................................................................................63

Computational screen for SBTFs .....................................................................................63
Analysis of protein complexes and interactions involving SBTFs ..................................64
Analysis of repetitive element binding by SBTFs ..............................................................64
Screening SBTF binding profiles for correlation with hda1Δ sensitivity .............................65
Dilution assays for growth of SBTF deletions on alternative carbon sources .........................66
Inferring SBTF modes of action by integrated analysis of matched binding and expression profiles .................................................................66
Identification of SBTF orthologs .......................................................................................67
Supplemental Figures ........................................................................................................69
Supplemental Tables .........................................................................................................74
Acknowledgements .........................................................................................................77
Chapter 4: Refinement of a transcriptional regulatory network using graphical probabilistic models

Abstract .................................................................................................................. 78
Introduction .......................................................................................................... 79
Results .................................................................................................................... 80
  Summary of physical regulatory models .......................................................... 80
  Experiment selection ......................................................................................... 83
  Model validation ............................................................................................... 85
  Automated model refinement .......................................................................... 87
  Learning curve analysis ................................................................................... 88
Discussion ............................................................................................................. 89
Conclusions ......................................................................................................... 91
Methods .............................................................................................................. 92
  Model building and inference ......................................................................... 92
  Experiment scoring ......................................................................................... 92
  Expression profiling ......................................................................................... 93
  Expression coherence ..................................................................................... 94
Tables .................................................................................................................. 96
Supplemental Tables .......................................................................................... 98
Acknowledgements .......................................................................................... 102

(continued next page)
Chapter 5: Cataloging network models in a searchable online database

Abstract ................................................................................................................................. 103
Introduction .......................................................................................................................... 104
Results ................................................................................................................................ 106
   A spectrum of network models ..................................................................................... 106
   Database coverage and assembly ................................................................................. 108
   Network model query ................................................................................................. 109
   Meta-analysis of models ............................................................................................. 111
Discussion ........................................................................................................................... 113
Methods .............................................................................................................................. 114
   Data processing ........................................................................................................... 114
   Web interface ............................................................................................................... 114
   Scoring models for Gene Ontology (GO) annotation ................................................. 115
   Scoring similarity between publications ................................................................... 115
Tables ................................................................................................................................ 116
Supplemental Tables ........................................................................................................... 117
Acknowledgements ............................................................................................................. 118

Chapter 6: Conclusions

Physiologically relevant interpretation of high-throughput data ............................. 119
Harnessing complementary relationships ................................................................. 120
Controlling for missing data and spurious data ......................................................... 122
Remaining grounded in biology ................................................................................... 123
Beyond wiring diagrams .............................................................................................. 125
Final remarks .................................................................................................................... 127

References ......................................................................................................................... 128
# LIST OF FIGURES

Figure 2.1. Overview of the systems approach.........................................................13
Figure 2.2. TF selection, chIP-chip, and expression profiling experiments. ...........15
Figure 2.3. Overlap in transcription factor binding patterns in response to MMS....17
Figure 2.4. Deletion buffering of the MMS response. ............................................20
Figure 2.5. A network of regulatory interactions connecting transcription factors to deletion-buffered genes. .................................................................22

Supplemental Figure 2.1. MMS sensitivity assay of transcription factor deletions strains. .............................................................................................................36
Supplemental Figure 2.2. Scoring overlap in TF binding between conditions. .......37
Supplemental Figure 2.3. Clustering of MMS expression responses with other stresses. .................................................................................................................38
Supplemental Figure 2.4. MMS responsive genes implicate G1/S checkpoint. ........39
Supplemental Figure 2.5. Crt1 deletion-buffering P-values....................................40
Supplemental Figure 2.6. Number of genes buffered by each transcription factor deletion (buffering p-value <0.005). .........................................................41
Supplemental Figure 2.7. Correspondence between expression and regulation of the RNR genes by Crt1p.................................................................42
Supplemental Figure 2.8. Direct binding interactions supported by deletion-buffering effects. .................................................................43

Figure 3.1. Transcription Factors that Preferentially Bind Sequences at Subtelomeres.................................................................48
Figure 3.2. Example SBTFs that Show Dynamic Binding Preferences as a Function of Growth or Stress Condition. ..........................50
Figure 3.3. Subtelomeric Binding Preference is Dynamic and Distributed into Distinct Clusters. ........................................................51
Figure 3.4. Interactions Between SBTFs and Telomere-related Genes. .................53
Figure 3.5. Binding Profiles, Expression, and Growth Phenotypes Link SBTFs to Hda1p.................................................................57
Figure 3.6. Dynamic Binding and Expression Profiles Suggest Models of SBTF Function and Link SBTFs to Poorly Characterized Genes at Subtelomeres. .................59

(continued next page)
Supplemental Figure 3.1. Estimating the error in identifying SBTFs. ..........................69
Supplemental Figure 3.2. A map of the subtelomeric regulatory circuitry. ....................70
Supplemental Figure 3.3. Analysis of protein complexes. .............................................71
Supplemental Figure 3.4. Screening expression profiles of single gene deletions for
enrichment of subtelomeric genes. .................................................................72
Supplemental Figure 3.5. Expression responses of subtelomeric genes to
environmental perturbations.................................................................73
Figure 4.1. Wiring diagrams for example network models. ............................................81
Figure 4.2. Schematic of the experimental design approach. .........................................84
Figure 4.3. Validation and refinement of Swi4 transcriptional cascades. .......................86
Figure 4.4. Simulated learning curves of three experimental design methods. ..............89
Figure 5.1. The need for a new type of database. .......................................................104
Figure 5.2. Representative network models stored in CellCircuits. .............................107
Figure 5.3. Web interface (http://www.cellcircuits.org). ...........................................110
Figure 5.4. Meta-analysis of models.........................................................................111
Figure 6.1. Contributions of the dissertation to each stage of the network lifecycle.119
LIST OF TABLES

Table 2.1. DNA sequence motifs found in promoters bound in only one condition.....35
Table 3.1. Summary of Subtelomere Binding Transcription Factors..........................68
Supplemental Table 3.1. List of telomere-related genes curated from the literature....74
Supplemental Table 3.2. Correlation of binding profiles with the presence of
    genomic features. .............................................................................75
Supplemental Table 3.3. Orthologs of SBTFs in other species. .......................76
Table 4.1. Internal validation for 21 of the 38 inferred models. .......................96
Table 4.2. Top-ranking knock-out experiments proposed for model discrimination. .97
Supplemental Table 4.1. Internal validation for 17 out of 24 restricted network
    models. ............................................................................................98
Supplemental Table 4.2. Correlations between swi4Δ and gcn4Δ data from
    Rosetta and the new experiments. ..............................................99
Supplemental Table 4.3. Restricted subsets used to evaluate the reproducibility.....100
Supplemental Table 4.4. Gene sets for external validation. .............................101
Table 5.1. Sources of data. ............................................................................116
Supplemental Table 5.1. Figures curated. ......................................................117
ACKNOWLEDGEMENTS

I thank Trey Ideker, Chris Workman, Scott McCuine, Chen-Hsiang Yeang, and Mike Daly for the opportunity to collaborate on this research. Your insight, advice, and work were critical and will continue to shape my thinking in the future. To my advisor, Trey, I owe the opportunity to work in the lab, continued guidance and advice, and an uncompromising standard of excellence.

I am grateful to Tim Ravasi, Kai Tan, Sourav Bandyopadhyay, Ryan Kelley, Dwight Kuo, and Bianca Gruebel for their discussions, comments, and contributions to my research. I would also like to thank Sandi Jacobson, Scott McCuine, Colin Luo, Kate Licon, and George Kassavites for experimental assistance. Sandi Jacobson and Lorraine Pillus welcomed me into their lab and shared their love of yeast, chromatin, and telomeres – Thank you. I am grateful to Joe Ecker, Amy Kiger, Richard Kolodner, and Bill Loomis for helpful advice and discussions. I also owe thanks to Geoff Rosenfeld and Valentina Perissi for an initial foray into the biology of nuclear receptors. That exposure has turned out to be invaluable. Finally, I thank everyone in the Ideker lab. Over the years, you all shaped my work for the better in one way or another.

Vasanthan Dasan, Tae Kim, and Ann Togasaki helped me get to where I am today and led by example along the way. Thank you.

My time in La Jolla was greatly enriched, and my sanity preserved, by friends who helped remind me about life outside of biology. I thank John, Hudson, Dan, Lauren, Dave Shear, Ambika, and Dave and Naomi Spinak.

I am indebted to my family, Nina, Mom, Dad, and T, for encouragement, patience, and an unwavering foundation from which to go forward. Last, but not least, I owe everything to Kiran for keeping it real. This is for all of you.
Chapter 2, in full, is a reprint of material as it appeared on May 19, 2006 in Science, 2006, vol. 312, p. 1054-1059. Workman, Christopher T.; Mak, H. Craig; McCuine, Scott; Tagne, Jean-Bosco; Agarwal, Maya; Ozier, Owen; Begley, Thomas J.; Samson, Leona D.; Ideker, Trey. The dissertation author was one of three primary investigators (with C.T. Workman and S. McCuine) and two primary authors (with C.T. Workman) of this paper.

Chapter 3, in full, has been submitted for publication as it may appear in Genome Research, 2008, Mak, H. Craig; Pillus, Lorraine; Ideker, Trey. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of material as it appeared on July 1, 2005 in Genome Biology, 2005, 6:R62. Yeang, Chen-Hsiang; Mak, H. Craig; McCuine, Scott; Workman, Christopher; Jaakkola, Tommi; Ideker, Trey. The dissertation author was one of two primary investigators and authors (with C.H. Yeang) of this paper.

Chapter 5, in full, is a reprint of material as it appeared on November 29, 2006 in Nucleic Acids Research, 2006, vol. 35, D538–D545, Database issue 2007. Mak, H. Craig; Daly, Mike; Gruebel, Bianca; Ideker, Trey. The dissertation author was one of two primary investigators and authors (with M. Daly) of this paper.
VITA

1999 Bachelor of Science, Massachusetts Institute of Technology
1999-2002 Member of Technical Staff, Sun Microsystems Inc.
2003-2005 Teaching Assistant, University of California, San Diego
2002-2008 Research Assistant, University of California, San Diego
2008 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

“Validation and refinement of gene regulatory pathways on a network of physical interactions” Genome Biology, Volume 6:R62, July 2005


“Dynamic reprogramming of transcription factors to and from the subtelomere” Genome Research, submitted, August 2008

FIELDS OF STUDY

Major Field: Biology (Computational and Systems Biology)

Studies in Computational and Systems Biology
Professor Trey Ideker
ABSTRACT OF THE DISSERTATION

ASSEMBLING, ANALYZING, REFINING, AND CATALOGING MOLECULAR INTERACTION NETWORKS

by

Huajiang Craig Mak

Doctor of Philosophy in Biology

University of California, San Diego, 2008

Professor Trey Ideker, Chair

Life within an organism is sustained by biomolecular interactions. Mapping networks of interactions and deciphering their functions – at both individual and global scales – is a long-standing challenge. Network analyses promise to illuminate how complex behaviors emerge from collections of individual interactions. Many disease states and physiological responses, for instance, emerge from the complex interconnections between biological pathways.
This dissertation addresses four challenges inherent in deriving biological meaning from networks. We studied transcriptional regulatory networks, although the insights gained are more widely applicable. First, we investigated methods for assembling regulatory networks by integrating physical and functional data. Second, we analyzed individual regulatory pathways and global properties of networks. Third, we investigated methods for efficient network validation and refinement. Fourth, we describe a database of network models: its value, its implementation, and its potential uses as a research tool.

We present the first large-scale network model of the transcriptional response to a DNA damaging agent in a eukaryotic cell. We developed novel methods for statistically assigning confidence scores to high-throughput data and for integrating physical and functional data to generate a network that is induced by an environmental perturbation. In a second study, we developed computational methods to discover and characterize a group of 22 yeast transcription factors that specifically bind to targets in the subtelomeric regions near the ends of chromosomes. In a third study, we used a measure of information gain to prioritize potential experiments for validating a network. We then performed three top-ranking experiments and used the results to refine the network. In a fourth study, we describe CellCircuits, an online, searchable database of network models containing over 1000 models from eleven published studies. This database was made available to the wider research community.

As a result of our work, we identified general guidelines and themes for working with molecular interaction networks. We conclude with remarks on grounding computational research in the realities of biology.
Chapter 1: Introduction

Networks: the output of high-throughput experiments

Imagine the molecules in a cell. Little machines that carry out life’s tasks: breaking, synthesizing, modifying, transporting. Replicating. Each has specific roles, but none work in isolation. Each has the potential to work with many others, but not at the same time or same place within the cell. If we could map the web of interactions between molecules, we would have a network: a record of chemical affinities, of what influences what, of how cells get things done.

Today we can, in fact, map networks, albeit to an incomplete but ever improving extent. Key advances in technology have enabled biological measurement on a massive, so called high-throughput, scale. Whereas the Southern blot has been used since 1975 to measure the amount of messenger RNA produced from a single gene, the development of the microarray in 1995 allowed us to measure in a single experiment the RNA produced from all genes. The measurement capacity of microarrays has been paired with other experimental protocols from molecular biology to identify the genomic locations bound by regulatory proteins, the targets of kinase enzymes, and the individual cells among a pool of a thousand variants that grow the fastest. Microarrays derive their power by performing many experiments in parallel. Parallelization is a powerful strategy that has also recently been applied to other experimental techniques including DNA sequencing, generating mutants with gene deletions, and mutant phenotyping.

The net results of these parallel experiments are millions of data points. Each experimental technique illuminates a different facet of life in a cell, possibly at different times, possibly under different environmental conditions, possibly with different components of the cell modified or disabled. Each experiment produces a
The challenges inherent in extracting meaningful information from networks are the central subjects of this dissertation. We conceptually group these into four stages: network assembly, network analysis, network refinement, and network cataloging. Together these four stages constitute the lifecycle of a network.

**The Network Lifecycle, Part I: Assemble and Analyze**

In assembling a network, data collection is the first step. At this point the network may contain an impressive amount of raw data, but it is not very useful. Among the thousands of genes and interactions represented by the network, only a handful may be involved in a biochemical pathway or disease being studied. Filtering becomes an important task. And as alluded to previously, the activities within a cell are multi-faceted, involving protein, DNA, RNA, and metabolite interactions. Some of which represent direct contact between molecules, while others represent the indirect effects observed when one or more molecules are disrupted. Integrating these different interactions becomes a second important task. Together, data collection, filtering, and integration are the challenges faced in network assembly.

Filtering is often simple. It is aimed at answering questions such as: Where is my favorite gene in the network? And what does it interact with? Software tools have been developed that make it easy to answer these simple questions\textsuperscript{10,11}. Online databases of molecular interactions also provide the ability to retrieve data for a specific gene or molecular pathway\textsuperscript{12-16}. Moving beyond single genes, networks have been analyzed for global properties – consistent organizing principles that illuminate, for instance, how a cell functions as a system, how it is robust to failures, how it adapts to the environment, or how one external signal can reprogram several internal
pathways\textsuperscript{17}.

For example, one property of a protein in a network is the number of other proteins with which it interacts. Most proteins have a small number of interacting partners, while fewer have many partners, forming a hub-and-spoke topology\textsuperscript{18-21}. The hubs have been interpreted as “master regulators,” proteins that sit at the top of a regulatory hierarchy, responsible for propagating one signal to many downstream targets in the network. To the spokes are ascribed more specialized roles: perhaps functioning in one tissue in the body or one biochemical pathway among many\textsuperscript{22}.

On top of this hub-and-spoke model other data can be layered. By layering the time during the cell cycle when each protein is expressed, for instance, we observe that some hub proteins act as gatekeepers, regulating progression through the cell cycle\textsuperscript{23-25}. Yet other hubs appear to interact with the same sets of partners, but only one hub is ever expressed in the cell at a given time, suggesting that their roles may be interchangeable, but not redundant\textsuperscript{19}. Layering additional information on to a network is an example of what is referred to as data integration.

Data integration is usually more complex than filtering. But what, exactly, does it involve? Integration. Layering. These are words borrowed from other disciplines – mathematics or stone masonry perhaps – to describe the biologist’s abstract aspirations, not actual actions. They describe what we feel must be possible, but do not know exactly how to do yet, otherwise we would have a more precise word. We think: new technology generates so much data; there must be a way to integrate it all to better understand a living organism. Somehow, someway we should be able to mush it all together, wringing out biological insight. But how? What way?

The methods used to date have been typically grounded in either statistical pattern finding or in a search for patterns directed by biologically inspired templates. Computer scientists use the terms ‘unsupervised’ versus ‘supervised’ to express
this distinction in methodology. Both methods have been successful. Statistical
pattern finding approaches use techniques from a branch of computer science known
as machine learning and have successfully identified small, repeated patterns of
interactions (so called network motifs), groups of genes and regulators involved
in different cancer types, and modules of highly interconnected genes that
respond coherently to external stimuli. Patterns directed by biologically inspired
templates have been used to search for gene expression signatures among large
collections of microarray data, evolutionarily conserved DNA regulatory motifs,
interconnectivity between molecular machines in protein interaction networks,
and pathways conserved across the networks of multiple species.

Two themes emerge from past efforts at data integration. First, data integration
often produces additional networks – either composite networks made of interactions
drawn from several data sources, focused networks made of a subset of genes, or a
network both composite and focused. Second, network assembly and analyses are
often interconnected activities. The act of integration is, itself, one of analysis – an
application of rules or heuristics that formalize the biologist’s insight about how one
data type is related to another.

In this dissertation we posit that data integration can be defined as the act of
formalizing these relationships. To integrate, to layer data onto the network, to mush
it all together is to articulate a deep understanding of how those data are related.
If we observe an interaction between proteins, what does that tell us about their
functional relationship? If two interacting proteins are expressed at different times or
at different locations in the cell, what does that imply? If deleting a regulatory protein
only affects the expression of its targets under specific environmental conditions, do
we expect that regulator to directly bind the promoters of those targets? The variations
on these questions are endless. Proposing answers, codifying, and then testing them
against the data collected is the methodology we will use to extract meaning from networks.

The biological focus of this dissertation will be on transcriptional regulatory networks. These networks represent interactions between proteins that regulate gene expression and their target genes. Regulatory proteins may include transcription factors (TFs), signaling molecules, chromatin proteins that package DNA, or the RNA polymerase holoenzyme. We articulate relationships between direct (physical binding) and indirect (functional dependence) interactions linking TFs to their targets (Chapters 2 and 4). We also articulate relationships between the physical location of a gene along the chromosome and TF-target binding interactions (Chapter 3). We describe computational methods that codify these relationships and then apply these methods to study transcriptional responses to genetic perturbations (Chapter 4), responses to DNA damaging chemicals (Chapter 2), and properties related to the organization of genes along chromosomes (Chapter 3).

The Network Lifecycle, Part II: Refine and Catalog

Next, we turn to the challenges of working with a network after it has been assembled and perhaps analyzed. Among the questions the biologist might ask when presented with a network: What is new and what is already known? How do the network results compare with previous findings? Why is a well-known observation not represented in the network? What conclusions can be drawn from this network about the function of my favorite gene?

As a result of exploring these questions, the network may need to be refined, either by adding new data, revisiting analysis methods, or throwing it all out and starting over. In many cases, the network may be known to be incomplete. It is widely recognized, for example, that networks assembled using data from high-
throughput experiments often contain many false negatives (missing interactions) and false positives (spurious interactions)\textsuperscript{34}. Alternatively, a network may be interpreted in more than one way that is consistent with the observed data\textsuperscript{35}. In all of the aforementioned scenarios we are interested in the best ways to refine the network – either by adding missing interactions or pruning erroneous ones. Network refinement is an open problem that is often overlooked in the rush to gather, assemble, and analyze.

Networks can be refined by quantitatively scoring the accuracy of individual interactions. Low scoring interactions may be discarded, leaving only the proteins and interactions judged most likely to be accurate. Krogan and Greenblatt identified a high-confidence set of protein interactions using a decision trees and Bayesian statistics – tools from a branch of computer science known as machine learning\textsuperscript{36}. Networks have also been compared to each other and to “gold standard” data sets of interactions known to be true. Collins and Krogan used a probabilistic scoring method to assess the accuracy of two protein interaction networks gathered using the same experimental technique, affinity purification followed by mass spectrometry\textsuperscript{37}. Reguly and colleagues curated a network of over 33,000 interactions from the published literature and compared it to networks obtained by high-throughput methods\textsuperscript{38}. After applying methods like these, the resulting refined networks are more likely to contain true biological interactions and less likely to contain interactions measured in error.

A second approach to network refinement involves using the network to predict a biological response – perhaps to a gene deletion or to an environmental stimulus. Then, experiments are performed to observe the actual response. The agreement between the predicted and the observed responses is a measure of how well the network represents the real biological system. Next, computational approaches can be used to refine the network by adding or removing interactions, and then using
the refined network to predict the response. The best refinements will improve the network’s predictive power. This predict-compare-refine strategy has been applied to the yeast metabolic network\textsuperscript{39} and to transcriptional regulatory networks\textsuperscript{40,41}. In Chapter 4, we apply this approach to refine transcriptional regulatory networks that model responses to genetic perturbations.

Biological discovery is an ongoing process in which databases of biological information have been recognized as important tools. DNA sequences\textsuperscript{42,43}, protein structures\textsuperscript{44}, microarray results\textsuperscript{45,46}, and protein interactions\textsuperscript{13,15} are all stored and disseminated in online databases. The scientific community continually adds new and refines existing data. Databases of interaction networks, though, have only recently begun to emerge.

Designing a database for networks is primarily an engineering challenge. However, as we discuss in Chapter 5, the resulting database can be used to address scientific questions. Engineering design criteria include the mechanisms used to store networks, the interface for displaying networks to a user, and the tools provided for searching or browsing the repository. Networks are typically stored and presented in two ways: in a graphical format conducive to human interpretation or in a textual format that is easier to manipulate using computers. The software programs used for network analysis have typically used simple text formats\textsuperscript{10,11}. But specialized markup languages, such as BioPAX\textsuperscript{47} and SBML\textsuperscript{48}, have been developed to describe complex networks. In contrast, networks in journal articles are often presented in graphical figures, which may be easier for people to interpret, but are not amenable to further computational analysis.

A number of network-centric databases exist. The BioModels database stores
Contributions of this dissertation

In this dissertation, we describe four studies that address biological questions by using molecular interaction networks. These studies are broadly focused on gene regulatory networks – networks that model how proteins, typically transcription factors, regulate the expression of genes. We describe methods for assembling regulatory networks from high-throughput experiments that measures TF-promoter interactions, protein interactions, and gene expression levels. We also describe methods for analyzing regulatory networks. Several analysis methods seek to identify combinations of proteins that work together to regulate expression, either at same promoter or in the context of a regulatory pathway.

The experimental and computational tasks performed in these studies fall into four categories – network assembly, network analysis, network refinement, and network cataloging – that together can be thought of as the network lifecycle.

Assembly. In Chapter 2, we describe a map of the transcriptional response in yeast to a DNA damaging chemical, methyl-methanesulfonate (MMS). MMS alkylates DNA molecules in multiple places and is used as a model for understanding
chemotherapeutic drugs. We codified the relationship between two types of network data: (1) physical binding interactions between transcription factors and gene promoters, and (2) functional interactions derived from expression profiles of transcription factor deletions. We applied this knowledge to model how a select group of 30 transcription factors regulate DNA damage induced expression changes through a network of TF-promoter and protein-protein interactions. We present the first large scale regulatory network of the DNA damage response, and identify many instances of cross-talk, where a factor typically associated with one biological process in fact appears to regulate several processes.

**Analysis.** In Chapter 3, we describe an analysis of position effects in the yeast transcriptional regulatory network. We identified 22 transcription factors that preferentially bind the subtelomeric regions of chromosomes. Subtelomeric binding is a novel property of these transcription factors and of the transcriptional regulatory network. These findings suggest that genome position effects involve not only co-expression of neighboring genes along a genome but are also evident in the architecture, and perhaps evolution, of entire transcriptional regulatory networks.

**Refinement.** In Chapter 4, we describe a combined computational and experimental method for refining a transcriptional regulatory network. We used an information-theoretic score to rank genes in the network for further experimental investigation. We then performed four highly ranked experiments and assessed how well the observed responses were predicted by the network. We use simulations to demonstrate that our refinement method is theoretically superior to randomly picking genes and to picking ‘hub’ genes, those with many interaction partners.

**Cataloging.** In Chapter 5, we describe CellCircuits, a database of 1052 network models obtained from 11 published datasets. CellCircuits archives networks in graphical and text formats. It is accessible via the World Wide Web, and provides
an interface for searching for network models by a specific gene or Gene Ontology annotations. We demonstrate that a database of network models is useful for more than just storing models. By performing meta-analysis of the 1052 models in CellCircuits, we show how a database of models can be used to assess the novelty of new network results.
Chapter 2: Assembling a DNA damage response network

Abstract

Failure of cells to respond to DNA damage is a primary event associated with mutagenesis and environmental toxicity. To map the transcriptional network controlling the damage response, we measured genome-wide binding locations for 30 damage-related transcription factors (TFs) after exposure of yeast to methyl-methanesulfonate. The resulting 5,272 TF-target interactions revealed extensive changes in the pattern of promoter binding and identified damage-specific binding motifs. As systematic functional validation, we identified interactions for which the target changed expression in wild-type cells in response to MMS but was non-responsive in cells lacking the TF. Validated interactions were assembled into causal pathway models that provide global hypotheses of how signaling, transcription, and phenotype are integrated after damage.
Introduction

Exposure of cells to chemical and physical damaging agents can result in DNA lesions that contribute to the onset of cancer, aging, immune deficiencies, and other degenerative diseases\(^5\text{6}\). Initially, DNA damage is sensed by a highly-conserved mechanism consisting of the ATM and ATR protein kinases (the Tel1 and Mec1 enzymes in yeast) which aggregate at DNA lesions\(^5\text{7}\) and activate signaling cascades that include the Chk protein kinases (Chk1, Rad53, and Dun1 in yeast). These kinases, in turn, trigger both transcriptional and transcription-independent responses, including activation of DNA repair machinery and cell-cycle arrest\(^5\text{6}\).

Beyond the known DNA repair genes, genome-wide expression profiling in yeast has identified several hundred genes\(^5\text{8-60}\) that undergo increased or decreased expression in response to alkylation damage by methyl-methanesulfonate (MMS). At the level of growth phenotype, systematic deletion studies have also identified several hundred genes that are required for normal recovery from alkylation damage\(^6\text{1-63}\). Surprisingly, the set of genes that, when deleted, affect damage recovery is not enriched for genes whose transcript levels change upon damage exposure\(^6\text{2,64}\). Thus, transcriptional profiling alone, or genomic phenotyping alone, does not adequately define the cellular response to DNA damaging agents. However, these studies do suggest that the DNA damage response involves multiple levels of regulation, affecting not only DNA repair genes but also genes that influence protein and lipid turnover, cytoskeleton remodeling, and general stress pathways.
Results

Identifying transcription factors involved in the MMS response

To construct a global model of yeast transcriptional networks activated by MMS, we applied a systems approach\(^6^5\) that integrated data from genome-wide chromatin immunoprecipitation assays, expression profiling, systematic phenotyping, and protein interaction databases. First, we performed a systematic screen for transcription factors (TFs) involved in the MMS response. TFs were chosen from a set of 141 yeast DNA-binding factors\(^6^6\) and were selected according to any one of three criteria \(T\), \(B\), or \(S\) (Figure 2.1A). These criteria were TF expression (that is, the TF was differentially expressed after exposure to 0.03% MMS); expression of Bound genes (that is, the TF had been previously shown\(^6^6\) to bind the promoters of genes that were differentially expressed in the above MMS experiment); or Sensitivity (that is, deletion of the TF gene, if not lethal, caused growth sensitivity in MMS relative to nominal conditions, see Fig. S1).
A set of 23 TFs was identified (Figure 2.2A). Nine TFs were implicated by multiple criteria (Yap1, Gcn4, Cin5, Fkh2, Swi5, Swi6, Ixr1, Rim101, Rpn4), and three were encoded by essential genes (Hsf1, Mcm1, Ndd1). Two of the 14 MMS-sensitive TF-deletion strains ($sok2\Delta$ and $ecm22\Delta$) had not been reported in a previous genome-wide assay for MMS sensitivity$^{61}$. The set also included seven of the nine known cell-cycle regulators$^{25}$, as might be expected given that DNA damage affects cell cycle progression$^{56}$. A search of the literature identified 15 TFs that had been previously associated with regulation of DNA repair$^{67-69}$. Eight of these were also detected by our systematic criteria; the remaining seven were added to our list to yield a total of 30 TFs associated with the DNA damage response.

**Measuring the MMS-induced transcriptional network**

We then used the technique of chromatin immunoprecipitation coupled with microarray chip hybridization (chIP-chip) to identify the MMS-induced transcriptional network immediately downstream of each TF. Exponentially-growing yeast cultures were exposed to 0.03% MMS for one hour, resulting in ~50% cell viability$^{58,61}$. At a significance threshold of $p \leq 0.001$, the number of promoters bound by each TF ranged from 13 (Adr1 and Dig1) to 1,078 (Ino4) with an average of 214 per factor (Figure 2.2B). A total of 5,272 protein-DNA interactions with 2,599 distinct genes were identified for the 30 factors.

To reveal how the transcriptional network was reprogrammed between damaging and non-damaging conditions, we compared the MMS-induced interactions for each TF to the corresponding interactions as reported in nominal (non-induced) growth conditions by Lee et al.$^{66}$. Raw data from nominal conditions were reanalyzed using an identical data processing pipeline and significance threshold ($p \leq 0.001$) as for the MMS experiments, yielding a total of 4,996 protein-DNA interactions with
Figure 2.2. TF selection, chIP-chip, and expression profiling experiments.

(A) Results of the four criteria used to select the 30 TFs (TF expression, expression of bound genes, Sensitivity or Literature; see text). In column T, a ‘+’ or ‘−’ represents increased or decreased respectively. For column S, a ‘*’ denotes essential TFs. (B) Number of gene promoters bound by each TF. The three regions represent promoters bound exclusively in the absence of MMS (blue), presence of MMS (orange), or in both conditions (green). The proportion of genes in each region was compared to a negative control data set using Fisher’s exact test (see Methods). A red square in the left column ($P_c$, contracted) indicates that the proportion both/(both+absence) is significantly lower than expected in the negative control. Similarly, red in the right column ($P_e$, expanded) indicates that the proportion both/(both+presence) is lower.
2,588 distinct genes for the same 30 factors in nominal conditions.

For each TF, we applied a pairwise statistical analysis to score the overlap in promoter binding between the damage-induced and non-induced conditions. This method exploits the dependency in $p$-values of binding between two related chIP-chip data sets to identify TF-promoter interactions with more sensitivity than can be obtained if each data set were analyzed separately (see Methods and Supplemental Figure 2.2). Three to six factors bound significantly more genes in the presence of MMS, and conversely, eight factors bound significantly more genes under nominal growth conditions (Figure 2.2B).

Several promoter sets were enriched for DNA sequence binding motifs reported in the literature (Table 2.1)$^{70-72}$. The observed shifts in promoter binding for Cad1 and Hsf1 correlated with the known Cad1 and Hsf1 binding motifs, which were enriched in the sets of promoters bound before but not after MMS exposure. We also used the ANN-Spec algorithm$^{73}$ (see Methods) to search each set of promoters for motifs that had not been previously reported. Four such motifs were found, associated with Gcn4, Sko1, Ndd1, or Swi5, respectively (Table 2.1). The GCTCGAAAA motif was found upstream of 12 of 15 genes bound by Ndd1 in the presence of MMS, but was found upstream of only three of 30 genes bound in the absence of MMS. Such motifs may have been missed in past analyses because they are not active prior to damage exposure. They may represent DNA-binding sequences bound directly by the associated factor (Gcn4, Sko1, Ndd1, or Swi5) upon post-translational modification, or alternatively, they may be bound by a different TF that is coordinately recruited to the promoters after MMS treatment.

To identify combinations of TFs that regulate genes in common, we scored the significance of overlap between the gene sets bound by each pair of the 30 TFs (Figure 2.3A; hypergeometric test at $p \leq 0.01$). Cad1 was found to pair with Yap1 in
the absence of MMS but paired with Hsf1 in the presence of MMS. Several cell-cycle TFs that normally co-regulate large numbers of genes (such as Fkh2 and Swi6 or Ace2 and Swi5) no longer appeared to do so after MMS treatment, perhaps because DNA damage causes delayed progression through the cell cycle. A prominent combination that emerged after MMS treatment consisted of Ino4 and six other factors (Dal81, Mcm1, Rim101, Ecm22, Rpn4, and Uga3) from the “expanded” set, which regulated several hundred genes in common (Figure 2.3B). An additional set of >200 genes were targeted uniquely by Ino4 both before and after damage exposure (Figure 2.3B), supporting a previous hypothesis that Ino4 is actually a global regulator of gene expression.
Deletion-buffering analysis

Next, we sought to validate transcriptional effects of the measured binding interactions and to pinpoint the particular interaction pathways involved in transmission of the damage response signal. For this purpose, we used yeast genome microarrays to monitor MMS-induced gene expression changes across the viable knockout strains\(^75\) for the TFs found to be important in the DNA damage response (Methods). Of the 30 TFs identified, 27 were non-essential (Figure 2.2A column “S”) and could be profiled. Hierarchical clustering over all genes confirmed that these knockout profiles were globally more similar to the responses of wild-type cells to MMS [as measured in this study and previously\(^58,59\)] than to the expression responses of wild-type cells to other stress conditions\(^60,76\) (Supplemental Figure 2.3). Furthermore, although \(~20\%\) of the transcriptional response to MMS may be due to slowed cell cycle progression through G1/S phase\(^59,60\), the majority of responsive genes were not periodically expressed during the cell cycle (Supplemental Figure 2.4).

Processed expression data were analyzed to identify genes that were genetically “buffered” by one or more TF deletions. In this context, we define “deletion buffering” to mean an effect in which genes that are normally differentially expressed become unresponsive in a specific knockout background. For each gene-TF combination, wild-type and knockout profiles were analyzed to score the significance of the deletion-buffering effect using a Bayesian scoring scheme (Supplemental Figure 2.5 and Methods). At \(p \leq 0.005\), a total of 341 genes showed deletion buffering in the 27 knockouts, corresponding to 27 genes on average over a range of 90 genes for the Adr1 knockout to four genes for the Ecm22 knockout (Supplemental Figure 2.6).

As a positive control, we examined the deletion-buffering results for Crt1 (Rfx1), a transcriptional repressor of the ribonucleotide-diphosphate reductase (\(RNR\)) complex which catalyzes synthesis of new nucleotides during DNA repair\(^77,78\). As
expected, the expression levels of \textit{RNR2}, \textit{3}, and \textit{4} were deletion-buffered in the \textit{crt1Δ} strain but not in most other strains, and Crt1 bound the promoters of these genes before but not after MMS treatment (Figure 2.4A; Supplemental Figure 2.7).

Many of the remaining deletion-buffering events represent previously undocumented regulatory relationships. For example, contradictory to a prior report\textsuperscript{79} we found that both members of the Swi4-Swi6 complex could bind the \textit{DUN1} promoter and were required for the \textit{DUN1} transcriptional response (Figure 2.4B,C).

Beyond validation of individual interactions, the deletion-buffering analysis provided insights at the level of the damage response system as a whole. Paradoxically, the set of genes that are differentially expressed in response to DNA damage does not significantly overlap with the set of genes required for growth under damaging conditions\textsuperscript{62,64}. Our new expression data confirmed these findings for MMS, but also showed that the number of genes buffered by a TF knockout was highly correlated with its degree of MMS sensitivity ($r=0.72$; Figure 2.4D). For example, \textit{adr1Δ}, the most sensitive TF knockout in our study, also buffered the largest number of genes. Thus, the transcription factors most essential for cellular recovery after MMS exposure are, apparently, also the most central to the MMS transcriptional response.

The opposite of deletion-buffering is deletion-enhancement, that is, genes that are MMS-responsive in a TF deletion strain but not in wild type. Deletion enhancement was a much rarer event than deletion buffering. At the same $p$-value threshold, only 16 genes showed deletion enhancement whereas 341 showed deletion buffering. TFs associated with deletion-enhancement are apparently required to maintain stable expression of a set of genes, which become MMS responsive in their absence.
Building a network of regulatory pathways

Figure 2.4. Deletion buffering of the MMS response.

Differential expression and chIP-chip binding p-values of representative genes showing deletion-buffering for (A) Crt1, (B) Swi4, or (C) Swi6. Expression changes are colored yellow for up-regulation or cyan for down-regulation. Genes highlighted in red are those previously known to function in the DNA damage response. (D) Correlation between phenotypic sensitivity of each deletion strain in MMS (x-axis) versus the number of genes buffered by the TF-deletion (y-axis). Sensitivity scores were drawn from\(^n\). Scores range from 0 to 30, where high scores indicate that a TF-deletion was found to be sensitive to low concentrations (less than 0.03%) of MMS over replicate growth experiments. Only TFs with detectable sensitivity (score > 0) are shown.
Only 11% of the observed deletion-buffering events (37 out of 341) coincided with a direct chIP-chip binding interaction (Supplemental Figure 2.8). Such low overlap might occur for two reasons. First, failure to detect deletion buffering does not necessarily invalidate TF-promoter binding. Second, the observed deletion-buffering effect might be indirect, that is, mediated by longer regulatory pathways connecting the deleted TF to its regulated genes through one or more intermediate factors.

To identify these longer pathways, we applied a Bayesian modeling procedure to search the known physical network for the smallest set of paths (of two interactions) that were supported by the largest number of deletion-buffering events (Methods). Types of regulatory pathways identified are shown in Figure 2.5A. The physical network consisted of TF-promoter binding interactions from: [1] the 5,272 interactions measured in the presence of MMS for the 30 TFs in our study; [2] the 4,996 interactions measured for these TFs under nominal conditions; and [3] 5,903 interactions for the 74 additional TFs assayed by Lee et al., also in nominal conditions. We also included [4] a set of 14,319 high-throughput protein-protein interaction measurements from the Database of Interacting Proteins. These protein-protein interactions were measured in cells grown in the absence of MMS, and hence might or might not be present in MMS-treated cells. In the combined network, a link from protein $a$ to $b$ represented the observation that $a$ directly targets the promoter of the gene encoding $b$ (sources [1] through [3]) or that $a$ and $b$ physically interact (source [4]).

In total, we identified 68 buffering events that validated 88 longer paths. These paths were combined with the 37 direct effects identified earlier to formulate a model of the transcriptional response to MMS (Figure 2.5B). This model explains the MMS expression response of 82 genes and provides the basic scaffold on which inter-process
Figure 2.5. A network of regulatory interactions connecting transcription factors to deletion-buffered genes.

(A) Example regulatory paths in which deletion-buffering effects (squiggled arrows) support binding interactions (straight arrows). (B) The full validated model based on overlap between binding and buffering (see text). Buffering effects are omitted for clarity. Regions of the network are organized based on known functions of the deleted TFs. (C) Models resulting from follow-up experiments to generate buffering profiles for Mbp1 and Rtg1.
communication is achieved in the transcriptional response to an alkylating agent. At the core of the model are the known damage response genes \( RNR1, RNR2, RNR4, RFA1, RFA2, DIN7, DUN1, \) and \( MAG1 \) \(^{58}\). Although some of the TFs regulating these genes are expected based on previous studies (e.g., Swi4, Swi6, or Crt1), many others would not have been predicted, including those that have been previously associated with lipid metabolism (Ino4), stress response (e.g., Yap5), or cAMP-dependent signal transduction (Sok2). Overall, the model highlights extensive regulatory cross-talk among the processes of DNA replication and repair, cell cycle and cell-cycle arrest, stress responses, and metabolic pathways.

**Experimentally testing the model**

Every regulatory path of length two implicates an intermediate factor which is expected to regulate a similar set of genes as the deleted TF. A majority of these paths (49 out of 88), such as \( \text{Rtg3}\rightarrow\text{Ino4}\rightarrow\text{RFA2} \) and \( \text{Swi4}\rightarrow\text{Sok2}\rightarrow\text{MAG1} \), were already consistent with available data for both the source (e.g., Rtg3 or Swi4) and the intermediate (e.g., Ino4 or Sok2) factor, as both TFs were already included among the 27 assayed (“Reinforcing” in Figure 2.5A). Other paths included intermediate factors for which the transcription profiles were implied but untested (“Indirect” in Figure 2.5A), suggesting follow up experiments to refine the model.

Swi6 is thought to bind DNA in a complex with either Swi4 or Mbp1\(^{80}\). To discriminate which Swi6 targets were Mbp1 dependent, we analyzed an \( \text{mbp1} \) knockout strain. The set of genes deletion-buffered by \( \text{mbp1}\Delta \) was found to overlap with the \( \text{swi6}\Delta \) buffered set, including the genes \( RNR1 \) and \( DIN7 \) (Figure 2.5C). Thus \( RNR1 \) and \( DIN7 \) appear to depend on both Mbp1 and Swi6 for proper regulation.

Noting that loss of Rtg3 caused deletion-buffering of many downstream genes through the path \( \text{Rtg3}\rightarrow\text{Ino4} \), we investigated whether loss of Rtg1, a binding partner
of Rtg3\textsuperscript{81}, would produce a similar outcome. Indeed, loss of Rtg1 buffered many
genes that were bound by Ino4 (Figure 2.5C). However, the sets of Rtg1 versus Rtg3
deletion-buffered genes did not strongly overlap. Thus, in response to MMS exposure,
the three factors Rtg1, Rtg3, and Ino4 apparently collaborate to regulate a battery of
genes (whose products influence phospholipid metabolism and retrograde transport),
but their functional roles are not interchangeable.

We have integrated transcription factor binding profiles with genetic
perturbations, mRNA expression, and protein interaction data to reveal direct and
indirect interactions between transcription factors and MMS-responsive genes. The
result is a highly interconnected physical map of regulatory pathways supported by
both binding and deletion-buffering profiles. Some relations in this map are confirmed
by previous studies but most represent the basis for new hypotheses. As systems-level
approaches continue to map the connectivity of large cellular systems, an important
goal will be to make these maps even more integrative, and to learn how to use them
to predict the effects of different drugs, dosages, and genetic dispositions on pathway
function.
Methods

Strains and Media

Strains used in phenotyping and expression profiling were derived from the haploid BY4741. The parent strain was obtained from ATCC (Manassas, Virginia, USA), and all nonessential deletion strains constructed by the Saccharomyces Gene Deletion Project were obtained from Research Genetics (Huntsville, AL). Epitope-tagged strains (c-myc) used in construction of the physical regulatory network were derived from W303 and obtained from the laboratory of Dr. Richard A. Young at the Whitehead Institute (Cambridge, MA). Cells were cultured in standard yeast rich media (YPD) at 30°C except when noted.

Phenotypic Sensitivity Assays

Parent and all viable deletion strains harboring single knockouts of transcription factors (TFs) profiled in Lee et al. were arrayed into 96-well plates containing YPD broth and grown to saturation. Settled cells in each well were re-suspended to ensure homogeneity, and 5µl from each well was spotted simultaneously onto YPD agar plates with the use of a Hydra liquid handling apparatus (Robbins Scientific). The process was repeated on additional YPD agar plates containing a range (0.01% to 0.1%) of methyl methanesulfonate concentrations (Sigma Chemical Company) to determine the optimum concentration to adequately detect sensitive knockout strains. MMS was added to cooled agar and used fresh within one day to ensure the stability of the agent. Spotted plates were grown for 60 h at 30°C and imaged using a Gel Doc 1000 (BioRad) running Quantity One software, and all screens were performed in triplicate using fresh cultures.
Whole Genome Expression Analysis

Gene expression experiments were processed in parallel with at least two distinct biological samples from each strain (colonies of similar size picked from fresh YPD + G418 agar plates) grown to saturation in YPD overnight at 30°C. The overnight culture was diluted 1:100 in a flask of 100ml fresh YPD and grown in a shaking incubator at 30°C 180 rpm until the culture reached an OD$_{600}$ of 0.8 – 1.0. Every effort was made to closely match the final optical density of each of the samples. Once a culture reached the appropriate cell density, the sample was split in half, MMS added to a final concentration of 0.03% in the non-reference samples, and both cultures grown for an additional hour. Cells were harvested by centrifugation at 3000 rpm for five minutes at room temperature in a Legend RT centrifuge (Kendro Laboratory Products, Asheville, NC, USA). Cell pellets were immediately snap-frozen in liquid nitrogen to suspend gene expression (including any temperature stress response genes) and stored at −20°C prior to RNA extraction. Total RNA from each sample was isolated by hot acid phenol extraction and mRNA-purified via Poly(A)Pure kits (Ambion). Labeling of cDNA was performed in a dye-reversal scheme by direct incorporation using a CyScribe First-Strand cDNA Labeling Kit (Amersham Biosciences). Corresponding Cy-3 and Cy-5 labeled samples were co-hybridized to microarrays containing the Yeast Genome Oligo Set Version 1.1 (Qiagen).

Genome Wide Transcription Factor Binding Analysis

Samples were processed in parallel using three distinct biological replicates from each epitope-tagged strain and grown to saturation in YPD overnight at 30°C. The overnight culture was diluted 1:100 in a flask of 50 ml fresh YPD and grown in a shaking incubator at 30°C at 180 rpm until the culture reached an OD$_{600}$ of 0.8 – 1.0.
MMS was added to a final concentration of 0.03%, and the culture was grown for an additional hour. Protein-DNA binding locations were assayed as previously described by Lee et al. with corresponding IP-enriched and unenriched samples co-hybridized to a single cDNA microarray containing all yeast intergenic sequences derived from PCR amplification.

**Array Preparation and Hybridization**

Microarrays were spotted on UltraGAPS II slides (Corning) using an OmniGrid 100 microarrayer (GeneMachines). After spotting, arrays were baked at 80°C for 2 hours, cross-linked at 300 mJ in a UV Stratalinker 2400 (Stratagene), and stored under vacuum. Hybridizations were conducted at 42°C for 15 hours using the Lucidea SlidePro automated hybridization machine (Amersham Biosciences), and arrays scanned using GenePix 4000A (Axon Instruments) or ScanArray Express scanners (Perkin-Elmer) at a 10.0 μm resolution.

**Data Post-processing and Significance Assessment**

Scanned images were processed using GenePix Pro 3.0 (Axon Instruments) or QuantArray (PerkinElmer) software to obtain raw Cy-3 and Cy-5 foreground \((f_3, f_5)\) and background \((b_3, b_5)\) intensity measurements for each spot on the array. Background intensities were smoothed using a 7x7 median spatial filter, separately for the Cy-3 and Cy-5 channels to obtain \((b_3', b_5')\). Background-adjusted log ratios \(L = \log_{10}(\frac{f_3 - b_3'}{f_5 - b_5'})\) were corrected for cyanine-dye dependent bias using Qspline normalization. Corrected log ratios \(L'\) were spatially normalized by subtracting the median \(L\) value within a 9x9 window. The four replicate arrays for each gene expression experiment were processed using the VERA package to estimate multiplicative and additive errors and to associate a \(p\)-value of differential
expression with each gene. This approach was also applied for error modeling and significance assessment of the chIP-chip data. However, unlike for gene expression analysis, in which both increases and decreases in fluorescent intensity are of interest, DNA binding is indicated only for increases in intensity, representing increased promoter binding in the IP-enriched sample versus the IP-unenriched sample. Thus, significance of DNA binding must be assessed using a one-sided test. The VERA likelihood ratio test was therefore modified to use a one-sided statistic [force $\mu_x \geq \mu_y$ in the denominator of Eqn. 5 of reference$^{83}$].

**Overlap in Binding Between Conditions**

For each transcription factor, $p$-values of binding from $+\text{MMS}$ and $-\text{MMS}$ experiments are integrated to select sets of genes bound in $-\text{MMS}$ only, $+\text{MMS}$ only, or in both conditions, as displayed in Figure 2.1 of the main text. Typically, genes are selected in a single condition by including those with $p$-values below a certain threshold, such as $t = 0.01$. However, given that two conditions $c_1$ and $c_2$ are under consideration, the dependency in the corresponding pair of $p$-values $(p_1, p_2)$ measured for each gene can be exploited to achieve a more sensitive selection. Treating the pair of values as replicate measurements of binding, a compounded $p$-value of binding (in any/either condition) can be computed by multiplying $p$-values, i.e., genes for which $p_1 \cdot p_2 < t'$ are chosen as bound.

Based on this observation, we combine the one- and two-condition thresholding methods, associated with $t$ and $t'$ above, to partition the two-dimensional space of $p$-values into four regions as shown in Figure S2. Region (A) is defined by $(p_1 \leq t, p_2 > t'/t)$ and represents binding in $c_1$ only. Region (C) is symmetrically defined by $(p_1 > t'/t, p_2 \leq t)$ and represents binding in $c_2$ only. Region (B) is the area defined by $(p_1 \cdot p_2 \leq t, p_1 \leq t'/t, p_2 \leq t'/t)$ and represents binding in $c_1$ and $c_2$. The remaining values
of \((p_1, p_2)\) not covered by regions (A-C) represent lack of binding. Furthermore, for a given simple threshold \(t\), the compound threshold \(t'\) is set so that regions (A-C) are equal in area, which ensures that under the null hypothesis (i.e., no genes are bound in either condition) the expected numbers of genes which fall into each region are the same. This two-dimensional thresholding scheme is similar to one proposed by Zaykin \textit{et al.}\textsuperscript{84} for epidemiological studies.

**Significance of expansion or contraction**

Given overlap counts (A, B, C) for each TF, we computed the significance of a possible change in genome-wide binding pattern between untreated and MMS-treated conditions. To construct a null model in which the TF binding profiles did not change across conditions, additional chIP-chip experiments were performed in untreated conditions for each of three transcription factors Gcn4, Crt1, and Rpn4. As with all other experiments, three replicate experiments per TF were normalized and processed using the VERA/SAM package. These data were compared to the original chIP experiments for Gcn4, Crt1, and Rpn4 generated by the Young laboratory\textsuperscript{24,70}, also in untreated conditions.

Each of these untreated (original Young lab) vs. untreated (new Ideker lab) comparisons was analyzed to compute the overlap in binding, exactly as described above for the untreated (original Young lab) vs. treated (new Ideker lab) binding profiles. Promoter counts falling into regions A, B, or C were aggregated across Gcn4, Crt1, and Rpn4 yielding \(A=347, B=459, C=239\). These totals were used to derive a pooled estimate of \(B/(B+(A+C)/2) = 61\%\) for the proportion of promoters bound in both conditions in the negative control.

This expected value was compared to the \(B/(B+A)\) and \(B/(B+C)\) proportions for the -/+ MMS comparison across each of the 30 TFs in our study using a one-sided
Fisher’s Exact Test (FET). The $p$-value of contraction or expansion was defined as the FET significance that the first or second fraction, respectively, was lower than expected under the null model.

**Motif Enrichment in Transcription Factor Binding Data**

Transcription factor binding site motifs were defined as position-specific weight matrices (PWMs) of log-likelihood ratios. PWMs were compiled for 114 different individual TFs from Harbison et al. and public data bases. When more than one matrix was defined for the same TF, the PWM with the highest information content per position (relative entropy) was selected. Using the PWM scoring functionality of ANN-Spec, the score distribution for each motif was determined over all possible subsequences of the intergenic regions (as defined by the probes on the chIP-chip array) such that a score threshold could be selected to ensure a rate of $< 10^{-4}$ predicted sites per base pair.

ANN-Spec was used to discover potentially new motifs in promoter sets defined by the method described in “Overlap in Binding Between Conditions” ($t=10^{-3}$). This defined three non-overlapping sets for each TF: +MMS-only, –MMS-only, and BOTH. For each sequence set, 100 training runs were performed for pattern widths 8 to 14 using ANN-Spec’s discriminative mode against all intergenic regions. Training allowed for zero or more sites per alignment (so called ‘ZOPS’ occurrence model). Redundant or approximately redundant alignments were merged (correlation coefficient $>0.8$, over 6 or more contiguous alignment positions in either orientation) and the remaining alignments were filtered against the 114 known TF-motifs in a similar way. The remaining 222 alignments were converted to PWMs and represented our novel motif models (108 in –MMS-only, 71 in +MMS-only and 43 in BOTH).

For each TF and known and novel motif, PWM enrichments were calculated
(hypergeometric p-value) in either the –MMS-only or +MMS-only non-overlapping binding sets to identify differentially enriched motifs. Differential enrichment was defined when binding p-values $<10^{-7}$ ($<10^{-3}$ after Bonferroni correction) were observed in one condition, and p-values $>10^{-2}$ were observed in the other condition.

**Deletion Buffering Analysis**

Expression profiles (+/−MMS treatment for wild type and 27 TF knockout strains) were analyzed to identify genes that were differentially expressed in the vast majority of profiles but did not change in expression in a particular TF knockout background. In these cases, the gene was said to be deletion-buffered by the TF in question and the TF was “epistatic” to this gene.

For each (gene, TF) combination, the probability of buffering was computed using a Bayesian score function, as follows. Let $T_{ko}$ vs. $T_{other}$ represent the event of true differential expression of the gene in the TF knockout vs. all other strains. Similarly, let $p_{ko}$ and $p_{other}$ represent the corresponding observed values for the gene, which in this case are $p$-values of differential expression. The value for $p_{other}$ is computed as the combined $p$-value of differential expression over all profiles excluding the knockout under consideration, according to Fisher’s rule of multiplication $86$:

$$m = \prod_{i \neq ko} p_i$$

$$p_{other} = m \sum_{n=0}^{n=m} \left( -\ln m \right) / i!$$

Since most genes are likely regulated (directly or indirectly) by only a fraction of the TFs encoded by the genome, they are expected to behave similarly between the wild type and most of the 27 knockout backgrounds. The combined $p$-value thus provides a measure of differential expression in + vs. – MMS conditions that is more robust than
the single \( p \)-value obtained for a wild type experiment. Given measurements of \( p_{ko} \) and \( p_{other} \), the probability of deletion-buffering is equivalent to the probability that the gene is truly differentially expressed in other experiments but not in the knockout in question. Using Bayes’ rule:

\[
\Pr(I_{ko} = 0, I_{other} = 1 | p_{ko}, p_{other}) = \frac{\Pr(p_{ko}, p_{other} | I_{ko} = 0, I_{other} = 1) \Pr(I_{ko} = 0, I_{other} = 1)}{\Pr(p_{ko}, p_{other})} = \frac{\Pr(p_{ko} | p_{other}, I_{ko} = 0, I_{other} = 1) \Pr(p_{other} | I_{ko} = 0, I_{other} = 1)}{\Pr(p_{ko}, p_{other})}.
\]

[Because \( \Pr(I_{ko} = 0, I_{other} = 1) \) is constant]

\[
\equiv \frac{\Pr(p_{ko} | I_{ko} = 0) \Pr(p_{other} | I_{other} = 1)}{\Pr(p_{ko}, p_{other})}.
\]

[Given \( I, p \) is conditionally independent of other variables]

By identity, \( \Pr(p_{ko} | I_{ko} = 0) = p_{ko} \). The remaining two terms are estimated empirically from the available frequency distributions. \( \Pr(p_{ko}, p_{other}) \) is estimated by a two-dimensional histogram of the observed values over all genes and experimental conditions in our study. \( \Pr(p_{other} | I_{other} = 1) \) is estimated by a histogram of \( p_{other} \) restricted to “true” differentially-expressed genes, defined as the set of 255 genes for which the median absolute expression change was greater than two-fold over at least two of seven time points (5, 15, 30, 45, 60, 90, 120 min.) monitored by Gasch et al. after MMS addition.58

**Assembly of Transcriptional Pathways From Integrated Data**

Physical mechanisms of transcriptional regulation were modeled using an approach described previously35. Briefly, we postulated that the regulatory effects of deleting a gene are propagated along paths of physical interactions (protein-protein
and protein-DNA). We formalized the properties of these paths and interactions using a factor graph and found the most probable set of paths using the max-product algorithm \(^8^7\). The raw data used in the modeling procedure included: 7,055 promoter-binding interactions found in –MMS only; 3,851 promoter binding interactions found in both + and – MMS; and 1,431 promoter-binding interactions found in +MMS only; the set of all 15,166 pair-wise protein-protein interactions recorded in the Database of Interacting Proteins as of April 2004 \(^1^6\); and the 341 buffering interactions found using the method described in “\textit{Genome Wide Epistasis Analysis}” (buffering p-value ≤ 0.005). For the 30 transcription factors in our study (see Figure 2.2a), we used the method described in “\textit{Overlap in Binding Between Conditions}” to identify genes bound in either +MMS only, -MMS only, or both conditions. For 75 additional factors from Lee \textit{et. al.} for which we had data for the –MMS condition only, we used a strict p-value threshold of p ≤ 0.001 to select significant promoter-binding interactions and classified these as –MMS only interactions.

In Figure 2.5b of the main text all paths that directly or indirectly (via one intermediate TF) connect TFs to deletion-buffered genes are shown. In supplemental web figures “Integrated models by TF” these paths are partitioned to show the paths connecting each individual TF to genes that it buffers. To generate integrated models for the Mbp1 and Rtg1 validation experiments (Figure 2.5c), the same promoter-binding and protein-protein interaction network was used along with either 14 buffering interactions from the Mbp1 experiment or 20 from the Rtg1 experiment (buffering p-value ≤ 0.004).

\textbf{Assessing overlaps in binding between TF pairs}

Each of the sets of genes bound by a TF (in –MMS or +MMS conditions using a strict threshold of p ≤ 0.001) was systematically compared to the sets of genes bound
by each of the other TFs using the hypergeometric test. Significant overlaps between
sets (p ≤ 0.01 after Bonferroni correction) are displayed in Figure 2.3 of the main text.
TFs are linked by a green line if a significant number of genes were bound by both
TFs in –MMS and (a possibly different set of genes were bound) in +MMS conditions.
If two TFs only bound the same genes in either –MMS or +MMS, they are linked with
a blue or orange line, respectively. Hierarchical clustering was performed using the
ClustArray program (http://www.cbs.dtu.dk/services/DNAarray/).
Tables

Table 2.1. DNA sequence motifs found in promoters bound in only one condition

1A. Motifs found in -MMS, but not +MMS\(^1\)

<table>
<thead>
<tr>
<th>Consensus sequence</th>
<th>Promoter set</th>
<th># promoters w/motif (-MMS)(^4)</th>
<th># promoters w/motif (+MMS)(^4)</th>
<th>Motif source(^5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGACTC</td>
<td>Gcn4</td>
<td>20/31</td>
<td>0/5</td>
<td>GCN4(^\dagger), BAS1(^*)</td>
</tr>
<tr>
<td>ATTAGTAAGC</td>
<td>Cad1</td>
<td>18/26</td>
<td>0/69</td>
<td>CAD1(^*)</td>
</tr>
<tr>
<td>eGGGGG</td>
<td>Hsf1</td>
<td>35/143</td>
<td>1/41</td>
<td>ADR1(^*)</td>
</tr>
<tr>
<td>CGGGGGCACnCTcStCCG</td>
<td>Hsf1</td>
<td>43/143</td>
<td>3/41</td>
<td>GAL4(^\dagger), PUT3(^*)</td>
</tr>
<tr>
<td>TTCtannnnnTTC</td>
<td>Hsf1</td>
<td>44/143</td>
<td>7/41</td>
<td>HSF1(^*)</td>
</tr>
<tr>
<td>CGGGrnCnGGG</td>
<td>Hsf1</td>
<td>37/143</td>
<td>0/41</td>
<td>LEU3(^*)</td>
</tr>
<tr>
<td>CGGGGCGCCCGAn</td>
<td>Hsf1</td>
<td>43/143</td>
<td>6/41</td>
<td>RFA2</td>
</tr>
<tr>
<td>GCCSnGSCC</td>
<td>Hsf1</td>
<td>40/143</td>
<td>1/41</td>
<td>SKN7(^*)</td>
</tr>
<tr>
<td>GCGGCnnGCGGC</td>
<td>Hsf1</td>
<td>39/143</td>
<td>2/41</td>
<td>STP1(^*)</td>
</tr>
<tr>
<td>ACCCGTACAt</td>
<td>Yap5</td>
<td>44/104</td>
<td>0/3</td>
<td>SFP1(^*)</td>
</tr>
</tbody>
</table>

1B. Motifs found in +MMS, but not –MMS\(^6\)

<table>
<thead>
<tr>
<th>Consensus sequence</th>
<th>Promoter set</th>
<th># promoters w/motif (-MMS)(^4)</th>
<th># promoters w/motif (+MMS)(^4)</th>
<th>Motif source(^5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMSCTGCAAAnTT</td>
<td>Gcn4</td>
<td>0/31</td>
<td>5/5</td>
<td>Predicted (ANNspec)</td>
</tr>
<tr>
<td>GCTCGAAAA</td>
<td>Ndd1</td>
<td>3/30</td>
<td>12/15</td>
<td>Predicted (ANNspec)</td>
</tr>
<tr>
<td>TGAYTAACn</td>
<td>Sko1</td>
<td>7/153</td>
<td>11/14</td>
<td>Predicted (ANNspec)</td>
</tr>
<tr>
<td>TnTCnCTCAT</td>
<td>Swi5</td>
<td>6/61</td>
<td>9/10</td>
<td>Predicted (ANNspec)</td>
</tr>
<tr>
<td>GCGGCnnGCGGC</td>
<td>Pdr1</td>
<td>0/2</td>
<td>21/48</td>
<td>STP1(^*)</td>
</tr>
<tr>
<td>GCGGGGnCGG</td>
<td>Pdr1</td>
<td>0/2</td>
<td>19/48</td>
<td>SUT1(^*)</td>
</tr>
</tbody>
</table>

\(^1\) Enriched in -MMS only promoters (hypergeometric \(p < 10^{-7}\)) but not in +MMS only (\(p > 10^{-2}\))

\(^2\) Letter based on IUPAC code, lowercase used when information content < 0.3

\(^3\) The set of promoters bound by each TF was analyzed for the presence of known and novel motifs

\(^4\) Motif presence was scored using ANN-spec at a threshold that predicted sites at a rate of <10\(^{-4}\) overall intergenic regions.

\(^5\) Novel motifs were identified using ANN-spec. Known motifs were drawn from the \(\dagger\) SCPD databases or from \(^*\) Harbison et al.

\(^6\) Enriched in +MMS only promoters (hypergeometric \(p < 10^{-7}\)) but not in –MMS only (\(p > 10^{-2}\))
Supplemental Figure 2.1. MMS sensitivity assay of transcription factor deletions strains.

Light images of yeast colonies on agar plates are shown. Each position in the grid contains the colonies for TF-deletion or wild-type strains grown for 3 days in YPD (A) and YPD +0.025% MMS (B). The key for the location of yeast strains is shown in (C). Sensitive mutants for this assay are indicated in red. Factors marked with '*' were not found to be sensitive in subsequent trials.
Supplemental Figure 2.2. Scoring overlap in TF binding between conditions.

For a given TF, each gene is assigned a pair of $p$-values for binding in +MMS vs. –MMS. Genes with $p$-value pairs that fall into the region (A), (B), or (C) are considered bound in +MMS only, both conditions, or –MMS only, respectively. The inset shows an enlargement of the plot for $p$-values < 0.25. The regions shown correspond to $p$-value thresholds of $t=0.01$ and $t'=0.002$. The thresholds used in the actual study were $t=0.001$ and $t'=1.5\times10^{-4}$.
Supplemental Figure 2.3. Clustering of MMS expression responses with other stresses.

Expression profiles generated in this study (wild type and TF knockouts +/- 0.03% MMS) were clustered together with previous expression data for MMS [Gasch et al. 58 and Jelinsky et al. 59] and other environmental stresses60. Hierarchical clustering was performed using ClustArray (http://www.cbs.dtu.dk/services/DNAarray/) to construct a dendrogram on the conditions. The dendrogram shows that the +/- MMS wild type profile from this study is in relative agreement with previous ones. Moreover, all but two (cad1Δ and dal81Δ) of the +/- MMS knockout profiles are more similar to the MMS wild type response than to other stress responses.
**Supplemental Figure 2.4. MMS responsive genes implicate G1/S checkpoint.**

(A) Approximately 22% of the MMS-responsive genes (111 of 514) have been shown to be temporally regulated during the cell cycle\(^{205}\). (B) Distributions of the median log-ratios (over the 27 TF knockouts and wild-type) for genes found to be up-regulated in G1, S, S/G2, G2/M, M/G1. Boxplot colors correspond to the color scale in C. (C) Clustering these MMS responsive genes with representative profiles from cell cycle phases G1, S, G2, M and M/G1 \(^{205}\) shows that the MMS profile is most similar to that of late G1 and S phases. Further evidence of G1 and S phase arrest can be found in the expression responses of *ACE2* and *SWI5*, cell cycle regulators expressed in M/G1 phase\(^{205}\). Both factors are down regulated in response to MMS (Figure 2.2a in the main text), and have contracting binding profiles (Figure 2.2b). Moreover, the Ace2 and Swi5 TFs only show significant binding overlap in −MMS (Figure 2.3). Together, these findings confirm that exposure to MMS causes cells to progress more slowly through G1/S and S phase\(^{60}\).
Supplemental Figure 2.5. Crt1 deletion-buffering $P$-values.

Deletion buffering $P$-values (colors ranging from green to white) are computed for each gene (black points) based on their $P$-values of differential expression for $crt1\Delta$ (x-axis) versus all experiments (y-axis).
Supplemental Figure 2.6. Number of genes buffered by each transcription factor deletion (buffering p-value <0.005).

Because more than one transcription factor may buffer each gene, the number of buffered genes (341) is less than the total number of buffering interactions observed (757). As indicated by the dashed grey line, the average number of genes buffered by each TF is 26.5.
Supplemental Figure 2.7. Correspondence between expression and regulation of the RNR genes by Crt1p.

(A) Distributions of normalized spot intensities (cy3 and cy5 channels) for the four RNR genes in crt1Δ + vs. −MMS over four replicate microarrays (16 spot intensities for each distribution). The wild-type expression levels show an induction of these genes after exposure to MMS while the crt1Δ shows constitutive expression both + and −MMS. (B) The binding data show strong evidence for Crt1p binding upstream of RNR2,4 (and marginal evidence for RNR1,3) in nominal growth condition while binding evidence suggests a lack of Crt1p binding for all RNR genes after exposure to MMS. These data support the induction (by derepression) observed in the first two distributions in (A).
Supplemental Figure 2.8. Direct binding interactions supported by deletion-buffering effects.

Circles or squares represent deleted TFs or target genes, respectively. Protein-DNA interactions between TFs and the gene promoters they bind are represented as arrows colored blue (bound in −MMS only), orange (+MMS only), or green (both +/−MMS). Genes are colored orange or blue, representing induction or repression in response to MMS. Each binding interaction shown is supported by a deletion-buffering effect. For example, deletion of the transcription factor SWI6 results in lack of expression of DIN7 and DUN1 in response to MMS. In total, 42 binding interactions covering 37 distinct target genes are supported.
Acknowledgements

We acknowledge funding from the National Institute of Environmental Health Sciences (NIEHS), the National Cancer Institute (NCI), and the David and Lucille Packard Foundation. LDS is the Ellison American Cancer Society Research Professor. We thank R. Young, D. Odom, T.I. Lee, and B. Ren for assistance with chIP-chip analysis, and J. Kadonaga, W. McGinnis, and G. Kassavetis for consultations on the binding motif predictions.

Chapter 2, in full, is a reprint of material as it appeared on May 19 2006 in Science, 2006, vol. 312, p. 1054-1059. Workman, Christopher T.; Mak, H. Craig; McCuine, Scott; Tagne, Jean-Bosco; Agarwal, Maya; Ozier, Owen; Begley, Thomas J.; Samson, Leona D.; Ideker, Trey. The dissertation author was one of three primary investigators (with C.T. Workman and S. McCuine) and two primary authors (with C.T. Workman) of this paper.
Abstract

Transcription factors are most commonly thought of as proteins that regulate expression of specific genes, independently of the order of those genes along the chromosome. By screening genome-wide chromatin immunoprecipitation (chIP) profiles in yeast, we find that more than 10% of DNA-binding transcription factors concentrate at the subtelomeric regions near to chromosome ends. None of the proteins identified were previously implicated in regulation at telomeres, yet genomic and proteomic studies reveal that almost all have interactions with established telomere binding complexes. For many factors, the subtelomeric binding pattern is dynamic and undergoes flux toward or away from the telomere as physiological conditions shift. We find that subtelomeric binding is dependent on environmental conditions and correlates with the induction of gene expression in response to stress. Taken together, these results suggest that genome position effects involve not only co-expression of neighboring genes along a genome but have implications for the architecture and evolution of entire transcriptional regulatory networks.
Introduction

It is becoming increasingly evident that gene order is not random\textsuperscript{88,89}. Expression profiling of diverse eukaryotic species has revealed that co-expressed genes are often clustered along the chromosome\textsuperscript{90-93}, and that such clusters of genes function to varying degrees in the same metabolic pathways\textsuperscript{94}. In yeast, for example, adjacent genes are co-regulated during the cell-cycle\textsuperscript{95} or in response to changing growth conditions\textsuperscript{96}. In multicellular eukaryotes such as flies and humans, extended tracts of co-expression have been observed encompassing up to thirty genes\textsuperscript{97,98}.

Broadly speaking, the effects of gene order on expression are known as “position effects.” Yet despite intensive study, the mechanisms controlling position effects are incompletely understood\textsuperscript{88}. The chromosome ends of the yeast \textit{Saccharomyces}, though, have been used as a model to link position effects to epigenetic factors such as the state of surrounding chromatin and the spatial compartmentalization of the nucleus\textsuperscript{99}. These epigenetic factors are, in part, assembled by DNA-binding transcription factors and chromatin modifying proteins\textsuperscript{100,101}. At chromosome ends, such transcription factors (including Rap1p) and chromatin modifiers (including Hda1p and the Sir silencing complex) appear to bind in a position-specific manner across multiple genes in a broad genomic region\textsuperscript{102-104}, but do not completely explain the observed expression patterns of genes located near these ends\textsuperscript{103,105,106}.

Transcription factors have traditionally been assumed to regulate specific gene promoters regardless of their genomic locations\textsuperscript{24,70,107}. The vast majority of transcription factors therefore have not been investigated for position-specific binding, which, if found, would suggest a broader role for those factors in position effects than has been previously appreciated. Thus, we developed methods for analyzing the wealth of available transcription binding profiles and applied them to yeast to
discern whether position-specific binding could be observed as a general phenomenon, focusing in particular on position effects at chromosome ends.

We find that a surprising number (more than 10%) of all profiled yeast DNA-binding transcription factors display a marked preference for binding genomic locations within 25 kilobases of the telomere, and much of this position-specific binding is responsive to changes in physiological conditions. We also assay the phenotypes of single and double deletions of these transcription factors in response to physiological challenges. None of these factors have been previously known to have any correlation with genome position, but we find that most interact with proteins with known telomere functions. Taken together, our findings suggest that genome position effects involve not only co-expression of neighboring genes along a genome but are also evident in the architecture, and perhaps evolution, of entire transcriptional regulatory networks.

Results

Discovery of a large family of Subtelomere Binding Transcription Factors

We screened the existing compendium of transcription factor binding locations gathered using the technique of chromatin immunoprecipitation followed by microarray analysis for each of 203 *S. cerevisiae* transcription factors\(^70\). These data were collected from yeast grown in rich medium and under stress conditions.

Each transcription factor (TF) was scored using a quantitative measure of telomere-proximal binding that we defined to be its Telomere Distance Profile (TDP). We computed the TDP for a TF by measuring the distance to the closest telomere for every target sequence reported to be bound by that TF, resulting in a distribution of distances. Then, we compared each TDP to a background Telomere Distance Profile consisting of all yeast genes (Figure 3.1A; see Methods).
Figure 3.1. Transcription Factors that Preferentially Bind Sequences at Subtelomeres.

(A) Background Telomere Distance Profile (TDP) for all yeast promoters. (B) Overview of TDPs for rich-medium promoter-binding profiles compared against the background distribution. Each dot represents data for one TF: its significance score on the y-axis (the P-value of a one-sided Kolmogorov-Smirnov test) versus the percentage of bound promoters that are subtelomeric on the x-axis. Subtelomere binding transcription factors (SBTFs) having bi-modal TDPs more significant than the P-value threshold of 0.001 are labeled and indicated in black. (C, D, E) Telomere Distance Profiles for (C) Gat3p – the statistically most significant SBTF, (D) Msn4p – a SBTF that is a master regulator of stress response genes, and (E) Ace2p – a cell cycle regulator that is representative of TFs having a TDP similar to the background. Blue dashed lines correspond to the blue dashed line in (A). The red dashed line indicates the 25 kb cutoff distance used to categorize subtelomeric genes. (F, G) Plots showing the location of promoters bound on the 16 yeast chromosomes by (F) Gat3p and (G) Msn4p. Red ticks indicate binding events located within 25 kb of a telomere. Black ticks indicate non-telomeric binding events. Small black dots mark the centromere on each chromosome. (H) The average percent of subtelomeric promoters bound by three groups of TFs: SBTFs, cell cycle TFs (those annotated with the “cell cycle” GeneOntology term GO:0007049), and the remaining TFs. Error bars indicate standard deviation.
Among the TF binding profiles from yeast grown in rich medium, seventeen TFs had a TDP that was significantly different from the background distribution according to a Kolmogorov-Smirnov test at \( P < 0.001 \) (Figure 3.1B). This \( P \)-value threshold corresponds to a false discovery rate\(^{108} \) of approximately 1%, meaning that none of the TDPs identified are expected to be false positives (Supplemental Figure 3.1).

Fifteen of these significant TDPs were distinctly bi-modal, with an unusually high number of target sequences located within 25 kb of the closest telomere (Figures 1C-D, F-G). This bi-modality was not simply due to low gene density 25 kb from the telomere, since the background distribution was not depleted for genes in this region (Figure 3.1A). Of note is that in \textit{S. cerevisiae} the chromosomal region known as the subtelomere is postulated to extend roughly 25 kb inwards from each telomere\(^{109} \). The subtelomere is a patchwork of sequence blocks that are repeated within and between the ends of chromosomes, forming the transition between telomeric repeats and chromosome-specific sequences\(^{110} \). Thus, we named these 15 factors SBTFs, for Subtelomere Binding Transcription Factors.

We next analyzed the 84 TF binding profiles that had been reported under various stress conditions such as rapamycin, butanol, or hydrogen peroxide\(^{70} \). Eleven SBTFs were identified that showed a subtelomeric binding preference under stress. Conversely, a number of SBTFs that had been identified in rich medium were found to lose their subtelomeric binding preferences in alternative conditions (Figure 3.2).

In total, this raised the number to 22 SBTFs identified (Figure 3.3A): seven TFs for which the subtelomeric binding preference was specific to a stress condition, six TFs for which it was specific to rich media, four TFs for which subtelomeric binding was observed under both stress and rich media, and five TFs for which no stress binding data were available. It is perhaps intriguing that 11% of \textit{S. cerevisiae}
Figure 3.2. Example SBTFs that Show Dynamic Binding Preferences as a Function of Growth or Stress Condition.

(A) SBTFs with Telomere Distance Profiles (TDP) that are significant only in a stress condition. TDPs are shown for each SBTF in two conditions: rich medium (top) and a stress condition (bottom). Genes considered subtelomeric are colored light red in rich-medium binding profiles or dark red in stress profiles. (B) SBTFs that show subtelomeric preference in rich medium. Stress conditions are abbreviated: RAPA (100 nM rapamycin), H2O2 Hi (4mM hydrogen peroxide), and galactose (2% in YEP medium).
Figure 3.3. Subtelomeric Binding Preference is Dynamic and Distributed into Distinct Clusters.

(A) Comparison of stress versus rich-medium SBTF promoter binding profiles. Dark and light horizontal red bars indicate the number of subtelomeric targets bound. Open bars indicate the total number of targets bound. Check marks on right indicate whether the SBTF shows a significant (P<0.001) subtelomeric preference in the (S) stress condition or (R) rich medium. (B) Hierarchical clustering of the 125 subtelomeric genes (columns) bound by one or more SBTFs (rows). Blue indicates binding. Colored bars at right correspond to the dynamic binding behaviors in panel A: SBTF displays a subtelomeric preference in stress only (S; dark red), rich media only (R; light red), or both stress and rich media (SR; bright red). Stress conditions are abbreviated: Acid (succinic acid, pH 4), BUT90 (1% butanol), GAL (2% galactose), RAFF (2% raffinose), H2O2Lo (0.4 mM hydrogen peroxide), H2O2Hi (4mM hydrogen peroxide), RAPA (100 nM rapamycin), SM (0.2 mg/ml sulfometuron methyl, an inhibitor of amino acid biosynthesis).
bound a completely different set of stress responsive promoters upstream of hexose transporters (HXT), flocculation genes (FLO, FSP2), and a sorbitol dehydrogenase (SOR1). Notable exceptions to these trends included Mig1p, a SBTF only in rich media but which bound targets in the ‘stress and rich media’ cluster; Yap6p, which bound different sets of subtelomeric genes in low versus high levels of hydrogen peroxide; and three SBTFs (Aft2p, Mal33p, and Yjl206c) that did not strongly cluster with other factors.

**SBTFs do not behave like known telomere-binding complexes or transcription factors but do interact with these proteins**

The telomere, as opposed to the subtelomere, is the target of several extensively studied protein complexes. The function of these complexes includes telomere replication, chromosome end protection, and transcriptional silencing.

To assess whether SBTFs have been linked to any of these functions, we examined protein complexes\(^{114}\) (Supplemental Figure 3.3), recent literature reviews\(^{99,115}\), and results from a genetic screen for telomere length mutants\(^{116}\). None indicated that SBTFs play known roles near the telomere (Supplemental Table 2.1). However, we also analyzed all of the protein interactions from the BioGRID database\(^{117}\) and did find that more than 80% of SBTFs (18 of 22) were linked via either physical or genetic interactions to genes with established functions at the telomere (Figure 3.4; see Methods).

We next compared the SBTFs to the most extensively studied telomere-binding transcription factor, Rap1p. Rap1p binds to repeated sequences (C\(^{1-3}\)A) that are found at all chromosome ends\(^{118}\) and plays roles in telomere silencing and length control\(^{115}\). We note that in our computational screen Rap1p was identified among the top subtelomere binding factors (\(P=0.011\); Table 2.1) but did not meet our stringent \(P\)-value cutoff. This is consistent with previous reports that found >90% of the target
Figure 3.4. Interactions Between SBTFs and Telomere-related Genes.

BioGRID\textsuperscript{117} interactions that connect SBTFs to any telomere-related gene listed in Supplemental Table 3.1. Also shown are all interactions between these telomere-related genes. Genes that function at the telomere, but do not interact with a SBTF based on data from BioGRID, are not shown (e.g. \textit{SIR2} or \textit{YKU70}). Ovals indicate that a gene has been associated with the telomere in recent literature reviews\textsuperscript{99,115}. Rectangles indicate annotation from the GO database or high-throughput experiments. Interaction types are described in detail in the Methods.
TFs (22/203) display a binding preference for subtelomeric genes, which themselves only comprise about 6% of the genome.

**SBTFs regulate stress and carbon metabolism in three broad clusters**

Roughly one third of the SBTFs we identified had been previously associated with cellular stress responses (e.g., Msn4p, Pdr1p, Phd1p, Rox1p, Xbp1p, Yap5p, Yap6p — see Table 2.1). SBTFs also included factors mediating glucose and nitrogen repression (Mig1p, Dal80p, Gzf3p, Nrg1p, Uga3p) as well as growth on alternative carbon sources (Gal4p, Mal33p, Hap4p) and metal uptake (Aft2p, Cup9p).

In contrast, TFs associated with most other cellular programs, such as the cell-cycle regulator Ace2p (Figure 3.1E), had TDPs that closely matched the background distribution (Figure 3.1H). Five SBTFs (Gat3p, Dat1p, Rgm1p, Yjl206c, Ypr196w) were poorly characterized or of unknown function. Several SBTFs such as Gal4p, Msn4p, and Pdr1p are among the most well studied yeast TFs. Their identification as subtelomeric binding factors is, to our knowledge, novel, but it is in concordance with their known roles in regulating stress or metabolic genes.

Hierarchical clustering of SBTF binding profiles indicated that SBTFs bound at least three distinct classes of subtelomeric genes (Figure 3.3B and Supplemental Figure 3.2). Stress-only SBTFs largely targeted alcohol dehydrogenases (AAD3 and ADH7) along with YRF genes, a family of putative helicases located within the subtelomeric Y’ repeated element which may function in telomere maintenance when telomerase is absent. Rich-media SBTFs bound YRF genes as well as members of the COS gene family, which are widely conserved and may function in salt resistance and the unfolded protein response, but are otherwise generally uncharacterized.

Intriguingly, SBTFs identified in both stress and rich media conditions
sequences bound by Rap1p lie outside the subtelomere in the promoters of many essential genes.

Nonetheless, we reasoned that if SBTFs are binding to repeated DNA elements in a manner analogous to Rap1p, then SBTF binding should be correlated with the presence of these elements. To check for coincidence with known repeated elements, we correlated SBTF binding with not only Rap1 binding sites but also the presence of all ten non-ORF sequence features annotated by the Saccharomyces Genome Database\textsuperscript{119}, which include the X, Y’, long terminal repeat, and autonomic replicating sequence elements (Supplemental Table 3.2). The binding profiles for two SBTFs, Yap5p and Msn4p, were correlated with the Y’ element (Pearson $P<0.05$ with $r = 0.75$ and $r = 0.73$, respectively) but no other significant correlations were found.

Because subtelomeric DNA is repetitive, we also considered the possibility that a single subtelomeric sequence bound by a SBTF might generate false-positive binding hits due to microarray cross-hybridization to other subtelomeric sequences. We expect that the impact of this effect is likely to be small, since binding was not correlated with the presence of known repeats.

We tested this possibility by applying TDP analyses to a filtered dataset in which binding targets with similar promoter sequences had been removed\textsuperscript{70}. Most SBTFs remained significant in this data set (15 of 22). Furthermore, the fact that most SBTFs have subtelomeric binding patterns that change dynamically depending on physiological state (Figure 3.3A) does not support an explanation whereby subtelomeric binding is predominantly an artifact of microarray cross-hybridization.

**Six SBTFs target heterochromatin domains that are derepressed during growth on alternative carbon sources**

Next, we sought to explore whether SBTFs might be associated with Hda1p, the only known regulator that specifically targets the subtelomere. Hda1p is a histone
deacetylase that establishes \textit{HDA1} associated subtelomeric (HAST) domains that repress about 40\% of all subtelomeric genes\textsuperscript{103}. It is plausible that SBTFs could help establish, maintain, or relieve repression of genes in HAST domains.

We examined the clusters of subtelomeric genes bound by SBTFs (Figure 3.3B) and found a significant number of HAST genes in cluster ‘SR’, the stress responsive genes bound by Mig1p, Nrg1p, Phd1p, and Yap6p in hydrogen peroxide, butanol, and rich conditions (Figure 3.5A; binomial test, $P<0.01$). These same HAST genes were also bound by Xbp1p in hydrogen peroxide. Interestingly, the other SBTF in the SR cluster, Yjl206cp, also bound HAST genes, albeit a different set.

We found that the subtelomeric genes bound by TFs in cluster SR were specifically upregulated in an \textit{hda1Δ} deletion strain\textsuperscript{120} (Figures 5B-D; KS test, $P<0.001$), consistent with Hda1p’s function as a repressor. However, SBTFs in this cluster did not appear to be required for establishing or maintaining repression because individually deleting each TF\textsuperscript{121} does not affect the expression of many subtelomeric targets in rich medium (Supplemental Figure 3.4).

Taken together, these findings suggest that SBTFs in cluster SR perhaps do not function at the subtelomere by differential binding but instead are modulated by other regulatory mechanisms such as post-translational modification. Alternatively, it is possible that they do function by differential binding but only in conditions other than those profiled to-date.

Since genes in HAST domains function during growth on alternative carbon sources\textsuperscript{103}, we tested the SBTF deletion strains \textit{yap6Δ}, \textit{phd1Δ}, \textit{nrg1Δ}, and \textit{yjl206cΔ} for growth defects on solid media supplemented with glucose, fructose, galactose, lactose, ethanol, maltose, or sucrose using a series of dilution assays. Although growth defects were not observed for any single mutant, a \textit{yap6Δ phd1Δ} double mutant exhibited a clear growth defect in galactose, ethanol, and glycerol (Figure 3.5E).
Figure 3.5. Binding Profiles, Expression, and Growth Phenotypes Link SBTFs to Hda1p.

(A) SBTFs that have a significant percentage of targets in HAST domains (P<0.01). Exact numbers of targets are above each bar. (B,C,D) Box-and-whisker plots of gene expression changes in an *hda1Δ* strain compared to wild type [data from 114] for targets bound by: (B) Yap6p (P=10^{-8}). (C) Phd1p (P=0.00006). (D) Yap1p (statistically unaffected). (E) *YAP6* and *PHD1* display a condition-specific genetic interaction during growth on non-glucose carbon sources. Five-fold serial dilutions were grown on synthetic complete (SC) medium supplemented with 2% of one of the following carbon sources: glucose, galactose, ethanol, or glycerol. Controls are BY4741, the parent strain used to create gene deletions, and gal4Δ, that does not grow on galactose.
This interaction likely depends on subtelomeric genes, as Yap6p and Phd1p bind seven common subtelomeric targets including the hexose transporter-like genes *HXT9*, *HXT15*, and *HXT16*. The gene *HXT15* was previously found to have higher expression levels during growth on ethanol and glycerol\(^{122}\).

In summary, we have linked six SBTFs to the chromatin-modifying enzyme Hda1p, which mediates Sir-independent silencing at the subtelomere. Two of these SBTFs, Yap6p and Phd1p, jointly contribute to growth on alternative carbon sources, a role consistent with existing models of Hda1p function.

**Dynamic subtelomeric binding is correlated with gene expression**

Given that subtelomeric genes are typically repressed\(^{105}\), a SBTF that preferentially binds the subtelomere in a stress condition (cluster S in Figure 3.3B) might function as an activator if its targets are upregulated in the same stress. Alternatively, a SBTF that moves away from the subtelomere in response to stress (cluster R) might function as a repressor if its targets are subsequently upregulated in the same condition.

To distinguish between these hypotheses, we analyzed published gene expression profiles\(^{76}\) to identify SBTFs whose subtelomeric target genes had condition-specific regulatory behaviors\(^{123,124}\).

We discovered a strong correlation between the binding and upregulation of targets of Aft2p in response to oxidative stress. This correlation was particularly striking for the 11 members of the subtelomeric *PAU* gene family that are bound by Aft2p under mild hydrogen peroxide treatment (Figure 3.6A), although the profiles of some family members might be affected by microarray cross-hybridization with other *PAU*-family genes since these sequences are over 80% similar to *PAU1* at the nucleotide level. *PAU* genes and Aft2p have each been implicated in oxidative stress
Figure 3.6. Dynamic Binding and Expression Profiles Suggest Models of SBTF Function and Link SBTFs to Poorly Characterized Genes at Subtelomeres.

TF binding profiles were matched to expression profiles gathered under similar environmental perturbations. (A) Aft2p binds upstream of 12 subtelomeric PAU genes under oxidative stress conditions (blue boxes). Heatmap shows induction of the PAU genes in oxidative stress conditions. (B) In the presence of rapamycin, which simulates nitrogen depletion by antagonizing the TOR kinases, Gzf3p, Uga3p, and Dal80p bind upstream of genes that are induced under conditions of nitrogen limitation. (C) Genes targeted by Nrg1p, Yap6p, and/or Rox1p under hydrogen peroxide but not in untreated conditions are upregulated in response to three oxidative stress agents (hydrogen peroxide, menadione, and diamide). For A-C, the asterisk (*) indicates the approximate time that the binding data were collected. (D) SBTFs that display a preference for the subtelomere only in stress conditions may positively regulate gene expression. (E) SBTFs that bind the subtelomere in rich media conditions perhaps either contribute to repression or are ‘poised for action,’ subsequently functioning in different stress conditions (see text).
resistance\textsuperscript{125,126}, but our results provide the first evidence that Aft2p upregulates \textit{PAU} genes.

The other stress-only (S) and rich-media-only (R) SBTFs bound two overlapping sets of genes (Figure 3.3B). These included 30 genes similar to the \textit{YRF} family that we found to be largely unresponsive in stress conditions (Supplemental Figure 3.5). However, excluding the \textit{YRF}-like genes, three stress-dependent SBTFs (Gzf3p, Uga3p, Dal80p) displayed behaviors consistent with functions as activators in conditions of nitrogen depletion (Figure 3.6B).

The expression analysis also identified a set of subtelomeric genes that were targeted by Yap6p, Nrg1p, or Rox1p under hydrogen peroxide but not in untreated conditions and which were upregulated in response to a broad array of oxidative stress agents (Figure 3.6C). For these targets, these three SBTFs appear to behave like ‘stress only’ SBTFs, although Yap6p and Nrg1p had been classified as ‘stress-and-rich-media’ factors based on their global binding patterns in Figure 3.3a. Intriguingly, many of the putative stress-regulated targets of these SBTFs are poorly characterized and unnamed, although \textit{YML131W} is similar in sequence to oxidoreductases\textsuperscript{119}. In contrast to the above evidence for SBTFs as transcriptional activators, we did not find any strong evidence for SBTFs as repressors of gene expression.

**Discussion**

We have identified 22 yeast transcription factors that display a clear binding preference for subtelomeric regions (Figure 3.1 and Table 3.1). Almost all of these factors physically or genetically interact with genes encoding proteins with known telomeric functions (Figure 3.4). The majority of binding patterns are dynamic, such that the factor is concentrated at the subtelomere only under certain conditions (Figures 3.2 and 3.3). This observation, combined with our analysis of stress-induced
expression profiles, suggests several models for how SBTFs may function. Most stress-only SBTFs (cluster S) appear to be activators of subtelomeric genes in response to stress (modeled in Figure 3.6D) and may relocalize to different target genes under other conditions.

The role of rich-medium SBTFs (clusters R and SR) is less clear. We initially hypothesized that these might function as repressors of genes that are derepressed under specific stress conditions (modeled in Figure 3.6E). We did not find such evidence of repressor function, although it remains possible that derepression occurs in conditions different than those for which data are currently available. In particular, many targets in the SR cluster function in alternative carbon metabolism, and we did find a combination of SBTFs that bind this cluster and promote growth in non-glucose conditions (Figure 3.5).

We also considered that a SBTF might be neither an activator nor a repressor. Instead, it may be sequestered to subtelomeres without effecting changes in subtelomeric gene expression as a means of holding it in reserve from other genomic locations. Additionally, there may be roles for SBTFs at the telomere that are dependent on sequestration as suggested for Rap1p\textsuperscript{127,128} and shown for the telomeric Ku proteins that become mobilized to sites of DNA damage to facilitate repair\textsuperscript{129}.

A growing body of evidence suggests that genome organization, nuclear architecture, and gene expression are interrelated\textsuperscript{101,130,131}. Thus, an intriguing alternative possibility is that rich-media SBTFs may be residing in repressive regions within the nucleus. Foci at the nuclear periphery contain telomeres\textsuperscript{132} along with other regions of silent heterochromatin\textsuperscript{130,133}.

Transcription factors at the nuclear periphery are also located at sites of active transcription tethered to nuclear pore complexes\textsuperscript{101,130}. In a process termed ‘reverse recruitment’ genes to be transcribed appear to be recruited to these sites of
transcription. Since Rap1p has been implicated in this process\textsuperscript{134}, it is tempting to ascribe a similar role to SBTFs. Under this model, the condition-dependent binding seen in Figure 3.3 would be interpreted not as SBTFs themselves ‘moving towards’ or ‘away’ from the subtelomere, but rather as different subtelomeric genes moving in or out of sites of gene activation or repression at the nuclear periphery.

As more binding profiles are generated in yeast and other species, it may be fruitful to screen those data for unexpected binding preferences at the subtelomere as well as more generally at other genomic regions. For instance, such a strategy has been used to show that Sgo1p and Rec8p, two cohesin proteins that function in chromosome segregation, localize to a 50 kb region around the centromere\textsuperscript{135}. Our work demonstrates that many other such analyses may be productive, and it provides an example of transcription factor binding patterns that reflect genome organization.

Further work will be required to determine the extent to which condition-specific targeting of the subtelomere is a conserved regulatory strategy. Orthologs for all SBTFs have been identified in one or more other yeast species, and seven SBTFs appear to be widely conserved across fungi, invertebrates, fish, and mammals (Supplemental Table 3.3). Moreover, SBTFs that cluster together by binding profile (Figure 3.3B) also appear to have roughly similar patterns of conservation across species.

Intriguingly, paralogs of most SBTFs were themselves not found to be SBTFs, with the exception of Dal80p / Gzf3p. This suggests there may be selective pressure that maintains the subtelomeric binding pattern of one paralog, but that redundancy is typically not necessary. It is tempting to speculate that SBTFs and their dynamic localization may contribute to the evolutionary plasticity that has previously only been attributed to the subtelomeric genes themselves\textsuperscript{110,136}. 

Methods

Computational screen for SBTFs

SBTFs were screened from published genome-wide transcription factor (TF) binding data. A \( P \)-value threshold of 0.001 was used to identify the set of gene promoters putatively bound by each TF in a particular environmental condition (either rich medium or one of 12 additional stress or nutritional conditions).

To compute the distance from each gene to the closest telomere, we first computed \( \text{mid}_g \), the midpoint of the starting and ending chromosomal coordinates for gene \( g \), which were downloaded from the Saccharomyces Genome Database (http://www.yeastgenome.org). Then, the distance to the closest telomere is \( d_g = \min(\text{mid}_g, \text{LengthChr}_g - \text{mid}_g) \), where \( \text{LengthChr}_g \) is the length of the chromosome on which \( g \) is located.

The distribution of distances for all genes targeted by a TF in a particular binding experiment was defined to be the Telomere Distance Profile (TDP). Using R (http://www.r-project.org), a one-sided Kolmogorov-Smirnov (KS) test was used to compare the TDP for each binding experiment to a background distribution of the TDP for all yeast genes. To estimate the false discovery rate (FDR), the set of all KS \( P \)-values was used as input to the Q-value software.

Statistically significant TDPs were identified at \( P \leq 0.001 \), corresponding to a FDR of approximately 1%. Subtelomeric genes were defined to be those with \( d_g \leq 25000 \). Hierarchical clustering of SBTF binding profiles was performed using Cluster 3.0 and visualized using Java TreeView. To simplify data visualization, clustering was limited to subtelomeric genes bound by at least one SBTF.
Analysis of protein complexes and interactions involving SBTFs

A comprehensive set of yeast protein-protein interactions was obtained from the BioGRID database\textsuperscript{117} version 2.0.40 (May 2008). Interactions that connected a SBTF to any known telomere-related gene were identified, and then manually inspected using Cytoscape\textsuperscript{11}. Physical interactions were classified as those in BioGRID labeled: Affinity Capture, Co-crystal Structure, Co-fractionation, Co-purification, FRET, Far Western, Protein-peptide, Protein-RNA, Reconstituted Complex, or Two-hybrid. Genetic interactions were classified as those in BioGRID labeled: Dosage Growth Defect, Dosage Lethality, Dosage Rescue, Phenotypic Enhancement, Phenotypic Suppression, Synthetic Growth Defect, Synthetic Lethality, or Synthetic Rescue.

A compendium of 547 non-overlapping yeast protein complexes\textsuperscript{114} was checked to see if any complex contained one or more SBTFs and any of the known telomere-related genes listed in Supplemental Table 3.1. Telomere-related genes were obtained from two sources: 1) any gene mentioned in recent reviews of yeast telomeres\textsuperscript{115} or the telomere position effect\textsuperscript{99}, or discovered in genetic screens for telomere length mutants\textsuperscript{116}, and 2) any gene annotated in the Gene Ontology (GO) database under the categories “telomere organization and biogenesis” (GO:0032200) or “chromatin-silencing-at-telomere” (GO:0006348). In total, 426 telomere-related genes were identified, 321 from GO analyses and an additional 105 from literature reviews.

Analysis of repetitive element binding by SBTFs

Yeast genome sequence features were downloaded from the Saccharomyces Genome Database (http://www.yeastgenome.org). All features labeled Open Reading
Frame (ORF) were discarded, leaving ten different types of non-ORF features. Next, we determined the presence or absence of one or more occurrences of each feature on the 32 *S. cerevisiae* chromosome arms. We then determined whether each SBTF bound at least one promoter on the 32 chromosome arms. These analyses generated two sets of vectors consisting of 32 elements each: one set corresponding to SBTF binding and another set corresponding to sequence features. Using the Pearson correlation test in R, the similarity of each of the binding vectors was compared to each of the ten non-ORF feature vectors. *P*-values were adjusted for multiple hypotheses testing using the Bonferroni correction.

**Screening SBTF binding profiles for correlation with *hda1Δ* sensitivity**

We used a method by Steinfeld and colleagues to identify transcription factor (TF) and chromatin modifying enzyme (CM) pairs that function in concert. Briefly, gene expression profiles of CM deletions are analyzed to determine whether the targets of a TF, identified by ChIP-chip, are preferentially affected. The rationale being that if a TF and CM function together, then deletion of the CM should preferentially affect the genes targeted by that TF.

We extended this approach by reasoning that if a TF and CM function together to specifically regulate the subtelomeric targets of a TF, then those subtelomeric targets should be preferentially expressed compared to the non-subtelomeric targets in a CM deletion. We applied this approach to *hda1Δ* expression data and the sets of subtelomeric and non-subtelomeric genes targeted by each SBTF. As in Steinfeld, we used the Kolmogorov-Smirnov test to assess the statistical significance of the difference in expression between the bound vs. unbound subtelomeric targets for each SBTF. The significance threshold was a Bonferroni-corrected *P*-value less than 0.001.
Dilution assays for growth of SBTF deletions on alternative carbon sources

Single deletion yeast strains were obtained from the Yeast Deletion Collection (Open Biosystems). Double deletions were a gift from S. Bandypadhay in the Ideker lab, and were constructed as described previously. Individual colonies growing on YPD agar plates were picked and cultured overnight to saturation in 2 mL YPD. Overnight cultures were diluted with sterile water in a 96-well microtiter plate in six 5-fold serial dilutions with a starting OD between 2 and 3. A 48-pin replica pinner was used to spot approximately 5 uL of each dilution onto solid agar plates of synthetic complete medium containing 2% of one of the following carbon sources: glucose, raffinose, glycerol, ethanol, galactose, fructose, maltose, or lactose. Plates were incubated at 30 C for at least three days. Initial screening results were confirmed with repeated experiments.

Inferring SBTF modes of action by integrated analysis of matched binding and expression profiles

Gene expression experiments from Gasch and colleagues were matched to TF binding experiments performed under similar environmental perturbations. The set of gene targets bound by SBTFs in the presence of hydrogen peroxide (H\textsubscript{2}O\textsubscript{2}) was matched to expression perturbations caused by the oxidative agents menadione, DTT, diamide, or hydrogen peroxide. Binding profiles from cells treated with rapamycin were matched to expression profiles labeled ‘Nitrogen depletion’. Rich-media binding profiles were matched to expression profiles labeled ‘YPD’, ‘diauxic shift’ and ‘steady state expression’.

Matched sets of binding and expression that passed both of the following criteria were identified for in-depth manual inspection: 1) the expression of the subtelomeric targets bound by each SBTF was different than the unbound
subtelomeric genes, as assessed by the Kolmogorov-Smirnov test at P<0.05. 2) the SBTF bound the differentially expressed targets only in stress conditions, or the SBTF bound these genes only in rich medium. To identify YRF-like genes, the protein sequence encoded by \textit{YRF1-1} was used as an input to BLASTP, and matches were selected at an E-value cutoff of $10^{-10}$.

**Identification of SBTF orthologs**

Fungal orthologs of each \textit{S. cerevisiae} gene were extracted from the ‘Pillars. tab’ file downloaded from the Yeast Gene Order Browser version 2.0\textsuperscript{141} (http://wolfe.gen.tcd.ie/ygob/data/Version1.0_Nature-2006/Pillars.tab). Orthologs in other species were extracted from raw datafiles downloaded from version 6.0 of InParanoid\textsuperscript{142} (http://inparanoid.sbc.su.se/download/old_versions/6.0/sqltables/). The reported number of orthologs for an ORF in another species is the total number of distinct orthologs for that ORF and for the \textit{S. cerevisiae} paralogs of that ORF defined in the YGOB.
### Table 3.1. Summary of Subtelomere Binding Transcription Factors

<table>
<thead>
<tr>
<th>Genes bound #</th>
<th>Saccharomyces Genome Database annotation</th>
<th>HAST</th>
<th>COS</th>
<th>PAU</th>
<th>YRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Transcriptional activator related to Msn2p; activated in stress conditions, which results in translocation from the cytoplasm to the nucleus; binds DNA at stress response elements of responsive genes, inducing gene expression.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Zinc cluster protein that is a master regulator involved in recruiting other zinc cluster proteins to pleiotropic drug response elements (PDRs) to fine tune the regulation of multidrug resistance genes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Transcriptional repressor that binds to promoter sequences of the cyclin genes, CYCS, and SMF2; expression is induced by stress or starvation during mitosis, and late in meiosis; member of the Styx/Mbp1 family; potential CdC28b substrate.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>DNA repair and TFIIH regulator, required for both nucleotide excision repair (NER) and RNA polymerase II (RNAP II) transcription; involved in telomere maintenance.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Transcriptional repressor necessary for gamma-amino-butyrate (GABA)-dependent induction of GABA genes (such as UGA1, UGA2, UGA4); zinc-finger transcription factor of the Zn(2)-Cys(6) binuclear cluster domain type; localized to the nucleus.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Transcriptional activator necessary for gamma-aminobutyrate (GABA)-dependent induction of GABA genes (such as UGA1, UGA2, UGA4); zinc-finger transcription factor of the Zn(2)-Cys(6) binuclear cluster domain type; localized to the nucleus.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>DNA-binding transcription factor required for the activation of the GAL genes in response to galactose; repressed by Gal80p and activated by Gal3p.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>MAL-activator protein, part of complex locus MAL3; nonfunctional in generically reference strain S288C.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Subunit of the heme-activated, glucose-repressed Hap2p/3p/4p/5p CCAT-binding complex, a transcriptional activator and global regulator of respiratory gene expression; provides the principal activation function of the complex.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Iron-regulated transcriptional activator; activates genes involved in intracellular iron use and required for iron homeostasis and resistance to oxidative stress; similar to Mt18p.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>DNA-binding protein that recognizes oligo(A) oligo(E) tracts; Arg side chain in its N-terminal pentad Gly-Arg-Lys-Pro-Gly repeat is required for DNA-binding; not essential for viability.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Putative transcriptional repressor with proline-rich zinc fingers; overproduction impairs cell growth.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Putative transcriptional activator related to Msn2p; activated in stress conditions, which results in translocation from the cytoplasm to the nucleus; binds DNA at stress response elements of responsive genes, inducing gene expression.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Putative transcriptional activator related to Msn2p; activated in stress conditions, which results in translocation from the cytoplasm to the nucleus; binds DNA at stress response elements of responsive genes, inducing gene expression.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Nuclear protein (putative).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>DNA-binding protein involved in either activation or repression of transcription, depending on binding site context; also binds telomere sequences and plays a role in telomeric position effect (siling) and telomere structure.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Met18p and Rap1p did not meet the stringent SBTF P-value cutoff, but are included here for reference because they play known roles in telomere biology.
Supplemental Figures

Supplemental Figure 3.1. Estimating the error in identifying SBTFs.

The Q-value software package\textsuperscript{108} was used to account for multiple-hypothesis testing by estimating the false discovery rate for the telomere distance profile tests (see Figure 3.1 and text). (A) P-values plotted against their corresponding q-values. As indicated by the red dashed lines, the P-value threshold of 0.001 used in this study corresponds to a q-value (false discovery rate) of approximately 0.01, meaning that about 1% of the TF binding profiles with a significant subtelomeric preference are expected to be false positives. (B) At $P \leq 0.001$, 31 tests are considered significant. (C) Based on the q-value method, less than one (~0.3) of the 31 TF binding profiles with subtelomeric bias is expected to be a false positive.
Supplemental Figure 3.2. A map of the subtelomeric regulatory circuitry.

All TF-promoter binding interactions from significant SBTF binding profiles are visualized with Cytoscape\textsuperscript{11}. See Legend for details.
Supplemental Figure 3.3. Analysis of protein complexes.

(A) Protein complexes that contain SBTFs. (B) Examples of protein complexes identified in the same screen that contain telomere-related proteins identified in recent reviews of yeast telomeres.
Supplemental Figure 3.4. Screening expression profiles of single gene deletions for enrichment of subtelomeric genes.

Gene expression profiles of single gene deletions\textsuperscript{29,121} were analyzed with the same Telomere Distance Profile analysis method used to identify SBTFs. The set of genes differentially expressed in each deletion mutant was identified using a $P$-value cutoff of 0.001, as described in the original publications. Bars show the number of subtelomeric genes affected by each TF deletion. Grey indicates deletions that affect an unexpectedly large number of subtelomeric genes (KS-test, $P<0.001$). Black bars show data for the 15 rich-media SBTFs. No deletions of the stress condition SBTFs were available in these data sets. Numbers next to each bar indicate the total number of genes differentially expressed in a deletion.
Supplemental Figure 3.5. Expression responses of subtelomeric genes to environmental perturbations.

Expression data are from Gasch and colleagues\(^7\). Subtelomeric genes (columns) are aligned with clustered binding patterns from Figure 3.3b. Unnamed stress responsive ORFs from Figure 3.6 (panels b and c) are indicated by maroon ticks and include \(YNR068C\), \(YML131W\), \(YMR315W\), \(YCR102C\), and \(YLR460C\).
Supplemental Tables

Supplemental Table 3.1. List of telomere-related genes curated from the literature.

| ABF1 | CUP9 | HAT2 | MAK3 | PFD1 | RP89 | SIF2 | TEN1 | YEL057C |
| ADE12 | CYC8 | HCH1 | MAK31 | PGD1 | RP80 | SIN3 | THP2 | YGL039W |
| ADO1 | DALT80 | HCM1 | MAL33 | PHD1 | RPL12B | SIR1 | TFI1 | YGR042W |
| AFT2 | DAT1 | HCR1 | MCM10 | PHO80 | RPL13B | SIR2 | TLC1 | YHL012W |
| AGP2 | DCC1 | HDA2 | MCM5 | PHO85 | RPL16 | SIR3 | TOM5 | YL042C |
| AHC2 | DEP1 | HEK1 | MDM10 | PHO87 | RPL34B | SIR4 | TOP3 | YL206C |
| APE3 | DEP1 | HEK2 | MEC1 | PHO88 | RPN4 | SIT4 | TPD3 | YKU70 |
| APN | DIG1 | HFI1 | MEC3 | PIF1 | RPP1A | SIW14 | TRK1 | YKU80 |
| ARD1 | DLS1 | HFF1 | MED1 | PKC1 | RPS10A | SLA2 | UBP10 | YML035C-A |
| ARF1 | DNA2 | HFF2 | MET18 | PKP1 | RPS11B | SLL15 | UGA3 | YMR031W-A |
| ARG2 | DQA4 | HHT1 | MET7 | PHT3 | RPS14A | SLT2 | UG01 | YMR269W |
| ARV1 | DOT1 | HHT2 | MFT1 | PNC1 | RPS16A | SLX5 | UFP1 | YOL138C |
| ASC1 | DOTS | HFI1 | MIG1 | POC4 | RPS16B | SLX8 | UFP2 | YOR008C-A |
| ASF1 | DOT6 | HIR1 | MKK1 | POL1 | RPS17A | SM1 | UFP3 | YOR1 |
| ATC1 | DPB3 | HIR2 | MLH1 | POL12 | RPS17B | SNC2 | URE2 | YOR222C |
| ATG11 | DPB4 | HIT1 | MMM1 | POL2 | RPS18A | SNF2 | VAM6 | YPL041C |
| ATG17 | DUN1 | HMO1 | MMS19 | POL30 | RPS19A | SNF7 | VAM7 | YPL105C |
| BCK1 | EAP1 | HPR1 | MOT2 | POL32 | RPS21A | SNF8 | VPS15 | YPL144W |
| BDF1 | EBS1 | HSC82 | MOT3 | PPE1 | RPS22A | SOH1 | VPS18 | YPL205C |
| BFM2 | ELG1 | HSP104 | MRC1 | PRO1 | RPS23A | SOL2 | VPS22 | YPR196W |
| BFM4 | EPL1 | HSP82 | MRE11 | PRS3 | RPS23B | SPP1 | VPS23 | YPT7 |
| BRE1 | ERG2 | HST1 | MM2 | PRS5 | RPS27B | SPS100 | VPS25 | YRB2 |
| BRE2 | ERJ5 | HST2 | MTR38 | PTC1 | RPS28B | SPT21 | VPS28 | YRF1-1 |
| BRO1 | ESC1 | HST3 | MLP44 | PXR1 | RPS30B | SPT4 | VPS3 | YRF1-2 |
| BUD16 | ESC8 | HST4 | MRT4 | RAD27 | RPS4A | SRB2 | VPS32 | YRF1-3 |
| BUD21 | EST1 | HTA1 | MSN4 | RAD5 | RPS4B | SRB5 | VPS34 | YRF1-4 |
| BUD23 | EST2 | HTA2 | NAM7 | RAD50 | RFT4 | SRB8 | VPS36 | YRF1-5 |
| BUD30 | EST3 | HTL1 | NAT1 | RAD51 | RFT6 | SRE1 | VPS39 | YRF1-6 |
| BUD32 | EXO1 | HTZ1 | NAT3 | RAD52 | RR13 | SSH1 | VPS32 | YRF1-7 |
| CAC1 | FMP26 | HUR1 | NFI1 | RAD54 | RRP8 | SSN2 | VPS34 | YSP3 |
| CAC2 | FUS3 | IES3 | NMD2 | RAD57 | RSA1 | SSN3 | VPS65 | YTA7 |
| CAC3 | FYV12 | JPH1 | NPL6 | RAD59 | RSC2 | SSN8 | VPS75 | ZDS1 |
| CAX4 | FYV4 | IMG2 | NPT1 | RAD6 | RFT1 | STE11 | VPS9 | ZDS2 |
| CBC2 | FYV6 | ISA1 | NRG1 | RAP1 | RTT106 | STE12 | WHI2 | ZEO1 |
| CCW14 | GAL11 | JSW2 | RNR1 | REF2 | SAP30 | STE7 | WTM1 | |
| CDC13 | GAL4 | JTC1 | NUP2 | RA1 | SA1 | STM1 | WTM2 | |
| CDC45 | GAT3 | KAI1 | NUP60 | RAF2 | SA3 | STM1 | WTM3 | |
| CDC46 | GBP2 | KEM1 | NUT1 | RAF3 | SA4 | STO1 | XBP1 | |
| CDC6 | GC3 | KRE21 | OGG1 | RFM1 | SA5 | STP22 | XDJ1 | |
| CDC7 | GLO4 | KRE28 | OPI1 | RGM1 | SBA1 | SUB2 | XRS2 | |
| CDC73 | GON7 | LCD1 | ORC2 | RIF1 | SCP160 | SUM1 | YAF9 | |
| CDC1 | GPB2 | LDB19 | ORC5 | RIF2 | SCS2 | SUR4 | YAP5 | |
| CGL121 | GTR1 | LDB7 | PAF1 | RKR1 | SCS22 | SWA2 | YAP6 | |
| CHO2 | GTR2 | LEA1 | PB2 | RNA1 | SDC1 | SWD1 | YBR284W | |
| CSM1 | GUP1 | LEO1 | PCNA | RNH35 | SEM1 | SWD2 | YDJ1 | |
| CSR2 | GUP2 | LRP1 | PDC1 | RNR1 | SET1 | SWD3 | YDL118W | |
| CST6 | GZF3 | LSM7 | PDR1 | ROX1 | SGS1 | TAT2 | YDR115W | |
| CTF8 | HAP4 | LST7 | POX3 | RPA14 | SHG1 | TEL1 | YDR532C | |
| CTK1 | HAT1 | MAK10 | PEP3 | RPB4 | SHP1 | TEL2 | YEL033W | |
Supplemental Table 3.2. Correlation of binding profiles with the presence of genomic features.

<table>
<thead>
<tr>
<th>SBTF</th>
<th>ARS</th>
<th>ORF</th>
<th>X_element_combinatorial_repeats</th>
<th>X_element_core_sequence</th>
<th>Y_element</th>
<th>long_terminal_repeat</th>
<th>multigenic_locus</th>
<th>pseudogene</th>
<th>retrotransposon</th>
<th>RNA</th>
<th>telomere</th>
<th>telomeric_repeat</th>
<th>transposable_element_gene</th>
<th>log10 P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAL4</td>
<td>1.16</td>
<td>NA</td>
<td>1.80 NA</td>
<td>0.27</td>
<td>2.16</td>
<td>0.69</td>
<td>1.70</td>
<td>1.99</td>
<td>1.70 NA</td>
<td>0.76</td>
<td>2.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAT3</td>
<td>1.99</td>
<td>NA</td>
<td>2.29 NA</td>
<td>-0.82</td>
<td>1.98</td>
<td>2.04</td>
<td>2.10</td>
<td>2.29</td>
<td>2.29 NA</td>
<td>1.16</td>
<td>1.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHD1</td>
<td>0.63</td>
<td>NA</td>
<td>1.90 NA</td>
<td>1.81</td>
<td>2.18</td>
<td>1.66</td>
<td>2.11</td>
<td>1.82</td>
<td>2.05 NA</td>
<td>1.90</td>
<td>2.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDR1</td>
<td>1.17</td>
<td>NA</td>
<td>2.00 NA</td>
<td>-0.37</td>
<td>2.21</td>
<td>1.83</td>
<td>1.84</td>
<td>2.19</td>
<td>1.20 NA</td>
<td>1.49</td>
<td>1.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAP5</td>
<td>2.16</td>
<td>NA</td>
<td>2.24 NA</td>
<td>-3.80</td>
<td>2.17</td>
<td>2.09</td>
<td>2.16</td>
<td>2.26</td>
<td>1.55 NA</td>
<td>0.27</td>
<td>1.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAT1</td>
<td>1.62</td>
<td>NA</td>
<td>1.73 NA</td>
<td>1.01</td>
<td>2.27</td>
<td>1.84</td>
<td>2.25</td>
<td>1.95</td>
<td>1.82 NA</td>
<td>1.12</td>
<td>2.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CUP9</td>
<td>1.15</td>
<td>NA</td>
<td>1.72 NA</td>
<td>1.98</td>
<td>1.35</td>
<td>2.05</td>
<td>1.94</td>
<td>2.10</td>
<td>2.10 NA</td>
<td>2.00</td>
<td>2.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSN4</td>
<td>1.48</td>
<td>NA</td>
<td>2.06 NA</td>
<td>-3.33</td>
<td>1.86</td>
<td>2.23</td>
<td>2.25</td>
<td>1.73</td>
<td>2.15 NA</td>
<td>1.53</td>
<td>1.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAP4</td>
<td>1.00</td>
<td>NA</td>
<td>1.57 NA</td>
<td>0.29</td>
<td>1.97</td>
<td>2.21</td>
<td>1.24</td>
<td>1.98</td>
<td>1.87 NA</td>
<td>2.17</td>
<td>2.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGM1</td>
<td>0.81</td>
<td>NA</td>
<td>1.93 NA</td>
<td>0.95</td>
<td>1.93</td>
<td>2.00</td>
<td>0.94</td>
<td>2.06</td>
<td>2.06 NA</td>
<td>2.16</td>
<td>2.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTG1</td>
<td>1.87</td>
<td>NA</td>
<td>2.16 NA</td>
<td>2.23</td>
<td>1.46</td>
<td>1.89</td>
<td>2.26</td>
<td>1.99</td>
<td>1.99 NA</td>
<td>2.24</td>
<td>2.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YIL206C</td>
<td>0.81</td>
<td>NA</td>
<td>1.93 NA</td>
<td>2.24</td>
<td>1.93</td>
<td>1.72</td>
<td>2.06</td>
<td>2.06</td>
<td>2.06 NA</td>
<td>2.16</td>
<td>2.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NRG1</td>
<td>1.76</td>
<td>NA</td>
<td>2.00 NA</td>
<td>2.01</td>
<td>2.09</td>
<td>1.44</td>
<td>1.80</td>
<td>1.67</td>
<td>1.67 NA</td>
<td>1.58</td>
<td>1.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YPR196W</td>
<td>2.04</td>
<td>NA</td>
<td>2.00 NA</td>
<td>1.26</td>
<td>1.91</td>
<td>2.05</td>
<td>1.94</td>
<td>2.10</td>
<td>2.10 NA</td>
<td>2.00</td>
<td>2.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAP6</td>
<td>2.14</td>
<td>NA</td>
<td>2.24 NA</td>
<td>1.52</td>
<td>1.97</td>
<td>1.72</td>
<td>2.02</td>
<td>1.87</td>
<td>1.87 NA</td>
<td>1.99</td>
<td>2.02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Pearson correlation**

<table>
<thead>
<tr>
<th>SBTF</th>
<th>GAL4</th>
<th>GAT3</th>
<th>PHD1</th>
<th>PDR1</th>
<th>YAP5</th>
<th>DAT1</th>
<th>CUP9</th>
<th>MSN4</th>
<th>HAP4</th>
<th>RGM1</th>
<th>MTG1</th>
<th>YIL206C</th>
<th>NRG1</th>
<th>YPR196W</th>
<th>YAP6</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAL4</td>
<td>-0.32</td>
<td>-0.13</td>
<td>0.40</td>
<td>-0.32</td>
<td>-0.06</td>
<td>-0.23</td>
<td>0.32</td>
<td>-0.26</td>
<td>-0.35</td>
<td>-0.38</td>
<td>-0.38</td>
<td>0.16</td>
<td>-0.38</td>
<td>-0.11</td>
<td>-0.07</td>
</tr>
<tr>
<td>GAT3</td>
<td></td>
<td>0.18</td>
<td>0.56</td>
<td>0.39</td>
<td>0.06</td>
<td>0.52</td>
<td>0.45</td>
<td>-0.10</td>
<td>0.04</td>
<td>0.36</td>
<td>0.14</td>
<td>0.36</td>
<td>0.36</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>PHD1</td>
<td></td>
<td></td>
<td>0.67</td>
<td>0.35</td>
<td>-0.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDR1</td>
<td></td>
<td></td>
<td></td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAP5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAT1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CUP9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSN4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>HAP4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>RGM1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTG1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YIL206C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NRG1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YPR196W</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAP6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cells marked in blue are significant at P < 0.05
Supplemental Table 3.3. Orthologs of SBTFs in other species.

<table>
<thead>
<tr>
<th>SBTF</th>
<th>Ohnolog #</th>
<th>Binding cluster (see Fig 3b)</th>
<th>Total orthologs</th>
<th>Yeast</th>
<th>Plant</th>
<th>Worm</th>
<th>Fly</th>
<th>Fish</th>
<th>Mammal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUP9</td>
<td>TOS8</td>
<td>-</td>
<td>44</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>RGM1</td>
<td>YPL320W</td>
<td>R</td>
<td>27</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MSN4</td>
<td>MSN2</td>
<td>R</td>
<td>26</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>NRG1</td>
<td>NRG2</td>
<td>SR</td>
<td>19</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MIG1</td>
<td>NIG1</td>
<td>SR</td>
<td>17</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PHYD1</td>
<td>SOK2</td>
<td>SR</td>
<td>18</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>RGRX1</td>
<td>-</td>
<td>S</td>
<td>14</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YAP6</td>
<td>CIN5</td>
<td>SR</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GZF3</td>
<td>DAL80</td>
<td>S</td>
<td>11</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DAL80</td>
<td>GZF3</td>
<td>S</td>
<td>11</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AFT2</td>
<td>AFT1</td>
<td>S</td>
<td>11</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YAP5</td>
<td>YAP7</td>
<td>R</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>XBP1</td>
<td>-</td>
<td>S</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PDR3</td>
<td>-</td>
<td>R</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HAP4</td>
<td>-</td>
<td>R</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>UGA3</td>
<td>-</td>
<td>S</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GAL4</td>
<td>-</td>
<td>R</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GAT1</td>
<td>-</td>
<td>R</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YUL256C</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MAL33</td>
<td>-</td>
<td>S</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YPR1195W</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

#  Ohnologs are defined to be paralogous genes (in this case in *S. cerevisiae*) that arose from a whole genome duplication event.\(^{141}\)

¶ Ortholog data from the Yeast Gene Order Brower\(^{141}\)

§ Ortholog data from InParanoid\(^{142}\) version 6.0
Acknowledgements

We are grateful to S. Bandyopadhyay for providing the double deletion strains used in this project; S. Jacobson for assisting with dilution assays; S. Bandyopadhyay, D. Gottschling, S. Jacobson, J. Karlseder, T. Ravasi, K. Tan, B. Trask, C. Workman, and C.H. Yeang for critical comments on the manuscript. This work was funded by grant ES14811 to T. Ideker from the NIEHS. T. Ideker is a David and Lucille Packard Fellow.

Chapter 3, in full, has been submitted for publication as it may appear in *Genome Research*, 2008, Mak, H. Craig; Pillus, Lorraine; Ideker, Trey. The dissertation author was the primary investigator and author of this paper.
Chapter 4: Refinement of a transcriptional regulatory network using graphical probabilistic models

Abstract

As genome-scale measurements lead to increasingly complex models of gene regulation, systematic approaches are needed to validate and refine these models. Towards this goal, we describe an automated procedure for prioritizing genetic perturbations to optimally discriminate among alternative models of a gene regulatory network. Using this procedure, we evaluate 38 candidate regulatory networks in yeast and perform four high priority gene knockout experiments. The refined networks support previously-unknown regulatory mechanisms downstream of SOK2 and SWI4.
Introduction

Recent advances in genomics and computational biology are enabling construction of large-scale models of gene regulatory networks. High-throughput technologies such as automated sequencing\(^{143}\), gene expression arrays\(^{144}\), chromatin immunoprecipitation\(^{24}\), and yeast two-hybrid assays\(^{21}\) each probe different aspects of the gene regulatory system through genome-wide data sets. These data have spawned a variety of methods to infer the structure of gene regulatory networks or to study their high-level properties, as recently reviewed\(^{145}\).

Regulatory network models generated thus far in *E. coli* and yeast have been most often validated against functional databases or prior literature\(^ {146,147}\). In contrast, only a few studies have attempted to validate or refine models systematically. However, if we are to accurately model large gene networks in complex organisms including fly, worm, mouse, and human, automated procedures will be essential for analyzing the network, choosing the best new experiments to test the model, conducting the experiments, and integrating the resulting data.

The problem of choosing the best experiments to estimate a model, termed *experimental design* or *active learning*, has been a significant area of research in statistics and machine learning\(^ {148-150}\). Automating the experimental design process can greatly accelerate data collection and model building, leading to substantial savings in time, materials, and human effort. For these reasons, many industries such as electronic circuit fabrication and airplane manufacturing incorporate experimental design as an integral step in the design process\(^ {151,152}\). A promising application of experimental design for biological systems was presented by King et al.\(^ {153}\), who integrated computational modeling and experimental design to reconstruct a small, well-studied metabolic pathway. Whether automated experimental design can be useful in a large and poorly-characterized biological system with noisy data remains
an open question.

Recently, we reported a procedure to infer gene regulatory network models by integrating gene expression profiles with high-throughput measurements of protein interactions\textsuperscript{35}. Here, we extend this procedure to incorporate automated design of new experiments. First, we use the previously-described modeling procedure to generate a library of models corresponding to different gene regulatory systems in yeast. Many of these models contain transcriptional interactions for which the regulatory effects (inducer vs. repressor) are ambiguous and cannot be determined from publicly-available expression profiles. Next, to address these ambiguities we implement a score function that ranks possible genetic perturbation experiments based on their projected information content over the models. We perform four of the highest-ranking perturbations experimentally and integrate the data back into the model. The new data support two out of three novel regulatory pathways predicted to mediate expression changes downstream of the yeast transcriptional regulator SWI4.

Results

Summary of physical regulatory models

We applied a previously-described network modeling procedure\textsuperscript{35} to integrate three complementary sources of gene regulatory information in yeast: 5558 promoter-binding interactions for 106 transcription factors measured using the ChIP-chip approach\textsuperscript{24}; the set of all 15,116 pair-wise protein-protein interactions recorded in the Database of Interacting Proteins as of April 2004\textsuperscript{154}; and a panel of mRNA expression profiles for 273 individual gene deletion experiments\textsuperscript{29}. Software for performing the network modeling procedure is available as a plug-in to the Cytoscape package\textsuperscript{11} on our Supplemental Website (http://www.cellcircuits.org/Yeang2005/).

For each gene deletion experiment, the modeling procedure identified the most
Figure 4.1. Wiring diagrams for example network models.

Two examples are shown: Model 0, showing regulatory pathways that have unique functional annotations [a]; and Model 1, showing regulatory pathways downstream of SWI4 and SOK2 with ambiguous functional annotations (several would be consistent with the observed expression responses: two possibilities are shown in [b1] and [b2]). In the models, a connection from gene a to b represents the experimental observation that [1] the proteins encoded by a and b physically interact in a protein-protein interaction (dotted links); or [2] the protein encoded by a binds the promoter of b (solid links). Each gene is either a knockout cause (red nodes), a differentially-expressed effect (tan nodes), or a signal transducer that was chosen for follow-up perturbation (gray nodes). Functional annotations (edge colors) are uniquely determined [a] or else multiple annotations are possible [b1, b2] based on the available data. Diagram layout is performed automatically using the Cytoscape package\textsuperscript{11}.
probable paths of protein-protein and promoter-binding interactions that connect the deleted gene (the perturbation) to genes that were differentially expressed in response to the deletion (the effects of perturbation). Thus, a path represented one possible physical explanation by which a deleted gene regulates a second gene downstream. Based on the expression data, each interaction on a path was annotated with (1) its probable direction of information flow and (2) its probable regulatory effect as an inducer or repressor.

For example, the model in Figure 4.1[a] (top center) includes a path from \textit{GLN3} through \textit{GCN4} to a block of downstream affected genes. This model integrates evidence that: (1) Gln3p binds the promoter of \textit{GCN4} with high significance in a ChIP-chip assay \cite{3} (p ≤ 8·10^{-4}); (2) Gcn4 binds the promoters of many genes in the ChIP-chip assay (\textit{RIB5}, \textit{YJL200C}, and others in the downstream block); and (3) a significant number of genes in the block are up-regulated in a \textit{gln3Δ} knockout but down-regulated in a \textit{gcn4Δ} knockout\textsuperscript{29}. Together, these integrated data confirm Gcn4p as an activator of downstream genes\textsuperscript{155} and lead to a (novel) annotation that Gln3p likely regulates \textit{GCN4} via transcriptional repression.

In total, the modeling process generated 4836 paths, each explaining expression changes for a particular gene in one or more knockout experiments. Of the 965 interactions covered by paths, 194 had regulatory effects that were uniquely determined by the data, while regulatory effects of the remaining 771 interactions were ambiguous. For example, Figure 4.1b includes ambiguous interaction paths through \textit{SWI4}, \textit{SOK2}, and \textit{MSN4} explaining the observation that many genes for which the promoters are bound by \textit{MSN4} are up-regulated in a \textit{swi4Δ} knockout. This observation can be explained by several alternative annotations: one scenario is that \textit{SWI4} activates \textit{SOK2} and \textit{SOK2} represses \textit{MSN4} (Figure 4.1b1), while another is that \textit{SWI4} represses \textit{SOK2} and \textit{SOK2} activates \textit{MSN4} (Figure 4.1b2). These regulatory
annotations could be uniquely determined by measuring the expression changes of genes downstream of MSN4 in the model in response to a sok2Δ deletion and an msn4Δ deletion (see below).

Paths with ambiguous interactions were partitioned into 37 independent network models (numbered 1-37), where each model contained a distinct region of the physical network (see Methods). The remaining non-ambiguous paths were grouped into a single model (referred to as Model 0). As shown in Table 4.1, 21 of all models (55%) contained pathways that had been well documented in the literature or were significantly enriched for genes belonging to specific MIPS functional categories. Of 132 protein-DNA interactions incorporated into Model 0, we found that 50 had been confirmed in classical (low-throughput) assays as reported in the Proteome BioKnowledge Library. Moreover, the inferred regulatory roles (induction or repression) for 48 out of 50 of these interactions agreed with their experimentally-determined roles (96%, binomial p-value < 1.22×10⁻⁷). Wiring diagrams for Models 0 and 1 are given in Figure 4.1; diagrams for all other regulatory network models are provided in Supplemental Materials and on the Supplemental Website (http://www.cellcircuits.org/Yeang2005).

Experiment selection

As shown in Figure 4.2, we implemented an information-theoretic approach to discriminate among ambiguous model annotations using the fewest additional gene expression experiments. All non-lethal single gene knockout experiments were ranked by their projected information content based on the inferred models (see Methods). Table 4.2 reports the list of top-ranking experiments. This list coincides roughly with biological intuition, in the sense that informative target genes typically encode proteins
Figure 4.2. Schematic of the experimental design approach.

The input to the approach is a set of alternative representations of a gene regulatory model, each of which is equally likely given current expression data. In the present work, the alternatives arise due to ambiguities in the regulatory roles of interactions in the model as inducers or repressors of downstream genes. Next, a scoring procedure is used to rank candidate perturbations according to their expected information gain over the model alternatives. High-ranking perturbations are applied to the system and characterized using gene expression microarrays. The resulting expression profiles validate or invalidate particular connections in the model and reduce the set of model alternatives to those that are consistent with both old and new expression measurements.
that are network “hubs”, each having a large number of regulatory interactions with
downstream genes in the models. However, as discussed later, knocking out “hubs”
only is not as effective as using the information-theoretic criteria.

Among the highest priority experiments, Model 1 (Figure 4.1b) was the most
often targeted, containing 3 of the top 10 highest scoring genetic perturbations: sok2Δ,
yap6Δ, and msn4Δ. A fourth perturbation to Model 1, hap4Δ, was also highly ranked
(rank 34). Therefore, Model 1 was chosen for further experimentation.

Model validation

Knockout strains corresponding to the high-ranking perturbations sok2Δ,
yap6Δ, hap4Δ, and msn4Δ were grown in quadruplicate under conditions identical to
those for the initial 273 knockouts by Hughes et al29. Gene expression profiles were
obtained for each knockout culture versus wild-type using yeast-genome microarrays.
We sought to test the three regulatory cascades leading from SWI4 to SOK2 to either
MSN4, HAP4, or YAP6 (Figure 4.1b). To verify these cascades independently of the
model, we analyzed the expression patterns of gene sets known to be directly regulated
by MSN4, HAP4, or YAP6 (obtained from the Proteome BioKnowledge Library156; see
Supplemental Table 4.4).

To normalize between our microarray procedures and those of Hughes et al.,
we also repeated the original swi4Δ expression profile, and filtered the above sets to
select only those genes with expression changes that were reproducible (i.e., same
direction of change) between the Hughes et al. swi4Δ profile and our new profile.
Expression changes were reproducible for 28 of 42 Msn4-regulated genes, 11 of 29
Hap4-regulated genes, and 64 of 119 Yap6-regulated genes. Expression similarity
among the genes in each filtered set was captured formally in a measure called
“coherence”; details about the computation of expression coherence and the selection of the gene sets are described further in Methods and the Supplemental Website (http://www.cellcircuits.org/Yeang2005).

As shown in Figure 4.3a, the gene set downstream of MSN4 showed coherent

![Figure 4.3](image-url)

**Figure 4.3. Validation and refinement of Swi4 transcriptional cascades.**

Yeast genome microarrays were used to explore three transcriptional cascades from Model 1 involving the transcriptional regulators SWI4, SOK2, and either [a] MSN4, [b] HAP4, or [c] YAP6. Bar charts show the expression coherence of genes regulated by MSN4, HAP4, or YAP6 in knockout strains swi4Δ, sok2Δ, msn4Δ, hap4Δ, and yap6Δ. Coherence scores more extreme than ±0.7 are significant (p < 0.01; dotted lines). Results are also shown for genes bound by MSN1 [d] as representative of an unrelated model not targeted by these perturbations. This analysis provides validation for the MSN4 and HAP4 pathways and disambiguates the role of each pathway interaction as activating (SWI4 interactions) or repressing (SOK2 interactions) downstream genes [e]. The YAP6 pathway hypothesis is not supported by this analysis.
up-regulation in the \(swi4\Delta\) (\(p \leq 10^{-4}\)) and \(sok2\Delta\) (\(p \leq 10^{-4}\)) knockouts, but down-regulation in the \(msn4\Delta\) (\(p \leq 8\times10^{-4}\)) knockout. This result supports the existence of a regulatory cascade leading from \(SWI4\) to \(SOK2\) to \(MSN4\). Furthermore, in the context of the present regulatory cascade, \(MSN4\) appears to be an inducer since its downstream gene set was down-regulated in the \(msn4\Delta\) experiment. In contrast, \(SOK2\) appears to be a repressor of \(MSN4\) since a \(sok2\Delta\) deletion experiment up-regulates the same set of genes. Finally, \(SWI4\) appears to be an inducer of \(SOK2\) since the \(swi4\Delta\) knockout has the same effect as \(sok2\Delta\) (i.e., up-regulation).

Results were qualitatively similar for the \(HAP4\) pathway (Figure 4.3b). The gene set downstream of \(HAP4\) was up-regulated in the \(swi4\Delta\) (\(p \leq 10^{-2}\)) and \(sok2\Delta\) (\(p \leq 9\times10^{-4}\)) knockouts but down-regulated in \(hap4\Delta\) (\(p \leq 10^{-4}\)). These results suggest that \(swi4\Delta\), \(sok2\Delta\), and \(hap4\Delta\) deletions affect the set of genes immediately downstream of \(HAP4\), supporting the \(SWI4\)-\(SOK2\)-\(HAP4\) regulatory pathway hypothesis.

In contrast to the \(MSN4\) and \(HAP4\) pathways, the gene set downstream of \(YAP6\) had insignificant responses to all follow-up knockout experiments (Figure 4.3c). Thus, the existence of the \(SWI4\)-\(SOK2\)-\(YAP6\) regulatory pathway was not supported by our validation experiments.

**Automated model refinement**

We used our modeling procedure to construct a new physical network model using the original 273 knockout gene expression experiments of Hughes et al. combined with the new \(sok2\Delta\), \(hap4\Delta\), \(msn4\Delta\), and \(yap6\Delta\) profiles. Overall, 60 protein-DNA interactions were disambiguated by our data: 50 interactions were resolved as definite inducers or repressors, whereas ten interactions were removed
from the model because the expression of downstream genes did not change as a result of the knockout. In the updated Model 1, *MSN4* and *HAP4* were unambiguously annotated as inducers of downstream genes, *SOK2* was annotated as a repressor of *MSN4* and *HAP4*, and *SWI4* was annotated as an inducer of *SOK2* (Figure 4.3e). These results agree with our previous manually-derived annotations (see Model Validation above).

**Learning curve analysis**

We quantified the efficiency of our information-based approach by comparing it to two other methods of prioritizing gene knockout experiments: prioritizing hubs and prioritizing genes randomly. First, we generated a “reference” model by fixing each ambiguous interaction in Models 1-37 to be an inducer or repressor. Assignments were chosen arbitrarily from the set of annotations that were consistent with the original knockout data. Next, we used each method (information, hub, or random) to iteratively “learn” these assignments. In each iteration, we selected the highest-priority knockout experiment, simulated the resulting expression changes (up/down) using the “reference” model, updated the inferred model, and recorded the number of ambiguous interactions that were resolved. This iterative learning procedure was repeated 100 times.

As shown in Figure 4.4, the mutual information criterion significantly outperformed hub-based and random selection. The learning curves also provide an estimate of the number of additional experiments needed to reduce model ambiguity below a given level. For example, using the information-based score, ten knockout experiments are needed to reduce the number of ambiguous interactions by 50%. In contrast, over 25 experiments are needed according to the hub-based method. Figure
4.4 suggests that performing 40 additional experiments selected using the information-based score will clarify the regulatory roles of about 70% of the ambiguous interactions. The learning rate of the final 30% becomes very slow because these interactions are isolated in the physical network, unconnected to others, and thus require separate knockouts to decipher each of them.

![Figure 4.4. Simulated learning curves of three experimental design methods.](image)

Three different methods of selecting experiments are compared: mutual information scores (triangles), hub selection (circles), and random selection (squares). We performed 100 simulated trials and show the average number of ambiguous interactions remaining in the inferred model after each simulated knockout experiment. Vertical bars indicate the standard deviations for the random selection method. The standard deviations for the information and hub selection curves are less than five and are not shown for clarity.

**Discussion**

We have used global expression profiles to validate models of transcriptional regulation inferred from protein-protein interactions, genome-wide location analysis, and expression data. A previously described network inference algorithm\textsuperscript{35} identifies probable paths of physical interactions connecting a gene knockout to genes that are differentially-expressed as a result of that knockout. The proposed validation strategy uses information gain as a criterion for choosing optimal knockouts to profile using microarray experiments. This strategy agrees with intuition, in that optimal knockouts typically target intermediate genes along the pathways under consideration.
If an intermediate gene knockout fails to affect downstream genes in a pathway, that pathway is removed from the model.

The validated pathways point to a combination of previously documented and novel findings. First, in agreement with previous literature, we confirm that \textit{MSN4} and \textit{HAP4} are inducers\textsuperscript{157,158} and that \textit{SOK2} is a repressor\textsuperscript{159}. For instance, \textit{SOK2} is known to act downstream of PKA to repress genes involved in stress response, glycogen storage, and pseudohyphal growth\textsuperscript{159}. However, although \textit{SOK2} is thought to control these pathways via a transcriptional cascade, the components of this cascade have remained unclear. Here, we provide evidence for a model in which \textit{SOK2} acts as a negative regulator upstream of \textit{MSN4} and \textit{HAP4}. Interestingly, \textit{MSN4} has been shown to activate stress response genes\textsuperscript{158}, and \textit{HAP4} has been shown to activate genes involved in energy conservation and oxidative carbohydrate metabolism\textsuperscript{157}. Thus, we have identified a candidate model for the transcriptional cascade downstream of PKA signaling that mediates stress response. This model includes two novel regulatory pathways from \textit{SWI4} to \textit{SOK2} to \textit{MSN4} and from \textit{SWI4} to \textit{SOK2} to \textit{HAP4}. The validation experiments do not support the third predicted pathway from \textit{SWI4} to \textit{SOK2} to \textit{YAP6}.

In model simulations, choosing new gene knock out experiments with an information-theoretic approach significantly outperformed both random and hub-based selection. It also outperformed the observed experimental results: approximately 280 interactions were disambiguated after four simulated knockouts (Figure 4.4), whereas only 60 interactions were resolved due to the four actual knockouts \textit{sok2Δ, hap4Δ, msn4Δ, and yap6Δ}. This difference in performance stems from key differences between the simulated and actual scenarios. In simulation, the four experiments are performed independently and iteratively, selecting the absolute highest ranking knockout each time. In the actual study, four high-ranking experiments (but not the
highest) are chosen to interrogate and maximally resolve a single pathway model, resulting in experiments that are highly co-dependent and performed simultaneously without intervening rounds of inference and experimental design. In addition, the simulation assumes that all interactions in the model are correct, along with one of the initial sets of inducer/repressor annotations. It therefore isolates the process of learning regulatory role annotations, whereas the actual procedure also serves to distinguish interactions as true- versus false-positives. Nevertheless, the simulation provides a useful comparison of experimental design methods relative to each other.

An important limitation of the single-gene knockout approach is that single perturbations do not identify pathway intermediates for which loss of function can be compensated by another gene. Furthermore, our approach may not identify regulatory pathways in which several transcription factors independently activate gene expression. Applying knockouts in combination may prove fruitful in these cases. For instance, approximately 4000 double knockouts have been reported in yeast that lead to synthetic lethality, i.e. a lethal phenotype that is not observed in either of the single knockouts individually\textsuperscript{160}. These interactions suggest regulatory relationships which could be incorporated in future work.

**Conclusions**

Scientific discovery is an iterative process of building models to explain experimental observations and validating models with new experiments\textsuperscript{161}. Experimental design is the essential link between these two aspects. Here, we have explored a framework for modeling transcriptional networks, in which experimental design and validation are central features. This framework is based on computational analysis and expression microarrays, both of which are amenable to automation, suggesting a high-throughput strategy for mapping gene regulatory pathways.
Methods

Model building and inference

Physical mechanisms of transcriptional regulation were modeled using an approach described previously. Briefly, we postulated that the regulatory effects of deleting a gene are propagated along paths of physical interactions (protein-protein and protein-DNA). We formalized the properties of these paths and interactions using a factor graph and found the most probable set of paths using the max-product algorithm. The resulting set of paths was partitioned into independent network models, also as described previously. The raw data used in the modeling procedure included: 5558 promoter-binding interactions (at p-value < 0.001) for 106 transcription factors; the set of all 15,166 pair-wise protein-protein interactions recorded in the Database of Interacting Proteins as of April 2004; and mRNA expression profiles for 273 individual gene deletion experiments. Expression changes with a p-value < 0.02 were considered significant.

Experiment scoring

We calculated the expected information gain for each of the 4756 possible non-lethal single-gene deletion experiments that were not included in the set of 273 deletions used to generate our network models. Intuitively, information gain measures (the logarithm of) the number of ambiguous annotations in the model that are likely to be determined after generating a yeast-genome expression profile in response to a particular gene deletion under consideration. Each gene deletion experiment predicts a distinct expression profile given a particular configuration of model annotations. Experiments with high information gain are those for which the predicted expression profiles are highly variable over the set of possible annotations. In these cases, only one (or at most a few) of the predicted profiles will match the true observed profile,
efficiently constraining the space of possible model annotations.

The information gain discussed above arises from the expected value of information calculations in statistical decision theory. Here we describe the score more directly in terms of reduction of model entropy. The entropy of a set of ambiguous model annotations is given by:

\[ H(M) = -\sum_m P(M = m) \log_2 P(M = m) \]

The expected information gain is the difference between the entropies before and after a hypothetical experiment:

\[ I(M; Y^e) = H(M) - H(M | Y^e) \]

\[ = H(M) + \sum P(M = m, Y^e = y) \log_2 P(M = m | Y^e = y) \]

where \( Y^e \) denotes the vector of predicted expression changes for each gene in the model under experiment \( e \). The conditional entropy \( H(M | Y^e) \) requires us to consider all possible models and corresponding outcomes resulting from experiment \( e \). Direct enumeration of all values of \( M \) and \( Y^e \) is impractical; instead, we make several simplifying approximations as described on the Supplemental Website (http://www.cellcircuits.org/Yeang2005).

**Expression profiling**

Expression profiling experiments were based on the wild-type diploid BY4743 and homozygous gene knock-out strains derived from this parent (Invitrogen), with cultures grown identically to Hughes et al. Labeled cDNA from each gene knockout strain was co-hybridized versus wild type cDNA in quadruplicate two-color microarray hybridizations. Total RNA was isolated by hot acid phenol extraction, purified to mRNA (Ambion PolyAPure kits), and labeled with Cy3 or Cy5 by
direct incorporation (Amersham CyScribe First-Strand cDNA Labeling Kit). DNA microarrays were spotted from the Yeast Genome Oligo Set v1.1 (Qiagen) on Corning UltraGAPS slides using an OmniGrid® 100 robot (Genomic Solutions). Lyophilized Cy3- and Cy5-labeled samples were resuspended in 50μL buffer (5X SSC, 0.1% SDS, 1X Denhardt’s solution, 25% formamide) and co-hybridized at 42°C beneath a coverslip for 15 hours. Arrays were imaged at 10μm resolution using a ScanArray Lite instrument (PerkinElmer). Raw quantitated background intensities were smoothed using a 7x7 median filter, separately for the Cy3 and Cy5 channels, and data were corrected for cyanine-dye dependent bias using a Qspline normalization. The VERA/SAM package was utilized to assign a log-likelihood statistic λ with each gene, indicating its significance of differential expression in each experiment. Microarray expression data are deposited in the ArrayExpress database under accession numbers A-MEXP-217 (Arrays) and E-MEXP-351 (Experiments).

Expression coherence

The expression coherence of a set of genes measures whether the expression levels of these genes behave similarly in a particular experiment. Each gene $i$ in gene-deletion experiment $e$ has an expression ratio $r_{ie}$ (versus wild-type) and associated p-value $p_{ie}$ of differential expression. First, we filter out insignificant expression changes with a p-value > 0.5. Then, we use the inverse Gaussian cumulative distribution function, $\Phi^{-1}$, to convert each remaining p-value into a z-score:

$$z_{ie} = \Phi^{-1}(1-p_{ie})$$

Next, we compute a “signed z-score” by multiplying $z$ by +1 if the expression level is increasing and by −1 if it is decreasing. The average signed z-score for a gene subset of size $N$ is computed as:
Gene sets with expression changes that are significant and in the same direction result in large $Z$-values. A distribution of $Z$ values obtained from random gene sets of size $N$ was used to determine a p-value for each expression coherence score.

$$Z_e = \left| \frac{1}{N} \sum_{i=1}^{N} \partial(z_{i \epsilon} > 0) \text{sgn}(r_{ie}) z_{ie} \right|$$
Tables

Table 4.1. Internal validation for 21 of the 38 inferred models.
The number of genes and variants are shown for each model along with the results of our preliminary validations. Each variant corresponds to a distinct set of functional annotations on the interactions in the model (directions and regulatory effects: see text). For Model 0, the expression data implied a unique set of annotations; for all other models multiple sets of annotations were possible. Each model was validated if its pathways were (wholly or partially) cited in previous studies or its downstream genes were significantly enriched for MIPS functional categories (p<0.05; hypergeometric test with Bonferroni correction).

<table>
<thead>
<tr>
<th>Model</th>
<th>#Genes</th>
<th>#Variants</th>
<th>Validated literature pathway</th>
<th>Enriched MIPS functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>130</td>
<td>1</td>
<td>Kss1/Fus3-Ste12 (mating response and filamentous growth)</td>
<td>cell fate (1.48x10^-7); metabolism (0.0067)</td>
</tr>
<tr>
<td>1</td>
<td>69</td>
<td>8</td>
<td>Sok2-Msn4 (PKA pathway)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>63</td>
<td>16</td>
<td>Tup1-Hlh1 (histone regulation)</td>
<td>protein synthesis (7.13x10^-8)</td>
</tr>
<tr>
<td>3</td>
<td>44</td>
<td>2</td>
<td>Tup1/Ssn6-Nrg1 (glucose metabolism)</td>
<td>transport (1.05x10^-5); metabolism (5.41x10^-4)</td>
</tr>
<tr>
<td>4</td>
<td>58</td>
<td>8</td>
<td>Tup1/Ssn6-α2/Mcm1 (mating response)</td>
<td>cell fate (1.12x10^-5)</td>
</tr>
<tr>
<td>5</td>
<td>58</td>
<td>4</td>
<td>Rpd3-Abf1 (histone modification)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>44</td>
<td>2</td>
<td>Swi4-Ndd1-Ace2 (cell cycle)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>26</td>
<td>4</td>
<td>Slt2-Rlm1/Swi4 (PKC pathway)</td>
<td>cell cycle (0.035)</td>
</tr>
<tr>
<td>8</td>
<td>36</td>
<td>8</td>
<td>Med2-Gal4/Gcn4 (general transcription)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>45</td>
<td>16</td>
<td>Med2-Gal11-Gal4 (general transcription)</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>9</td>
<td>4</td>
<td>Cmd1-Cna1-Skn7 (calcium signaling)</td>
<td>cell defense (6.33x10^-6)</td>
</tr>
<tr>
<td>23</td>
<td>17</td>
<td>2</td>
<td>Med2-Srb6-Gal4 (general transcription)</td>
<td>metabolism (1.49x10^-6); energy (0.04)</td>
</tr>
<tr>
<td>26</td>
<td>8</td>
<td>8</td>
<td>Med2-Srb5-Gal4 (general transcription)</td>
<td>cell defense (9.62x10^-5)</td>
</tr>
<tr>
<td>29</td>
<td>9</td>
<td>4</td>
<td>Med2-Srb6-Gal4 (general transcription)</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>12</td>
<td>4</td>
<td>Med2-Srb6-Gal4 (general transcription)</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>12</td>
<td>4</td>
<td>Med2-Srb6-Gal4 (general transcription)</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>12</td>
<td>4</td>
<td>Med2-Srb6-Gal4 (general transcription)</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>9</td>
<td>4</td>
<td>Ste12-Mcm1 (mating response)</td>
<td>cell fate (4.55x10^-8); homeostasis (0.0012); cell communication (0.0345)</td>
</tr>
<tr>
<td>36</td>
<td>7</td>
<td>4</td>
<td>Med2-Srb6-Gal4 (general transcription)</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>5</td>
<td>2</td>
<td>Med2-Srb6-Gal4 (general transcription)</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.2. Top-ranking knock-out experiments proposed for model discrimination.
Each proposed target gene is reported, along with its function, mutual information score, rank, and the model(s) it informs. All target genes are non-lethal in rich media. A ‘*’ indicates gene knockouts selected in this study.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
<th>Score</th>
<th>Downstream genes</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHF1</td>
<td>Histone</td>
<td>52.1429</td>
<td>74</td>
<td>1 2</td>
</tr>
<tr>
<td>SOK2*</td>
<td>regulator for meiosis and PKA pathway</td>
<td>45.0279</td>
<td>64</td>
<td>2 1</td>
</tr>
<tr>
<td>CKA1</td>
<td>protein kinase of cell cycle</td>
<td>45.0075</td>
<td>64</td>
<td>3 5</td>
</tr>
<tr>
<td>A2</td>
<td>mating response</td>
<td>40.9023</td>
<td>58</td>
<td>4 4</td>
</tr>
<tr>
<td>YAP6*</td>
<td>stress response regulator</td>
<td>35.1625</td>
<td>50</td>
<td>5 1 3</td>
</tr>
<tr>
<td>NRG1</td>
<td>regulator of glucose dependent genes</td>
<td>31.6501</td>
<td>45</td>
<td>6 3</td>
</tr>
<tr>
<td>FKH1</td>
<td>regulator of cell cycle</td>
<td>29.1194</td>
<td>41</td>
<td>7 2</td>
</tr>
<tr>
<td>FKH2</td>
<td>regulator of cell cycle</td>
<td>26.7131</td>
<td>38</td>
<td>8 7</td>
</tr>
<tr>
<td>SLT2</td>
<td>protein kinase of cell wall integrity pathway</td>
<td>23.4727</td>
<td>31</td>
<td>9 8</td>
</tr>
<tr>
<td>MSN4*</td>
<td>regulator of stress response</td>
<td>21.8224</td>
<td>31</td>
<td>10 1</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>HAP4*</td>
<td>regulator of cellular respiration</td>
<td>6.3310</td>
<td>9</td>
<td>34 1</td>
</tr>
</tbody>
</table>
Supplemental Tables

Supplemental Table 4.1. Internal validation for 17 out of 24 restricted network models.

<table>
<thead>
<tr>
<th>Model</th>
<th># Genes</th>
<th>Validated literature pathway</th>
<th>Refs.</th>
<th>Enriched MIPS Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>109</td>
<td>Kss1/Fus3-Ste12 (mating response and filamentous growth)</td>
<td>163</td>
<td>cell fate (1.48x10^-3); metabolism (0.0067)</td>
</tr>
<tr>
<td>1</td>
<td>69</td>
<td>Sok2-Msn4 (PKA pathway)</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>Tup1/Ssn6-Nrg1 (glucose metabolism)</td>
<td>164</td>
<td>transport (1.05x10^-4); metabolism (5.41x10^-4)</td>
</tr>
<tr>
<td>3</td>
<td>58</td>
<td>Tup1/Ssn6-α2/Mcm1 (mating response)</td>
<td>165</td>
<td>cell fate (1.12x10^-3)</td>
</tr>
<tr>
<td>4</td>
<td>44</td>
<td>Swi4-Ndd1-Ace2 (cell cycle)</td>
<td>196</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>36</td>
<td>Slt2-Rlm1/Swi4 (PKC pathway)</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>Cmd1-Cna1-Skn7 (calcium signaling)</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td></td>
<td></td>
<td>cell defense (6.33x10^-6)</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>Med2-Gal11-Gal4 (general transcription)</td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td></td>
<td></td>
<td>metabolism (0.026)</td>
</tr>
<tr>
<td>13</td>
<td>17</td>
<td></td>
<td></td>
<td>metabolism (1.49x10^-3)</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>Gcn4-Met4 (methionine synthesis)</td>
<td></td>
<td>metabolism (2.88x10^-4)</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>Ste12-Mcm1 (mating response)</td>
<td></td>
<td>cell fate (4.55x10^-4); homeostasis (0.0012); cell communication (0.0345)</td>
</tr>
<tr>
<td>17</td>
<td>13</td>
<td>Cka2-Abf1 (casein kinase pathway)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td></td>
<td></td>
<td>metabolism (0.0017)</td>
</tr>
<tr>
<td>22</td>
<td>8</td>
<td>Cka2-Abf1 (casein kinase pathway)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>Cka2-Abf1 (casein kinase pathway)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplemental Table 4.2. Correlations between *swi4Δ* and *gcn4Δ* data from Rosetta and the new experiments.

<table>
<thead>
<tr>
<th></th>
<th><em>swi4Δ</em> total</th>
<th><em>gcn4Δ</em> total</th>
</tr>
</thead>
<tbody>
<tr>
<td>correlation</td>
<td>-0.013</td>
<td>0.34</td>
</tr>
<tr>
<td>correlation p-value</td>
<td>0.84</td>
<td>&lt; 10^{-4}</td>
</tr>
<tr>
<td>rank correlation</td>
<td>-0.058</td>
<td>0.145</td>
</tr>
<tr>
<td>rank correlation p-value</td>
<td>1.0</td>
<td>&lt; 10^{-4}</td>
</tr>
<tr>
<td>hyper-geometric p-value</td>
<td>0.188</td>
<td>4.88×10^{-21}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th><em>swi4Δ</em> subset</th>
<th><em>gcn4Δ</em> subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>correlation</td>
<td>0.344</td>
<td>0.831</td>
</tr>
<tr>
<td>correlation p-value</td>
<td>8×10^{-5}</td>
<td>&lt; 10^{-5}</td>
</tr>
<tr>
<td>rank correlation</td>
<td>0.353</td>
<td>0.705</td>
</tr>
<tr>
<td>rank correlation p-value</td>
<td>9×10^{-5}</td>
<td>&lt; 10^{-5}</td>
</tr>
<tr>
<td>hyper-geometric p-value</td>
<td>1.05×10^{-5}</td>
<td>&lt; 1.8×10^{-13}</td>
</tr>
</tbody>
</table>
### Supplemental Table 4.3. Restricted subsets used to evaluate the reproducibility.

#### Swi4 subset

<table>
<thead>
<tr>
<th>Subset</th>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
<th>Gene 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAP1</td>
<td>LYS20</td>
<td>YAP6</td>
<td>YJL225C</td>
<td></td>
</tr>
<tr>
<td>ATR1</td>
<td>MNN1</td>
<td>YBL029W</td>
<td>YJR078W</td>
<td></td>
</tr>
<tr>
<td>BAT2</td>
<td>MSN4</td>
<td>YBL112C</td>
<td>YKL051W</td>
<td></td>
</tr>
<tr>
<td>CCP1</td>
<td>NCE102</td>
<td>YBL113C</td>
<td>YLR084C</td>
<td></td>
</tr>
<tr>
<td>CLB1</td>
<td>PCL7</td>
<td>YDR451C</td>
<td>YLR463C</td>
<td></td>
</tr>
<tr>
<td>CLB2</td>
<td>PDR12</td>
<td>YDR545W</td>
<td>YLR465C</td>
<td></td>
</tr>
<tr>
<td>CLB6</td>
<td>PRY2</td>
<td>YEL045C</td>
<td>YLR466W</td>
<td></td>
</tr>
<tr>
<td>COX9</td>
<td>QCR2</td>
<td>YEL047C</td>
<td>YLR467W</td>
<td></td>
</tr>
<tr>
<td>CPA2</td>
<td>RNR1</td>
<td>YEL077C</td>
<td>YML133C</td>
<td></td>
</tr>
<tr>
<td>CSI2</td>
<td>RPI1</td>
<td>YER045C</td>
<td>YNL339C</td>
<td></td>
</tr>
<tr>
<td>CUP9</td>
<td>RPL19B</td>
<td>YER189W</td>
<td>YNR067C</td>
<td></td>
</tr>
<tr>
<td>ECM13</td>
<td>SAT2</td>
<td>YER190W</td>
<td>YOL011W</td>
<td></td>
</tr>
<tr>
<td>ECM33</td>
<td>SCW10</td>
<td>YFR006W</td>
<td>YOR248W</td>
<td></td>
</tr>
<tr>
<td>ERG6</td>
<td>SGA1</td>
<td>YGR086C</td>
<td>YOR315W</td>
<td></td>
</tr>
<tr>
<td>EXG1</td>
<td>SNQ2</td>
<td>YGR153W</td>
<td>YOX1</td>
<td></td>
</tr>
<tr>
<td>HAP1</td>
<td>SOK2</td>
<td>YGR296W</td>
<td>YPL267W</td>
<td></td>
</tr>
<tr>
<td>HAP4</td>
<td>SRL1</td>
<td>YHL049C</td>
<td>YPL283C</td>
<td></td>
</tr>
<tr>
<td>HSC82</td>
<td>SVS1</td>
<td>YHR048W</td>
<td>YPR203W</td>
<td></td>
</tr>
<tr>
<td>HTA1</td>
<td>SWI4</td>
<td>YIL056W</td>
<td>YPS4</td>
<td></td>
</tr>
<tr>
<td>HTB1</td>
<td>TRP4</td>
<td>YIL177C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISU2</td>
<td>UTR2</td>
<td>YJL217W</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Gcn4 subset

<table>
<thead>
<tr>
<th>Subset</th>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADH5</td>
<td>HAD1</td>
<td>UGA3</td>
<td></td>
</tr>
<tr>
<td>APG1</td>
<td>HAL2</td>
<td>YER069W</td>
<td></td>
</tr>
<tr>
<td>ARG1</td>
<td>HIS1</td>
<td>YER073W</td>
<td></td>
</tr>
<tr>
<td>ARG11</td>
<td>HIS4</td>
<td>YGL184C</td>
<td></td>
</tr>
<tr>
<td>ARG3</td>
<td>HOM3</td>
<td>YHM1</td>
<td></td>
</tr>
<tr>
<td>ARG7</td>
<td>LEU4</td>
<td>YHR122W</td>
<td></td>
</tr>
<tr>
<td>ARG8</td>
<td>MET16</td>
<td>YHR162W</td>
<td></td>
</tr>
<tr>
<td>ARO1</td>
<td>PCL5</td>
<td>YJL200C</td>
<td></td>
</tr>
<tr>
<td>ARO3</td>
<td>RIB5</td>
<td>YLR152C</td>
<td></td>
</tr>
<tr>
<td>CPA2</td>
<td>TRP2</td>
<td>YMC1</td>
<td></td>
</tr>
<tr>
<td>GCN4</td>
<td>TRP3</td>
<td>YOL119C</td>
<td></td>
</tr>
</tbody>
</table>
Supplemental Table 4.4. Gene sets for external validation.

**Msn4 subset**

<table>
<thead>
<tr>
<th>ALD3</th>
<th>ATP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARA1</td>
<td>ATP7</td>
</tr>
<tr>
<td>CTT1</td>
<td>ATP16</td>
</tr>
<tr>
<td>YMR173W</td>
<td>COX4</td>
</tr>
<tr>
<td>DOT5</td>
<td>COX7</td>
</tr>
<tr>
<td>GLK1</td>
<td>COX12</td>
</tr>
<tr>
<td>GLO1</td>
<td>COX13</td>
</tr>
<tr>
<td>GRE3</td>
<td>QCR7</td>
</tr>
<tr>
<td>GSY1</td>
<td>TUF1</td>
</tr>
<tr>
<td>HOR2</td>
<td>QCR10</td>
</tr>
<tr>
<td>HSP12</td>
<td>YLR294C</td>
</tr>
<tr>
<td>HSP26</td>
<td></td>
</tr>
<tr>
<td>HSP104</td>
<td></td>
</tr>
<tr>
<td>HXK1</td>
<td></td>
</tr>
</tbody>
</table>

**LAP4**

<table>
<thead>
<tr>
<th>PGM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAS2</td>
</tr>
<tr>
<td>SOD2</td>
</tr>
<tr>
<td>SOL4</td>
</tr>
<tr>
<td>SPS100</td>
</tr>
<tr>
<td>SSA4</td>
</tr>
<tr>
<td>TKL2</td>
</tr>
<tr>
<td>TPS1</td>
</tr>
<tr>
<td>TTR1</td>
</tr>
<tr>
<td>YDL124W</td>
</tr>
<tr>
<td>YKR011C</td>
</tr>
<tr>
<td>YMR315W</td>
</tr>
<tr>
<td>YNL200C</td>
</tr>
</tbody>
</table>

**Hap4 subset**

<table>
<thead>
<tr>
<th>ATP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP7</td>
</tr>
<tr>
<td>ATP16</td>
</tr>
<tr>
<td>COX4</td>
</tr>
<tr>
<td>COX7</td>
</tr>
<tr>
<td>COX12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COX13</th>
</tr>
</thead>
<tbody>
<tr>
<td>QCR7</td>
</tr>
<tr>
<td>TUF1</td>
</tr>
<tr>
<td>QCR10</td>
</tr>
<tr>
<td>YLR294C</td>
</tr>
</tbody>
</table>
Acknowledgements

We gratefully acknowledge Owen Ozier, Ryan Kelley, and Rowan Christmas for their valuable assistance with model visualization, and Julia Zeitlinger for commenting on our manuscript. CM, CW and SM were supported by NIGMS grant GM070743-01 and NSF grant CCF-0425926. TI was supported by a David and Lucille Packard Fellowship award. CY and TJ were supported in part by NIH grant(s) GM68762 and GM69676.

Chapter 4, in full, is a reprint of material as it appeared on July 1 in Genome Biology, 2005, 6:R62. Yeang, Chen-Hsiang; Mak, H. Craig; McCuine, Scott; Workman, Christopher; Jaakkola, Tommi; Ideker, Trey. The dissertation author was one of two primary investigators and authors (with C.H. Yeang) of this paper.
Chapter 5: Cataloging network models in a searchable online database

Abstract

CellCircuits (http://www.cellcircuits.org) is an open-access database of molecular network models, designed to bridge the gap between databases of individual pair-wise molecular interactions and databases of validated pathways. CellCircuits captures the output from an increasing number of approaches that screen molecular interaction networks to identify functional subnetworks, based on their correspondence with expression or phenotypic data, their internal structure, or their conservation across species. This initial release catalogs 2019 computationally-derived models drawn from eleven journal articles and spanning five organisms (yeast, worm, fly, Plasmodium falciparum, and human). Models are available either as images or in machine-readable formats and can be queried by the names of proteins they contain or by their enriched biological functions. We envision CellCircuits as a clearinghouse in which theorists may distribute or revise models in need of validation and experimentalists may search for models or specific hypotheses relevant to their interests. We demonstrate how such a repository of network models is a novel systems biology resource by performing several meta-analyses not currently possible with existing databases.
Introduction

Presently, a great deal of biological information is represented as interactions between molecules. This information includes physical interactions that occur among proteins, DNA, RNA, and small molecules (e.g.\(^{38,70,170}\)); genetic interactions such as synthetic lethality, enhancement, or suppression\(^{171}\); and interactions due to co-expression\(^{172}\) or co-citation\(^{173}\). Modern analyses of interaction data typically accomplish two goals. The first goal is to clean the data, by filtering erroneous interactions that can be associated with high-throughput screens (false positives, e.g.\(^{34,174}\)) or by predicting new interactions that may have been previously missed (false negatives, e.g.\(^{175,176}\)). The second goal is to organize the interactions into biological network models — that is, collections of interactions hypothesized to work together towards a particular cellular function or within a common pathway\(^{177-179}\).

Interaction analysis is currently supported by two types of available databases (Figure 5.1). First, the raw material for analysis is provided by databases of molecular interactions including the Database of Interacting Proteins\(^{16}\), the Munich Center for Information on Protein Sequences\(^{14}\), the Biomolecular Interaction Network Database\(^{12}\). Interaction database icons represent (clockwise from top left) the Database of Interacting Proteins (DIP\(^{16}\)); the General Repository of Interaction Datasets (GRID\(^{17}\)); Molecular INTeractions Database (MINT\(^{203}\)); the IntAct molecular interactions database\(^{15}\); the interaction database at the Munich Information Center for Protein Sequences (MIPS\(^{14}\)); and Biomolecular Interaction Network Database (BIND\(^{12}\)). Pathway database icons represent Reactome\(^{50}\); Signal Transduction Knowledge Environment (STKE\(^{54}\)); Gene Networks database (GeNet\(^{53}\)); Biocarta (http://www.biocarta.com/genes); Kyoto Encyclopedia of Genes and Genomes (KEGG\(^{204}\)); and CellMap (http://cellmap.org).

Figure 5.1. The need for a new type of database.

The CellCircuits database is positioned between raw molecular interaction databases (left) and databases of rigorously validated cellular pathways (right). Interaction database icons represent (clockwise from top left) the Database of Interacting Proteins (DIP\(^{16}\)); the General Repository of Interaction Datasets (GRID\(^{17}\)); Molecular INTeractions Database (MINT\(^{203}\)); the IntAct molecular interactions database\(^{15}\); the interaction database at the Munich Information Center for Protein Sequences (MIPS\(^{14}\)); and Biomolecular Interaction Network Database (BIND\(^{12}\)). Pathway database icons represent Reactome\(^{50}\); Signal Transduction Knowledge Environment (STKE\(^{54}\)); Gene Networks database (GeNet\(^{53}\)); Biocarta (http://www.biocarta.com/genes); Kyoto Encyclopedia of Genes and Genomes (KEGG\(^{204}\)); and CellMap (http://cellmap.org).
Database\textsuperscript{12}, the BioGRID\textsuperscript{15}, and IntAct\textsuperscript{13}. Many of these databases provide confidence scores with each measured and predicted interaction. Second, there are a growing number of so-called pathway databases, in which canonical diagrams of metabolic, signaling, or regulatory pathways have been hand-curated from review articles and textbooks. Metabolic pathways are the focus of Reactome\textsuperscript{50}, MetaCyc\textsuperscript{51} and the Kyoto Encyclopedia of Genes and Genomes\textsuperscript{52}, while databases such as BioCarta (http://www.biocarta.com/genes), CellMap (http://cellmap.org), the Signal Transduction Knowledge Environment\textsuperscript{54}, GeNet\textsuperscript{53}, and TransPATH\textsuperscript{180} are primarily concerned with signaling and transcription. All of these pathway databases are relevant to the second and perhaps ultimate goal of interaction analysis — models of well-defined and well-validated functional relationships among genes, proteins, and/or metabolites.

Automatic inference of accurate and detailed molecular pathways, however, is well beyond the capability of current interaction analyses and integrative modeling approaches. While current approaches attempt to place interactions into subnetworks according to their putative function\textsuperscript{177-179}, such subnetworks are hypothetical in nature and thus inappropriate for entry into any of the existing databases of canonical pathways. Rather, the subnetwork models produced by automated approaches are typically embedded in figures, tables, or supplementary information in the primary published literature. Although it is certainly possible to read about the models, there are several problems with this traditional method of dissemination. First, the size and number of models from even a single publication can be overwhelming, making models relevant to a particular gene or function difficult to locate. Second, in many cases, network modeling papers target bioinformatic, rather than biological or medical, audiences. As a result, the models remain largely inaccessible to those who have the most knowledge to interpret them and the most to gain from their successful interpretation.
Recent opinion articles have recognized a related problem for the case of protein functional predictions, calling for a clearinghouse of hypotheses generated by bioinformatics analyses and searchable by experimental biologists. In the same vein, the BioModels Database has recently been adopted as a working repository for simulations of kinetic quantitative systems based on ordinary differential equations. Subnetworks inferred from genome-scale data, however, do not fall into this category.

Motivated by these considerations, we have designed CellCircuits as an open-access general repository of models distilled from protein networks. By aggregating models derived from many separate studies into a single resource, CellCircuits bridges the gap between databases of individual pairwise interactions and fully curated, biologically-validated pathway models. The CellCircuits database enables experimentalists to readily access and cross-reference models across multiple publications. It also enables the meta-analysis of the entire set of models to reveal inter-model relationships and to answer global questions: for instance, which models overlap in terms of the genes and/or cellular processes represented? How novel is a new result given the models that are already present in the database?

Results

A spectrum of network models

To date, interactions have been organized by searching for essentially three types of subnetworks: linear paths of interactions, interaction clusters, or parallel clusters. Representative models of each type are shown in Figure 5.2. Linear (or branching) paths of interactions have been used to represent biological pathways such as metabolic processes or regulatory cascades (Figure 5.2a). Clusters in an interaction network are regions of dense interconnections and are suggestive of functional protein complexes (Figure 5.2b). Parallel clusters are two (or
more) similar network clusters in which the proteins in one cluster are, in some way, associated with the proteins in the other cluster. Parallel clusters have been used to represent protein complexes conserved across species\textsuperscript{188-190} (Figure 5.2c), in which pairs of proteins spanning the two clusters are orthologs associated by sequence-similarity relationships. They have also been used to identify the physical basis for genetic interactions\textsuperscript{191} (Figure 5.2d), in which two protein interaction clusters are linked by many genetic interactions if the clusters perform redundant or synergistic cellular functions.

![Figure 5.2. Representative network models stored in CellCircuits.](image)

(a) A collection of linear regulatory pathways downstream of mating-type locus in yeast\textsuperscript{184} (b) An interaction cluster of co-expressed proteins suggestive of a functional complex\textsuperscript{23} (c) Parallel clusters conserved between \textit{Plasmodium falciparum} and yeast\textsuperscript{190} (d) Parallel clusters that are highly connected by genetic interactions\textsuperscript{191}.

Finally, integrating the interaction network with external data, such as gene expression profiles and other molecular states, has also been a key methodology used to identify significant subnetworks. For instance, these approaches have been used to find protein interaction clusters that exhibit coherent expression changes in response
to panels of perturbations\textsuperscript{28,62} or as a function of the cell cycle\textsuperscript{23}. Other works\textsuperscript{192} have reported network “motifs”, defined as patterns of interactions that occur more often in the network than expected by chance. However, these approaches (by design) focus on general patterns rather than subnetworks of particular proteins. Therefore, they are not considered here.

**Database coverage and assembly**

This CellCircuits initial release (version 1.0) was designed as proof-of-principle of the value of a searchable database of network models. We focused on providing a clear database interface and representative, albeit incomplete, coverage of the types of network models possible. For version 1.0, the database includes models from eleven publications, spanning linear, clustered, or parallel subnetworks, with priority given to publications with models available in both graphical representations and machine-readable formats (Table 5.1). Graphical representations of network models are a particularly valuable method of disseminating interactions and/or pathways, in much the same way that DNA sequence logos\textsuperscript{193} are used to visualize position-specific score matrices of DNA binding motifs. Conversely, machine-readable formats, such as SBML\textsuperscript{48}, BioPax\textsuperscript{47}, or the Cytoscape SIF format\textsuperscript{11}, greatly facilitate database entry, model curation, and subsequent computational analysis. Four publications provided models in both graphical and machine-readable formats\textsuperscript{185,189-191}. For the remaining seven, models were manually curated from published figures\textsuperscript{23,28,62,183,184,186,188}.

Manual curation involved downloading figures containing each network model, and then transcribing the genes and interactions in the models into a machine-readable format. For most publications, one figure, or each subpanel in a figure, contained a single network model. However, in three publications\textsuperscript{23,184,188} the figures contained multiple, unconnected networks that were not divided by the authors into
separate subpanels. In these cases, each unconnected component was entered as one model in CellCircuits, and in one case, networks were further subdivided into smaller models if they contained several sparsely connected, but functionally annotated, clusters of proteins (see Supplemental Table 5.1).

These curation activities resulted in a total of 2019 protein network models in the database. Models in the database include protein interactions from five organisms: yeast (*Saccharomyces cerevisiae*; 91% of all models), fly (*Drosophila melanogaster*; 58%), nematode worm (*Caenorhabditis elegans*; 27%), a malarial parasite (*Plasmodium falciparum*; 2%), and human (2%; these percentages total >100% due to cross-species comparisons covering multiple species in a single model). The models include up to four types of interactions (protein-protein, protein-DNA, genetic, and metabolic) as well as two types of external data (gene expression and gene deletion phenotypes).

**Network model query**

Models in the CellCircuits database are queried through a web-based interface. In the simplest use case, entering a standard gene name (e.g. RAD9) into the search field will return all models containing that gene. Wild-card searches are permitted (e.g., RAD* will search for models containing any gene with a name that begins with the letters RAD. See Figure 5.3). All gene queries are also checked against a list of gene name synonyms, which are drawn from the latest release of the Gene Ontology (GO) database. In addition, searches can be limited to models from specific publications or to models containing genes from specific organisms.

Searches based on gene function are also supported. The CellCircuits database automatically scores all models for GO functional enrichment using the hypergeometric test (see Methods). Such tests had been originally applied in only
three of the 11 curated publications. The enrichment results are stored with each model in the database as meta-data, allowing users to search for models that are enriched for genes having a particular annotation. For example, some of the same models retrieved by searching for RAD9 can also be retrieved by searching for GO annotations associated with this gene. Queries may include exact GO ID numbers (e.g., GO:0006974) or partial or complete GO term names (e.g. “DNA damage” or “integrity checkpoint”; these must be enclosed in double quotes).

More than one gene, GO annotation, or wild-card may be included in a query. If a model matches multiple search terms, it will be ranked higher in the results. All search results include graphical representations of the models, links to the original publication, the organism(s) modeled, the genes or GO annotations from the search query that were found in each model, and the hypergeometric p-value of enrichment for any GO annotations (Figure 5.3).

Figure 5.3. Web interface (http://www.cellcircuits.org).
Results using RAD* and “DNA binding” as the search query (circle 1). A total of 274 subnetwork models are returned. The search output includes a graphical representation of the model (circle 7), the genes and GO terms from the model that match the query (circle 6), alternative gene names or synonyms matching the query (circle 9), the total number of matches (circle 8), enriched GO terms (circle 5), and a link to view similar models (circle 4).
Meta-analysis of models

Collecting published network models within a single database allowed us to survey the state of computational analysis of large interaction datasets. Scoring all models for GO functional enrichment (described in the previous section) is an example of such analyses. Another example, the observed sizes of models from all eleven publications, is shown in Figure 5.4a. On average, the 2019 models in the database

Figure 5.4. Meta-analysis of models.

[a] Histogram of the number of genes or proteins per model. [b] Histogram of the number of genes (y-axis) that are contained in a given number of models (x-axis). The inset is an expanded view of the genes that span over 50 models. [c] Overlap between network modeling publications. Thicker lines represent greater similarity between the sets of models published in two publications (see legend). Similarity is measured by the number of distinct models that share one or more interactions (yeast interactions only) divided by the total number of models in both publications. Interactions are shared between almost every pair of publications, but for clarity, similarity scores less than 0.05 are not shown.
contained ~18 proteins and 36 interactions with 95% of models containing between five and 30 proteins. However, this distribution was heavily influenced by two publications\textsuperscript{189,191} which together contributed over 90% of the models in the database.

To assess the overlap between models, we examined the extent to which the same proteins appeared in multiple models (Figure 5.4b). Although a protein was shared by approximately nine models on average, the majority were found in only one or two models. Thirty-five proteins appeared in over 100 models (<5% of all models in the database). Interestingly, among these were all six of the yeast ATPases in the 26S proteasome (RPT1-6), components of the yeast and worm 20S proteasome, and several yeast, worm, and fly protein kinases. The pervasiveness of these proteins in models may reflect their broad evolutionary conservation across species, a high degree of connectivity in the protein network, their popularity in the biological literature, or their functional roles in many distinct biological processes (i.e., pleiotropy).

The results of our model overlap analyses are accessible through the web interface. Each model is annotated in the CellCircuits database with a list of similar models, defined as those that contain at least three of the same genes. Clicking the “View similar models” link in the search results will display these models (Figure 5.3, circle 4). Currently, only the number of shared genes is used to assess similarity between models. However, more complex measures could be envisioned, potentially making CellCircuits, itself, a resource for comparing several similar models (perhaps corresponding to the same biological process) and showing the differences between them.

On a broader scale, we also assessed the extent to which publications covered overlapping regions of the protein interactome using a pairwise similarity score (see Methods). Results are shown in Figure 5.4c. Although our similarity score was permissive such that some overlap was expected between every pair of publications,
only five of the 55 possible pairs showed over 25% similarity. Thus, it appears that
the different modeling publications are, to some degree, capturing different regions of
the protein interaction network (excluding \textsuperscript{189,191}, see Figure 5.4c). Furthermore, in the
future, this kind of meta-analysis could be used to determine how the results from new
publications differ from existing models.

**Discussion**

In summary, CellCircuits v1.0 provides a clearinghouse in which hypothetical
pathway models derived from large-scale protein networks may be easily accessed,
queried, and exported for further study. The eleven publications included in this
initial release were chosen to cover a broad range of network model types with a bias
towards publications that provided models in both graphical and machine-readable
format. Beyond this proof-of-principle, a significant question is whether, or to what
extent, all past and future network models might be incorporated.

On one hand, the field of network biology is still young such that the number
of relevant previous publications is probably less than fifty. On the other hand, the
rapid adoption of systems and network approaches will make capturing information
from all future works a daunting prospect if the models are not readily accessible.
CellCircuits complements existing efforts that have begun to address this challenge,
such as markup languages for describing models (BioPAX\textsuperscript{47} and SBML\textsuperscript{48}) and the
BioModels Database of quantitative, kinetic models\textsuperscript{49}. Like biological sequence and
microarray databases, we envision CellCircuits as a valuable resource for storing,
accessing, and updating network models across the wider biological research
community.
Methods

Data processing

A data processing pipeline was used to extract information from the textual representation of a model and store that information in a MySQL (http://www.mysql.org) relational database. The data processing pipeline required a digital image of each model and a text file containing the genes, proteins, metabolites, other small molecules, and interconnections represented in the model. In cases when a network model was published in graphical form only, the text file was manually transcribed (see Supplemental Table 5.1).

To ensure that the CellCircuits database used a consistent set of gene identifiers, we mapped each gene name found in the text file for a model to a GO gene id using tables from the GO database. Gene names found in a model but not in the GO database were automatically inserted into the appropriate database tables and flagged as being externally added. Future curation efforts could be directed towards handling these genes missing from the GO database. After models were entered into the database, they were scored using the hypergeometric test for GO annotation enrichment.

Web interface

We used Perl CGI scripts (http://www.perl.org) in conjunction with the Apache web server (http://httpd.apache.org), mod_perl (http://perl.apache.org), and Perl DBI (http://dbi.perl.org) to serve HTML content, handle user input, and query the MySQL database. Script.aculo.us version 1.61 (http://script.aculo.us), an open source JavaScript library, was used to generate visual effects on the web pages that display search results.
Scoring models for Gene Ontology (GO) annotation

Using the latest release of the GO database, models were scored for a statistically significant number of genes in the model that were annotated with a particular GO term. We first identified the complete set of genes associated with each GO term. This set included the genes directly annotated with that term as well as those annotated with any of the term’s descendents in the GO hierarchy. Next, we used the hypergeometric distribution\textsuperscript{195,196} to test the genes in each model against the genes annotated with each of the GO terms. The resulting P-values were stored in the database.

Scoring similarity between publications

For each pair of publications we compared all models in one publication to all of the models in the other. To capture model similarity as sensitively as possible, we defined two models to be similar if they shared at least one interaction. The similarity score of a pair of publications was defined to be the number of distinct models that participated in any overlap divided by the total number of models in the pair. For example, consider publication A containing 2 models and publication B containing 6. If model 1 in A overlaps with models 1-5 in B, and model 2 in A only overlaps with model 1 in B, then the total number of distinct overlapping models is 7, and the similarity score between publications is 7/8.
## Tables

### Table 5.1. Sources of data.

<table>
<thead>
<tr>
<th>Model source</th>
<th>Organism(s)</th>
<th>Models**</th>
<th>Genes§</th>
<th>Interactions†</th>
<th>States‡</th>
<th>Network patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernard183</td>
<td>Y</td>
<td>1</td>
<td>17</td>
<td>39</td>
<td>62</td>
<td>X</td>
</tr>
<tr>
<td>Hartemink184</td>
<td>Y</td>
<td>2</td>
<td>33</td>
<td>66</td>
<td>320</td>
<td>X</td>
</tr>
<tr>
<td>Yeang185</td>
<td>Y</td>
<td>38</td>
<td>602</td>
<td>110 708</td>
<td>273</td>
<td>X</td>
</tr>
<tr>
<td>Kelley191</td>
<td>Y</td>
<td>473</td>
<td>787</td>
<td>1246 95 1843 554</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Sharan189</td>
<td>Y</td>
<td>32</td>
<td>40</td>
<td>68</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Suthram190</td>
<td>Y</td>
<td>48</td>
<td>85</td>
<td>74</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Gandhi188</td>
<td>Y</td>
<td>31</td>
<td>719</td>
<td>712</td>
<td>66</td>
<td>X</td>
</tr>
<tr>
<td>de Lichtenberg223</td>
<td>Y</td>
<td>10</td>
<td>107</td>
<td>83 26</td>
<td>20</td>
<td>X</td>
</tr>
<tr>
<td>Ideker28</td>
<td>Y</td>
<td>6</td>
<td>130</td>
<td>129 12</td>
<td>26</td>
<td>X</td>
</tr>
<tr>
<td>Begley62</td>
<td>Y</td>
<td>8</td>
<td>280</td>
<td>85 410</td>
<td>13</td>
<td>X</td>
</tr>
<tr>
<td>Haugen186</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>TOTALS:</td>
<td>2019</td>
<td>3622</td>
<td>5312</td>
<td>1356 1843 567 772</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Y = Yeast ; W = Worm ; F = Fly ; H = Human ; P = Plasmodium falciparum

** counts refer to total number of models across all organisms modeled

§ counts refer to number of distinct genes in yeast only across all models.

† counts refer to number of distinct interactions in yeast only across all models.

‡ For gene expression, counts refer to number of profiles used.
## Supplemental Tables

### Supplemental Table 5.1. Figures curated.

<table>
<thead>
<tr>
<th><strong>Model Source</strong></th>
<th><strong>Figures curated</strong></th>
<th><strong>Curation method</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Begley\textsuperscript{62}</td>
<td>5b (subpanels #1-#6)</td>
<td>Each panel is one model (1:1)</td>
</tr>
<tr>
<td>Ideker\textsuperscript{28}</td>
<td>5 (subpanels #1a-e, #2-#7)</td>
<td>1:1</td>
</tr>
<tr>
<td>Haugen\textsuperscript{186}</td>
<td>1a-f, 5a, 5b</td>
<td>1:1</td>
</tr>
<tr>
<td>Bernard\textsuperscript{183}</td>
<td>3</td>
<td>1:1</td>
</tr>
<tr>
<td>Hartemink\textsuperscript{184}</td>
<td>2</td>
<td>Each unconnected component is one model</td>
</tr>
<tr>
<td>Gandhi\textsuperscript{188}</td>
<td>Suppl. Figs 1 – 3</td>
<td>Each unconnected component is one model</td>
</tr>
<tr>
<td>De Lichtenberg\textsuperscript{23}</td>
<td>1</td>
<td>Each unconnected component is one model. And, the following annotated (yet sparsely connected) components were subdivided into separate models: \textit{(APC, Cdc28-cyclin, MCM/ORC, Sister chromatid cohesion), one model with Tubulin related and SPB, one model with Nucleosome/bud formation, Histones, and DNA replication}</td>
</tr>
<tr>
<td></td>
<td>3a-c</td>
<td>1:1</td>
</tr>
</tbody>
</table>
Acknowledgements

We acknowledge funding from the National Science Foundation (NSF 0425926) and thank the members of the Ideker lab for testing and suggesting improvements to the web interface.

Chapter 5, in full, is a reprint of material as it appeared on November 29 2006 in *Nucleic Acids Research, 2006, vol. 35, D538–D545, Database issue 2007*. Mak, H. Craig; Daly, Mike; Gruebel, Bianca; Ideker, Trey. The dissertation author was one of two primary investigators and authors (with M. Daly) of this paper.
Chapter 6: Conclusions

Physiologically relevant interpretation of high-throughput data

In this dissertation, we sought to elucidate the structure and function of molecular interaction networks that determine gene expression. Large-scale studies of these networks have been made possible by advances in technology for gathering cellular data, such as protein interactions and gene expression profiles, on an unprecedented high-throughput level. Generating network models, however, is challenging because it is often unclear how to meaningfully interpret the wealth of data in ways that are physiologically relevant.

We adopted a four-stage strategy for extracting meaning from high-throughput data: network assembly, network analysis, network refinement, and network cataloging (Figure 6.1). Chapters 2 through 5 of this dissertation are each focused on one stage of this strategy, and our main findings are summarized at the end of Chapter 1. Here, we discuss our work in the context of three themes: harnessing complementary relationships between data, controlling for missing and spurious data, and remaining grounded in biology.

Figure 6.1. Contributions of the dissertation to each stage of the network lifecycle.
Harnessing complementary relationships

Since different experimental techniques measure complementary information about the inner-workings of a cell, we explored methods for data integration. We sought to harness the complementary relationships between data types to produce network models that more accurately represented a biological system. The raw data used for network assembly, analyses, and refinement typically fell into two categories: physical and inferred. Physical interactions represent chemical affinities between molecules, and include interactions between two proteins or interactions between a protein and DNA. Inferred (or functional) interactions represent functional dependencies that one gene or protein has on another, and are observed when a biological system is perturbed, perhaps by an environmental stimulus or by deleting a gene.

In Chapters 2 and 4, we employed the following strategy (Figure 6.1a) to integrate data from gene deletion experiments with a noisy network of physical interactions to identify high confidence regulatory pathways.

First, we articulated the precise biological relationships between data types. Specifically, it has been observed that the majority of physical TF-promoter binding events do not coincide with changes in gene expression\(^\text{197}\). Instead, we proposed that only those physical interactions that link a transcription factor (TF) to a gene whose expression changes when that factor is deleted should be interpreted as a functional regulatory pathway. Thus, a pathway represents one possible physical explanation of an effect observed when a TF is deleted. Such a regulatory pathway might include protein kinases in a signaling cascade and/or transcription factors in a regulatory hierarchy.

Second, we codified these relationships by augmenting and implementing a previously described modeling procedure\(^\text{15}\) as a plugin for Cytoscape\(^\text{11}\), a software
package for network analyses and visualization.

Third, we used our software to search for all instances of these relationships. In Chapter 4, we searched for regulatory pathways that function in cells growing in rich medium. In Chapter 2, we extended this methodology to search for pathways that are responsible for expression changes induced by MMS, a DNA damaging agent. We found 37 direct physical pathways in which an MMS-dependent functional interaction was explained by a TF binding to the promoter of a target. By searching for longer regulatory pathways that contained intermediate genes, we were able to explain 68 additional functional interactions, an almost 3-fold increase. In total, however, pathways through the physical network could only explain about 30% of the functional interactions (105 of 341).

Why were the remaining 70% of functional interactions not explained? Clearly, the biological relationships that we articulated between physical and functional interactions were insufficient for explaining all of the observed gene expression changes. The next generation of transcriptional network models will need to incorporate data representing other regulatory mechanisms such as mRNA degradation, nuclear export, and protein translation.130

Lastly, in Chapter 3, we integrated physical TF-promoter interactions with the genomic locations on the chromosome of the bound promoters. A few yeast TFs are known to bind near the telomeres at chromosome ends. We screened a compendium of yeast TF binding profiles for additional transcription factors that also bound near chromosome ends. We discovered 22 factors, which we named SBTFs (Subtelomere Binding Transcription Factors). For many SBTFs, the subtelomeric binding pattern was dynamic, meaning that the TF appeared to relocate to or away from the subtelomere as physiological conditions shifted.

Our study presented the first systematic evidence that genome position
(i.e. the location of a target gene along the chromosome) may influence patterns of transcription factor binding. This may be representative of a general principle governing the wiring of transcriptional networks. Our results support observations that genome position affects gene expression\textsuperscript{88-93}, and provide possible insights into the regulatory network underlying these expression patterns.

**Controlling for missing data and spurious data**

Missing data includes false negatives (phenomena that could have been measured, but were not, perhaps due to errors or technical limitations) and incomplete coverage (phenomena that were excluded from being measured, perhaps due to the study design). Almost all high-throughput studies performed to-date have been limited by incomplete coverage, primarily due to finite time, dollar, and reagent resources. Noise and experimental biases in high-throughput methods also generate spurious data (false positives), measurements that appear accurate but are not.

Recent studies of protein interaction screens\textsuperscript{38} and microarray expression profiles\textsuperscript{27,30,139} suggest that aggregating data from several studies may be a useful, and perhaps necessary, strategy to control for missing or spurious data during network assembly and analysis (Figure 6.1 a and b). In Chapter 2, we developed a statistical method (the truncated product method) that used TF-promoter binding interactions measured in more than one physiological condition to more accurately identify high-confidence interactions. We also used a Bayesian probabilistic method (deletion-buffering analysis) for combining many expression profiles from deletion mutants to infer functional dependencies. In Chapter 3, we supported our discovery of SBTFs, which was based solely on TF-promoter binding data, with high-throughput protein interactions, gene expression profiles, and phenotyping studies. In Chapter 4, we described a principled method for identifying highly informative follow up
experiments and incorporating the results to refine a network (Figure 6.1 c).

**Remaining grounded in biology**

Bioinformatics and computational biology can be seductively powerful in their ability to find patterns in huge amounts of data. It is important to consider whether the patterns found will be of benefit to the practicing biologist. Here, we discuss our work in the context of potential pitfalls that may undermine the biological utility of computational analyses: *excessive precision, challenges of validation, and limited dissemination.*

*Excessive precision* refers to the danger inherent in the ability of computational methods to quantify a biological system with almost infinitely fine-grained measurements that do not reflect biology reality. For example, it is possible to use statistical methods to analyze a microarray expression profile and rank every single gene from most to least expressed. Because of the inherent noisiness of the measurements, though, it is unlikely that the 3543\(^{\text{th}}\) ranked gene, for instance, is actually expressed at a different level than the 2987\(^{\text{th}}\) gene. Excessive precision is related to *over-fitting,* which refers to statistical models that represent (or ‘fit’) a set of observations too closely. As a result of both over-fitting and excessive precision, the actual behaviors and variations in the underlying biological system are not accurately captured, thereby potentially leading to misleading conclusions. Perhaps the best counter to these pitfalls is experimental validation, which we attempted in Chapters 2, 3, and 4 and discuss in more detail below.

*Challenges of validation.* The accuracy and value of computational studies are sometimes regarded with skepticism by biologists, especially if the conclusions
generated have not been extensively experimentally confirmed. Experimental validation is therefore an important component of computational modeling, but it is particularly challenging because of the overwhelming quantity of biological hypotheses that may be generated. A single computational study may identify hundreds of proteins warranting further research, each of which has a rich history of published literature, each of which may require specialized (and expensive) experimental assays and reagents, and each of which could lead to focused studies that might consume an individual investigator for weeks, months, or even years.

Our validation strategy in Chapter 4 identifies the theoretically optimal experiments that will provide the most useful data for validating and refining the network (Figure 6.1 c). However, follow-up experiments may still be time-consuming and expensive, and ideally, we would want to pursue all hypotheses generated by a computational analysis. Thus, going forward, it might be worthwhile to explore technology for automating, miniaturizing, parallelizing, or otherwise reducing the cost of follow-up experiments. It is an open question as to whether computational studies can be explicitly designed to generate results that are amenable to low-cost, rapid verification.

Limited dissemination. The biological utility of network models is currently limited because they are typically embedded in figures, tables, or supplementary information in the primary published literature. As a result, the models remain largely inaccessible to the biologists who have the best knowledge to interpret them.

To address these obstacles, we built the CellCircuits database of network models in Chapter 5. In its initial release, CellCircuits provided an online interface for searching and viewing over 1000 network models from eleven previously published studies. An online database makes models more accessible to the research community
and perhaps will encourage validation and as well as the assembly of the next
generation of better networks (Figure 6.1d).

CellCircuits also enabled us to analyze, for the first time, a diverse collection
of network models as a whole. Based on these meta-analyses, we identified genes
and pathways that rarely appear in network models and that therefore might be
good targets for additional experimental or computational analyses. Just as existing
databases of DNA sequences, protein structures, and microarray profiles have proven
valuable, CellCircuits demonstrates the potential value of a database of network
models.

**Beyond wiring diagrams**

Much of this dissertation has focused on building transcriptional networks to
answer “What regulates gene X?” and “What does gene X regulate?” Knowing what
is connected to what – the ‘wiring diagram’ of the cell – is important, but it provides
little insight into how the system works and why is it wired that way. Indeed, network
models promise to illuminate how complex behaviors emerge from collections
of individual interactions. To move beyond a gene- and protein-centric view,
we propose that a future challenge of computational biology will be to formulate
abstractions that represent instances of known biological patterns or mechanisms.

For example, it is known than transcription factors activate or repress target
genes via many regulatory mechanisms that include nuclear translocation, mRNA
degradation, post-translational modification, binding with co-factors, altered affinity
for as sequence motif, localization within the nucleus, and differential expression of
the factor itself. Focused experimental studies have revealed specific cases where
each regulatory mechanism is used, but few system-wide studies have explored
the extent that each mechanism is used in an organism, how multiple mechanisms
concurrently affect the activity of a single regulatory protein, or whether different organisms have evolved to prefer some regulatory mechanisms over others.

To state the question another way: which TFs regulate gene expression by which mechanisms, and why? Future research might first devise methods for identifying TFs that use each mechanism. For instance, nuclear localization signals embedded within the amino acid sequences of TFs could be to classify factors that are translocated to the nucleus\(^\text{199}\). Gene promoters that contain overlapping, repeated, inverted, or co-occurring binding sites could be used to identify TFs that function together\(^\text{70}\). Protein sequence and structural similarity could be used to identify TFs that function as homo- or hetero-dimers\(^\text{19}\). Expression profiles could be used to identify TFs that are present in the cell at different times\(^\text{200}\). Protein modification assays could be used to identify TFs that are regulated by phosphorylation, ubiquitination, or other post-translational modifications\(^\text{201}\).

Algorithms could then be developed to map regulatory mechanisms to TFs, target genes, and therefore specific biological pathways. Such knowledge might allow us to globally characterize transcriptional networks so that, in addition to knowing that gene X regulates gene Y, we also know how, and thus perhaps why, that regulation occurs.
**Final remarks**

We anticipate that our ability to collect biological measurements will continue to grow in both the scale and breadth. Concurrently, the need for computational methods to interpret these data will grow as well. The individual studies performed in this dissertation have addressed specific stages in the network lifecycle. But we also recognize that the stages are interconnected. New networks are assembled as a result of analyses. A catalog of network models facilitates their refinement. Refinement and analysis are steps during which new data can be incorporated to assemble new networks.

What underlies the interrelationships between these stages? Perhaps it is that biological discovery is an iterative, incremental process: advances towards solving disparate problems proceed in lockstep. Key pieces to one puzzle are provided by another. As such, we see value in future research that targets the individual challenges of network assembly, analysis, refinement, and cataloging, as well as the boundaries between them, which are blurred and yet rich in potential for crafting ever more comprehensive models of life.
References


