Title

Integrating physical and genetic interaction networks for biological pathway discovery

Permalink

https://escholarship.org/uc/item/3fn2t0t4

Author

Bandyopadhyay, Sourav

Publication Date

2010

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Integrating Physical and Genetic Interaction Networks for Biological Pathway Discovery

A dissertation submitted in partial satisfaction

of the requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Sourav Bandyopadhyay

Committee in charge:

Professor Trey Ideker, Chair
Professor Vineet Bafna, Co-chair
Professor Richard Kolodner
Professor Bing Ren
Professor Steven Wasserman

2010

The dissertation of Sourav Bandyopadhyay is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

Co-chair

____

Chair

University of California, San Diego

2010

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTAL FIGURES

LIST OF SUPPLEMENTAL TABLES

ACKNOWLEDGEMENTS

I would first like to acknowledge the support of my advisor, Dr. Trey Ideker, without whom none of this work would have been possible. From our very first meeting he has treated me as an equal and always valued my opinion, a trait I hope I can carry on into my future pursuits. Finally, he provided a sense of enthusiasm and dedication to research which was a source of inspiration. I would also like to thank Dr. Nevan Krogan who, in the last half of my graduate career was instrumental in providing a fresh perspective on my mostly computational work as well as acting as a catalyst for my transition into an experimental biologist.

I also owe a great debt to my many co-authors for their valuable scientific work. Dr. Ryan Kelley was the bioinformatics work-horse for a number of projects and his expertise in algorithms and programming was instrumental in seeing these projects through to publication. Dr. Assen Roguev is a brilliant scientist whose talents led to the development of an E-MAP in S. *pombe*, the analysis of which was a tremendous learning experience for me. I would also like to thank Dr. Sumit Chanda for including me on multiple ground-breaking studies involving HIV and Influenza as well as frank discussions of the highs and lows of academic research.

Chapter 2, in full, is the following manuscript currently under submission,

Bandyopadhyay S, Chiang C, Srivastava J, Gersten M, White S, Bell R, Kurschner C, Martin CH, Smoot M, Sahasrabudhe S, Barber DL, Chanda SK, Ideker T. A *Core Network of Human MAPK Interactions.* Submitted.

The dissertation author is the sole first author on this work, responsible for study design and data analysis.

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

Bandyopadhyay S, Sharan R, Ideker T. *Systematic identification of functional orthologs by protein network comparison.* **Genome Research** 2006; 16(3):428-35.

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

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

Bandyopadhyay S, Kelley RM, Krogan NJ, Ideker T. *Functional maps of protein complexes from quantitative genetic interaction data.* **PLoS Computational Biology** 2008; 18;4(4).

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

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

Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins SR, Qu H, Shales M, Park H, Hayles J, Hoe K, Kim D, Ideker T, Grewal SI, Weissman JS, Krogan NJ. *Conservation and Rewiring of Functional Modules Revealed by an Epistasis Map in Fission Yeast.* **Science** 2008;322(5900):405-10.

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

Chapter 6, in full, is the following manuscript currently submitted,

Bandyopadhyay S, Mehta M, Kuo D, Sung M, Jaehnig E, Chuang R, Bodenmiller B, Licon K, Copeland W, Shales M, Fiedler D, Shokat KM, Kolodner RD, Huh W, Aebersold R, Keogh MC, Krogan NJ, Ideker T. *DNA-damage induced rewiring of protein signaling revealed by a conditional epistatic interaction map (cE-MAP)*. Submitted.

The dissertation author is the sole first author on this work, responsible for study design and data analysis.

VITA

2002	Bachelor of Science Molecular Biology	University of Wisconsin-Madison
2002	Bachelor of Science Computer Science	University of Wisconsin-Madison
2010	Doctor of Philosophy	University of California at San Diego

PUBLICATIONS

Bandyopadhyay S, Mehta M, Kuo D, Sung M, Jaehnig E, Chuang R, Bodenmiller B, Licon K, Copeland W, Shales M, Fiedler D, Shokat KM, Kolodner RD, Huh W, Aebersold R, Keogh MC, Krogan NJ, Ideker T. *DNA-damage induced rewiring of protein signaling revealed by a conditional epistatic interaction map (cE-MAP)*. Submitted.

Bandyopadhyay S, Chiang C, Srivastava J, Gersten M, White S, Bell R, Kurschner C, Martin CH, Smoot M, Sahasrabudhe S, Barber DL, Chanda SK, Ideker T. A *Core Network of Human MAPK Interactions.* Submitted.

Konig R, Stertz S, Zhou Y, Inoue A, Hoffmann HH, Bhattacharyay S, Alamares J, Tscherne DM, Ortigoza MB, Liang Y, Gao Q, Andrews SE, Bandyopadhyay S, De Jesus P, Tu B, Pache L, Shih C, Orth A, Bonamy G, Miraglia L, Ideker T, Garcia-Sastre A, Young JAT, Palese P, Shaw ML, Chanda SK. *Human Host Factors Required for Influenza Virus Replication.* **Nature** 2009; Epub

Fossum E, Friedel CC, Rajagopala SV, Titz B, Baiker A, Schmidt T, Kraus T, Stelberger T, Rutenberg C, Suthram S, Bandyopadhyay S, Rose D, Von Brunn A, Uhlmann M, Zeretzke C, Dong Y, Boulet H, Koegl M, Bailer SM, Koszinowski U, Ideker T, Uetz P, Zimmer R, Haas J. *Evolutionarily conserved herpesviral protein interaction networks.* **PLoS Pathogens** 2009;5(9).

Bushman FD, Malani N, Fernandes J, D'Orso I, Cagney G, Diamond TL, Zhou H, Hazuda DJ, Espeseth AS, Konig R, Bandyopadhyay S, Ideker T, Goff S, Krogan N, Frankel A, Young JAT, Chanda SK. *Host cell factors in HIV replication: meta-analysis of genome-wide studies.* **PLoS Pathogens** 2009; 5(5).

Wilmes G*, Bergkessel M*, Bandyopadhyay S, Chan A, Braberg H, Shales M, Collins SR, Whitworth BG, Kress TL, Weissman JS, Ideker T, Guthrie C, Krogan NJ. Cover Article: *A Genetic Interaction Map of RNA Processing Factors Reveals Links Between Sem1/Dss1-Containing Complexes and mRNA Export and Splicing.* **Molecular Cell** 2008;32(5):735-46. *Equal Contribution

Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins SR, Qu H, Shales M, Park H, Hayles J, Hoe K, Kim D, Ideker T, Grewal SI, Weissman JS, Krogan NJ. *Conservation and Rewiring of Functional Modules Revealed by an Epistasis Map in Fission Yeast.* **Science** 2008;322(5900):405-10.

Konig R, Zhou Y, Elleder D, Diamond TL, Bonamy GMC, Irelan JT, Chiang C, Tu BP, De Jesus PD, Lilley CE, Seidel S, Opaluch AM, Caldwell JS, Weitzman MD, Kuhen KL, Bandyopadhyay S, Ideker T, Orth AP, Miraglia LJ, Bushman FD, Young JA, Chanda SK. *Global Analysis of Host-Pathogen Interactions that Regulate Early-Stage HIV-1 Replication.* **Cell** 2008;135(1):49-60.

Bandyopadhyay S, Kelley RM, Krogan NJ, Ideker T. *Functional maps of protein complexes from quantitative genetic interaction data.* **PLoS Computational Biology** 2008; 18;4(4).

Beyer A, Bandyopadhyay S, Ideker T. *Integrating physical and genetic maps: from genomes to interaction networks.* **Nature Reviews Genetics** 2007; 8(9):699-710.

Bandyopadhyay S, Sharan R, Ideker T. *Systematic identification of functional orthologs by protein network comparison.* **Genome Research** 2006; 16(3):428-35.

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

Canet-Aviles RM, Wilson MA, Miller DW, Ahmad R, McLendon C, Bandyopadhyay S, Baptista MJ, Ringe D, Petsko GA, Cookson MR. *The Parkinson's disease protein DJ-1 is neuroprotective due to cysteine-sulfinic acid-driven mitochondrial localization.* **PNAS** 2004;101(24):9103-8.

Bandyopadhyay S, Cookson MR. *Evolutionary and functional relationships within the DJ1 superfamily.* **BMC Evol Biol** 2004; 4(1):6.

ABSTRACT OF THE DISSERTATION

Integrating Physical and Genetic Interaction Networks for Biological Pathway Discovery

by

Sourav Bandyopadhyay

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego 2010

Professor Trey Ideker, Chair

Professor Vineet Bafna, Co-chair

The goal of understanding complex biological systems and how they are perturbed to cause disease has long been a central focus of biology. The past decade has seen the creation and maturation of a number of new technologies designed to study biological pathways on a genome-wide scale. Rather than obtaining information about the function of one gene or protein at a time, such approaches can offer insight into the activity of every gene and protein in the cell all in the context of one experiment.

One fundamental mode of gathering biological insight is through identifying which proteins in the cell interact physically, such as those which form protein complexes or biochemical pathways. Techniques such as yeast-two hybrid and co-

immunoprecipitation followed by mass spectrometry allow the determination of a physical interaction map which details binding interactions between proteins on a large scale. Another fundamental mode of biological discovery is through assaying genetic interactions which arise when mutations in two genes produce a phenotype that is surprising in light of each mutation's individual effects. For example a synthetic lethal genetic interaction is indicated when deletions in two genes which are not essential for viability cause lethality when deleted together. Genetic interaction maps can be determined in high-throughput via SGA (Synthetic Genetic Array) technology.

In Chapter 2 we derive and analyze a large physical protein interaction map centered on a set of human protein kinases and show how biological insight can be derived from such large-scale screens. In Chapter 3, we develop methods for the comparison of such physical protein interaction maps between species in order to identify proteins whose function is conserved throughout millions of years of evolution.

In Chapter 4 we develop algorithms to integrate both physical and genetic interactions together for the purpose of biological pathway discovery. Moreover, our approaches create maps of genetic interactions that provide a picture of the global organization of pathways and complexes within the cell, which we apply to create a map of functional relationships among protein complexes involved in chromosomal biology. In Chapter 6, we apply this approach in two different yeast species and discover that while physical protein interactions are largely conserved across species, many genetic interactions are rewired which gives us valuable insight into pathway architecture. Finally in Chapter 7, we focus on the discovery of genetic interactions involved in the DNA damage response by assaying how different gene mutants respond to a drug which causes

DNA damage and then demonstration how this elucidates pathways involved in this process.

Chapter 1. Introduction

A central goal of biology is to understand the web of molecular interactions which give rise to cellular form and function. Our understanding of biology took a great leap forward with the elucidation of the central dogma of molecular biology. DNA forms an inherited genetic code delineating thousands of genes which are transcribed into mRNA which are then ultimately translated to make proteins. This discovery has served to unify several disciplines of biology which focus on different aspects of this dogma. On the level of DNA, genetics is the study of genes and heredity to understand how variations in DNA can give rise to particular phenotypes such as disease. Molecular biology seeks to understand the complex web of interactions between molecules and proteins which govern cellular function. While the fields may be distinct disciplines, they are highly intertwined. For example, mutations in the DNA sequences of genes can be reflected in mutations in their protein products which affect their molecular function through changes in interactions with other molecules. In the worst case, such changes in the function of proteins through mutations in DNA can ultimately lead to disease and even lethality.

In the post-genomic era, mapping of physical and genetic networks have become an effective approach for understanding cellular function. Physical interactions dictate the architecture of the cell in terms of how direct associations between molecules constitute protein complexes, signal transduction pathways, and other cellular machinery. Genetic interactions define functional relationships between genes, which give insight into how this physical architecture translates into phenotype. Genetic interactions report the extent to which a phenotype caused by one mutation is affected by another mutation, indicating pairs of genes which jointly affect a phenotype. The complementarities between physical and genetic interactions has been strikingly demonstrated in yeast, for which less than 1% of genetic interactions of the synthetic lethal type are also observed physically[1]. It has also been exploited numerous times in classical genetics and biochemistry, in which a great many pathways have been understood only through integration of both physical and genetic interactions (the LIN-12/Notch signaling pathway[2] and the actin cytoskeleton[3] are excellent examples). The modes by which genetic and physical interactions complement one another have not yet been fully elucidated; however, a growing body of work has begun to reveal a complex but concrete set of principles governing their relationships. The work in this dissertation illustrates how interactions of these different types can be combined to assemble a more comprehensive picture of biological systems.

Protein-protein interactions mediate most all of the processes in the cell. In most cases, proteins act in concert with each other as part of pathways or larger molecular assemblies called complexes. Systematic discovery of these physical associations can expand our knowledge base and our understanding of cellular behavior in terms of outlining all the possible reactions or catalytic steps in the cell. Until recently, protein interactions were mainly discovered by small-scale methods such as co-immunoprecipitation and FRET microscopy which reveal only a small number of protein interactions in one experiment. Now, high-throughput techniques like yeast two-hybrid (Y2H)[4] and tandem affinity purification coupled with mass spectrometry (TAP-MS)[5] reveal protein interactions at the level of the whole proteome resulting in the generation of a large number of proteins interactions. In TAP-MS studies proteins are used as bait in a co-immunoprecipitation assay and the pulled down proteins are separated and identified

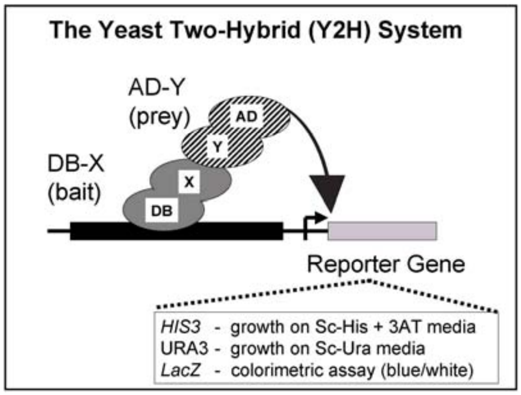

The Yeast Two-Hybrid (Y2H) System
AD-Y
(prey)
AD
Y
DB-X
(bait)
X
DB
Reporter Gene
HIS3 - growth on Sc-His + 3AT media
URA3 - growth on Sc-Ura media
LacZ - colorimetric assay (blue/white)

Figure 1.1: The yeast two-hybrid system.

The bait protein (X) is fused to the DNA binding domain (DB) of the transcription factor GAL4. The prey protein (Y) is fused to the transcription activation domain (AD) of a transcription factor. An interaction between bait and prey hybrid proteins (X-Y) results in the assembly of a functional transcription factor which drives the expression of a reporter gene (either HIS3, URA3, or lacZ).

using mass spectrometry. Y2H is a technique which is based on the functional reconstitution of an intact transcription factor that activates reporter gene expression (Figure 1.1). Y2H has been widely used to map interactomes of eukaryotic species such as that of S. *cerevisiae*[6,7], C. *elegans*[8] , D. *melanogaster*[9], and H. *sapiens*[10,11].

In Chapter 2 I describe the generation and analysis of a large scale Y2H network focused on human Mitogen Activated Protein Kinase (MAPK) pathways which form the backbone of signal transduction within the mammalian cell. I applied a systematic experimental and computational approach to map thousands of interactions among human MAPK-related proteins and assemble them into functional modules. Using this physical mapping approach I was able to predict a number of proteins which function in the MAPK pathway. Using a genetic technique of siRNA interference, I was able to show that many of these proteins function in the MAPK pathway. This study illustrates an approach for probing signaling pathways based on functional refinement of experimentally-derived protein interaction maps.

Evolutionary conservation is a fundamental principle in biology that is widely used to infer functional relationships among species. For example, conservation of protein/gene sequences across species is used to make function and domain assignments[12]. In the same vein, comparing protein interaction networks across species can highlight evolutionarily conserved and diverged pathways. In the MAPK study I was able to identify hundreds of human protein interactions which were conserved across species. A conserved interaction is identified if the two interacting human proteins have significant homology with two proteins in another species which have also been reported to interact. Based on a conserved interaction, one possible conclusion is that the two proteins encompass the same function in the two species. In Chapter 3, I describe an approach to compare interaction maps across different species in order to identify genes which are operating in a functionally identical fashion (i.e. functional orthologs). This approach is able to identify functional orthologs among large gene families where sequence-based information might not be adequate to identify which proteins across species are the most similar functionally.

In contrast to physical interactions, genetic interactions represent functional relationships between genes, in which the phenotypic effect of one gene is modified by another[13,14]. Genetic interactions are identified by comparing the effect of mutating each gene individually to the effect of the double mutant. For example, “synthetic sickness” (or in the extreme, “synthetic lethality”) is a negative genetic interaction in which the measured phenotype is growth, and mutating both genes results in slower growth than expected from either mutation alone (termed SSL, or synthetic sick/lethal). In yeast, large networks of genetic interactions are being measured using homozygous gene knockouts based on Synthetic Genetic Arrays (SGA)[15], diploid-based Synthetic Lethality Analysis on Microarrays (dSLAM)[16]. All of these methods allow the phenotypic consequences of double-mutant combinations to be assayed in high-throughput formats[17].

In its simplest form, Synthetic genetic array analysis involves a series of replica-pinning procedures, in which mating and meiotic recombination are used to convert an input array of single mutants into an output array of double mutants. The final transfer step results in an ordered array of double-mutant haploid strains, the growth rates of which can be quantitatively assessed. The SGA screening method has recently been complemented by a method termed E-MAP (Epistatic Miniarray Profile), which can quantify the wider spectrum of possible genetic relationships[13,14]. As outlined in Figure 1.2 and 1.3, epistasis also includes positive interactions where the double mutant grows better than expected from the growth of the two single mutants. Positive interactions (in addition to negative interactions) can provide valuable information about pathway organization based on pairwise genetic relationships. The full pattern of genetic interactions for a particular mutation can also provide more information[13,18-20] than individual SSL interactions, as it consists of measurements of the mutant phenotype in many different mutant backgrounds. These patterns of interactions can provide high-resolution phenotypic signatures which can be used to identify genes whose mutation has similar impact on cellular physiology. E-MAPs have been used to study genetic interactions in yeast among discrete subsets of proteins involved in the secretory

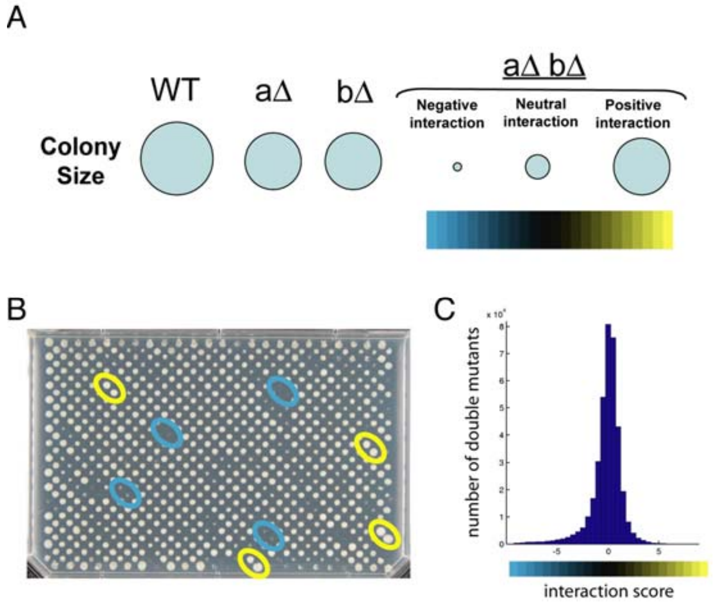


Figure 1.3: Uncovering Genetic Interactions

(A) The continuous spectrum of genetic interactions identified in E-MAP experiments. A neutral interaction is obtained when the growth of the double mutant corresponds to the product of the growth of the two single mutants. Negative (or synthetic sick/lethal) genetic interactions, represented in blue, arise when the double mutant grows at a rate that is less than the product of the two single mutants. Positive (epistatic or suppressive) interactions, represented in yellow, correspond to cases when the double mutant grows better than is expected based on growth of the two single mutants. **(B)** A representative plate of double mutant colonies pinned in duplicate. Examples of both positive (yellow circles) and negative (blue circles) interactions are highlighted. **(C)** Distribution of the sizes of the double mutant strains that we have collected to date. Negative and positive values correspond to negative and positive genetic interactions, respectively.

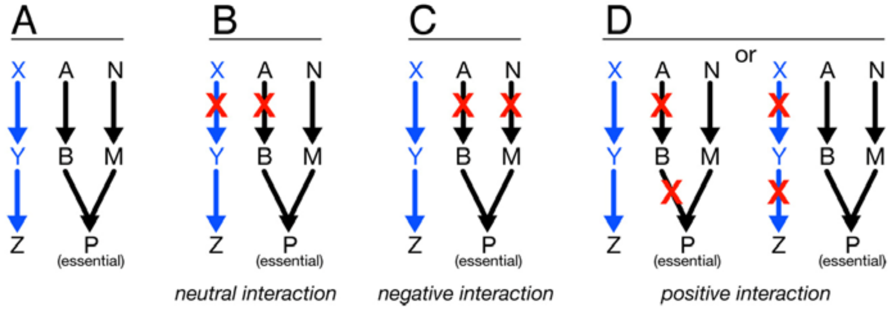


Figure 1.2: Pathways and Genetic Interactions

The various classes of E-MAP interactions can be illustrated in terms of two functionally distinct pathways responsible for cell function, one linear (blue) and one bifurcated (black). In this scenario, product Z is not essential for cell viability whereas product P is. In the wild type cell, both pathways are functional **(A)**. When genes Z and A are mutated **(B)**, this would manifest itself as a neutral genetic interaction since the pathways are unrelated. However, if both arms of this second pathway are impaired **(C)**, cell growth is affected, resulting in a negative interaction. When two genes from the same pathway (or arm of a pathway, or protein complex) are mutated, the result is often a positive interaction **(D)**. An E-MAP can detect all of these classes of genetic interaction.

pathway[13] and chromosomal biology[21].

Although an E-MAP provides a great deal of information about the genes within it, it is difficult to immediately organize this information into cellular pathways. Previous work has shown that genetic interactions fall predominantly 'within' and 'between-pathways'[18] (Figure 1.3). a notion which I have used in Chapter 4 to create an algorithm which partitions genetic interaction maps into maps of complexes and pathways. Not only is this approach effectively able to summarize genetic network information and identify functional connections between known pathways, it can also be used to discover novel pathways based on a combination of physical and genetic interaction evidence.

Just as conservation of protein-protein interaction networks can be used to identify genes which act through similar mechanisms across species, genetic interactions can be used to probe the conservation and divergence of gene functions. While nearly all previous genetic interaction screening has been performed in the budding yeast, S. *cereviasiae*, myself and others have pioneered a system for the rapid generation of such maps in the fission yeast S. *pombe*[22]. Chapter 5 describes the generation of a large genetic interaction map in fission yeast centered around chromosomal biology. A comparison of genetic interactions between the two yeasts revealed that while genetic interactions were significantly conserved between the two species, they were not as conserved as physical interactions were. Furthermore, among protein complexes we found significant rewiring of their genetic interaction inter-connections. This study had both positive and negative aspects. On a positive note, I discovered that genetic interactions were significantly conserved between species, an observation which supports a model of discovery of genetic interaction networks in human based on mapping in model organisms. The downside was that, given the similarity of the two species, the level of conservation was much less than we expected suggesting that this model of discovery might not be as effective as we had hoped.

The rewiring of genetic interaction networks between species led to questions about how these networks were rewired by different stresses in the same species. Chapter 6 describes the generation of a conditional genetic interaction map in budding yeast. By perturbing the cell with the DNA damaging agent methyl methanesulfonate (MMS), I sought to identify genetic interactions which point to pathways required for growth in response to DNA damage. This study puts forth a wealth of information about the genetic interactome of kinases, phosphatases and transcription factors in the cell and highlights their functional interdependencies. Since DNA damage is the primary driving event in the development of cancer, the goal is to uncover pathways which reveal mechanisms governing DNA damage and repair and could make future targets for anti-cancer drugs. In addition, I was able to establish a general framework through conditional interaction maps to probe cellular pathways which are critical for drug and stress responses.

With the emergence of physical and genetic interactions maps for a variety of species, my work points to new ways in which these maps can be constructed and used to create models of biological function which subsequently be tested in the laboratory. These methods include comparing networks across species, generating maps of complexes and pathways and their inter-relationships, and understanding the role of perturbations to these networks for pathway discovery. It is my hope that my systems-level contributions to this field will enable the greater understanding of the architecture of cellular networks and that my biological contributions spur on further research for years to come.

Bibliography

1. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Global mapping of the yeast genetic interaction network. *Science* 2004;303:808-13.

2. Greenwald I. LIN-12/Notch signaling in *C. elegans*. In: The*Celegans*ResearchCommunity, ed. WormBook: WormBook, doi/10.1895/wormbook.1.10.1, http://www.wormbook.org, August 4, 2005.

3. Botstein D, Amberg D, Mulholland J, Huffaker T, Adams A, Drubin D, Stearns T. The Yeast Cytoskeleton. In: Pringle J, Broach J, Jones E, eds. The Molecular and Cellular Biology of the Yeast *Saccharomyces*: Cell Cycle and Cell Biology. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 1997.

4. Fields S, Sternglanz R. The two-hybrid system: an assay for protein-protein interactions. *Trends Genet* 1994;10:286-92.

5. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;415:141-7.

6. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 2001;98:4569-74.

7. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 2000;403:623-7.

8. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M. A map of the interactome network of the metazoan C. elegans. *Science* 2004;303:540-3.

9. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL, Jr., White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM. A protein interaction map of Drosophila melanogaster. *Science* 2003;302:1727-36.

10. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;437:1173-8.

11. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005;122:957-68.

12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389-402.

13. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 2005;123:507-19.

14. Collins SR, Schuldiner M, Krogan NJ, Weissman JS. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol* 2006;7:R63.

15. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 2001;294:2364-8.

16. Ooi SL, Shoemaker DD, Boeke JD. DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nat Genet* 2003;35:277-86.

17. Boone CB, H. Andrews, B. Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics* 2007.

18. Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 2005;23:561-6.

19. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS. Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol Syst Biol* 2005;1:2005 0026.

20. Krogan NJ, Keogh MC, Datta N, Sawa C, Ryan OW, Ding H, Haw RA, Pootoolal J, Tong A, Canadien V, Richards DP, Wu X, Emili A, Hughes TR, Buratowski S, Greenblatt JF. A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol Cell* 2003;12:1565-76.

21. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z, Brown GW, Ingles CJ, Emili A, Allis CD, Toczyski DP, Weissman JS, Greenblatt JF, Krogan NJ. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 2007;446:806-10.

22. Roguev A, Wiren M, Weissman JS, Krogan NJ. High-throughput genetic interaction mapping in the fission yeast Schizosaccharomyces pombe. *Nat Methods* 2007;4:861-6.

Chapter 2. A Core Network of Human MAPK Interactions

Abstract

Mitogen Activated Protein Kinase (MAPK) pathways form the backbone of signal transduction within the mammalian cell. Here, we apply a systematic experimental and computational approach to map 2,269 interactions among human MAPK-related proteins and assemble them into functional modules. A core network of 641 interactions is supported by multiple lines of evidence including conservation with yeast. Using siRNA knockdowns, we interrogate novel members of the network to identify those that can functionally modulate signaling through p38 or AP-1. We uncover the Na-H exchanger NHE1 as a scaffold for a novel set of MAPKs, link HSP90 chaperones to MAPK pathways, and identify MUC12 as the human analogue to the yeast signaling mucin Msb2. This study illustrates an approach for probing signaling pathways based on functional refinement of experimentally-derived protein interaction maps.

Introduction

The MAPK pathways are a collection of protein signaling cascades stimulated by a wide variety of extra-cellular signals, including growth factors, cytokines, and environmental stress [1,2]. Upon activation, MAPK pathways regulate a large number of fundamental cellular functions, including differentiation, proliferation, and apoptosis, through the activation of specific transcription factors and other regulatory proteins[5]. Because of this central role in signal transduction, MAPK pathways have been repeatedly implicated in the pathogenesis of cancer and autoimmune diseases, leading to their selection as targets for drug development[6]. MAPK pathways are also well conserved over the eukaryotic kingdom from yeast to man, enabling study of their structure and kinetics through genetic analysis of model organisms[2].

A MAPK pathway minimally consists of three sequentially-activated MAPK family members: a MAPK kinase kinase (MAP3K) which activates a MAPK kinase (MAP2K) which, in turn, activates a MAP kinase (MAPK). However, the situation can be further complicated by inhibitor proteins, alternate MAPK and scaffold protein functions, and by extensive cross-talk between the individual MAPK cascades[8]. Therefore, it has been suggested that a systems-level approach will ultimately be necessary to map the complete MAPK network and to unravel its function[5,9].

To help unravel the complexity of MAPK signaling, we developed a combined experimental and computational approach based on mapping and functional exploration of protein interaction networks. We screen for physical interactions involving MAPK signaling proteins, establish several benchmarks of data quality, derive a high-confidence network based on evolutionary conservation, and begin developing this network into a compendium of signaling modules. The interactions guide the discovery of components of novel kinase scaffolds, including NHE1, as well as identify a conserved signaling cascade mediated by the signaling mucin, MUC12.

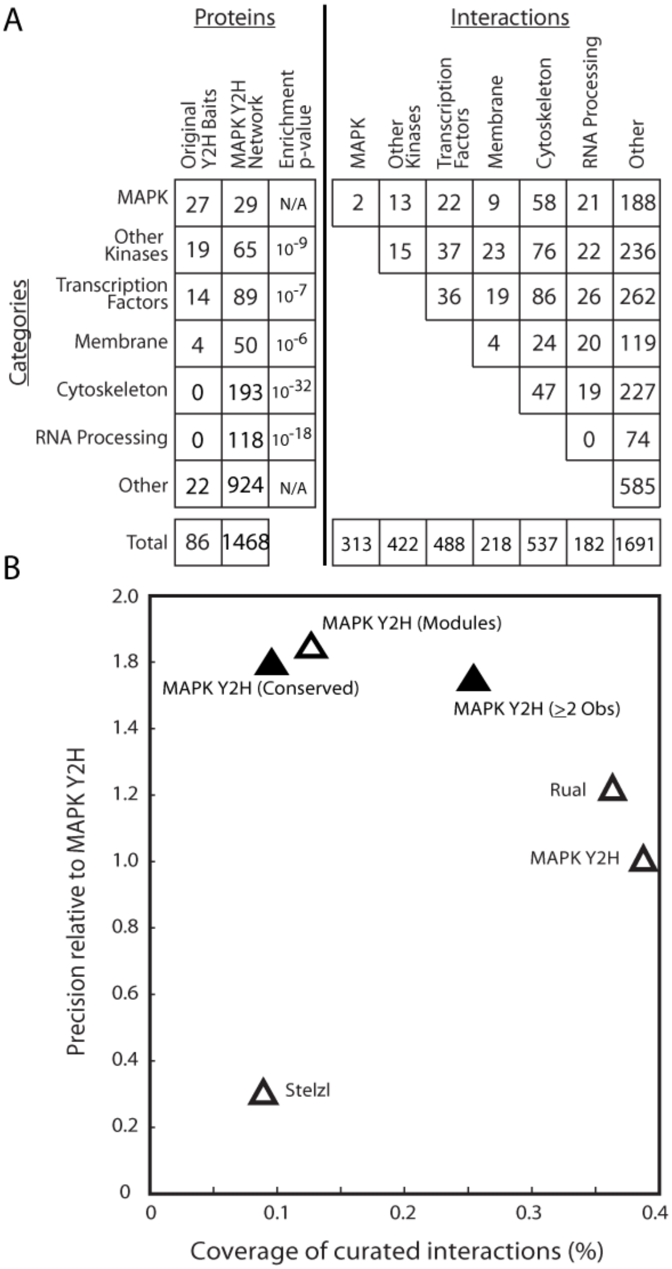
Results and Discussion

Characterization of a MAPK interactome

We assembled a human MAPK network comprised of protein interactions identified through a two-stage yeast two-hybrid (Y2H) screen. A Y2H network was derived from 86 MAPK-related bait proteins (Supplementary Figure S1) selected by literature curation (see Methods). Included in these baits were 46 kinases (of which 27 were known MAP-family kinases), 14 MAP-activated transcription factors downstream of MAPK signals, and four proteins associated with membrane receptors (Figure 2.1A). This effort yielded 1,496 protein examined 21 secondary baits that had been detected as preys in the first stage (based on their availability as bait cDNA clones). This round identified an additional 786 interactions for a total of 2,269 unique interactions among 1,468 proteins (termed the MAPK Y2H network) (Supplementary Table S1).

Analysis of the MAPK Y2H network revealed 313 interactions involving MAP-family kinases and 422 involving other kinases. After removing the original baits, the MAPK Y2H network was enriched for protein families known to be critical in MAPK signaling such as membrane proteins and transcription factors suggesting numerous possible upstream and downstream components of MAPK cascades (Figure 2.1A). The network was also highly enriched for proteins involved in cytoskeletal organization and RNA binding and processing. The significant number of cytoskeletal proteins identified in this study suggests the active regulation of microtubule dynamics by MAP kinases or the use of the cytoskeleton as an organizational scaffold in MAPK signaling; it has been postulated that up to one third of MAP-family kinases are associated with the microtubule cytoskeleton[10].

-



Categories	Proteins			Interactions						
	Original Y2H Baits	MAPK Y2H Network	Enrichment p-value	MAPK	Other Kinases	Transcription Factors	Membrane	Cytoskeleton	RNA Processing	Other
MAPK	27	29	N/A	2	13	22	9	58	21	188
Other Kinases	19	65	10^{-9}		15	37	23	76	22	236
Transcription Factors	14	89	10^{-7}			36	19	86	26	262
Membrane	4	50	10^{-6}				4	24	20	119
Cytoskeleton	0	193	10^{-32}					47	19	227
RNA Processing	0	118	10^{-18}						0	74
Other	22	924	N/A							585
Total	86	1468		313	422	488	218	537	182	1691

Figure 2.1: Functional Properties of the MAPK Network.

(A) Breakdown of network proteins (columns 1-3) and protein interactions (columns 4-10) by functional category. Enrichment p-value is based on the probability of identifying an equal or greater number of proteins in the same category at random assessed via the hypergeometric test with a background of 30,000 genes. (B) Different human protein interaction data sets were analyzed to assess their relative precision versus coverage. Protein interactions supported by additional evidence (filled triangles) were combined to form the core network.

protein interactions among 1,096 proteins. As follow-up to this primary screen, we

As a first method for evaluating the quality of the human MAPK Y2H network, we scored its enrichment for a reference set of literature-curated interactions over a random set of interactions (Methods). We analyzed this network in comparison to previous Y2H networks for human in Rual et al.[11] and Stelzl et al.[12]. The entire MAPK Y2H network was within the range of quality seen in the two previous human Y2H studies, with precision approximately three-fold higher than Stelzl et al. and 83% that of Rual et al. (Figure 2.1B). The coverage of the MAPK Y2H was approximately the same as for Rual et al. and approximately four-fold higher than that of Stelzl et al.

As an independent test of network quality, we determined the extent to which members of the MAPK network had been previously shown to function in MAPK signaling. In this regard, we found that the network (with 86 original baits removed) was highly enriched for proteins that were shown to be phosphorylated in response to stimulation of HeLa cells with epidermal growth factor[3] (EGF; 429 of 2,089 phosphorylated proteins; $P<10^{-170}$). Second, the MAPK network contained 734 genes with homologs in *Drosophila*, 92 of which were shown to be required for the activation of extracellular signal-regulated kinases (ERK) by an unbiased RNAi screen in fly [13] ($P<10^{-13}$).

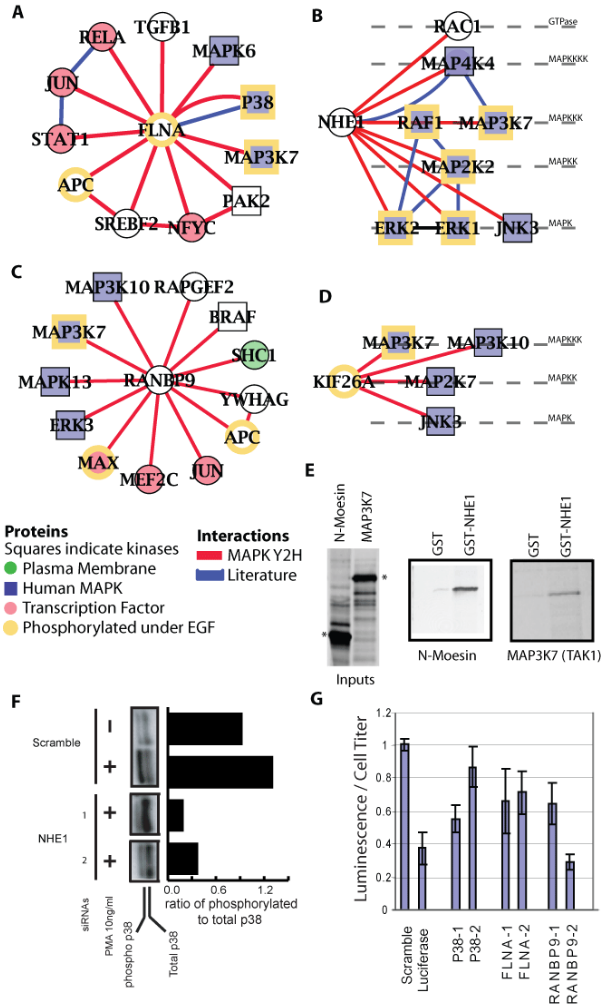


Figure 2.2: Protein Subnetworks Reveal Known and Putative MAPK Scaffolds.

A screen for proteins whose network neighborhoods are enriched for kinases reveals a number of novel and known scaffolding proteins. Such neighborhoods are shown for (A) Filamin protein FLNA, (B) the Na-H exchanger NHE1, (C) RAN binding protein RANBP9, and (D) the kinesin family member KIF26A. Newly-identified Y2H interactions (red) and interactions from literature are shown (blue). Proteins are colored based on their annotation as membrane (green), MAP-family kinase (blue), transcription factor (red), or phosphorylated when stimulated with EGF (yellow border) [3]. (E) Binding of invitro-translated MAP3K7 (TAK1) (metabolically labeled with 35S-methionine) to GST tagged C-terminus of NHE1 or GST alone. N-moesin, a previously identified NHE1 binding partner is used as a positive control [4]. Expressed input proteins used for in-vitro binding assays are marked with asterisks. (F) Phosphorylated (pp38) and total levels of p38, assayed with and without PMA stimulation and two different siRNA knockdowns of NHE1. The bars quantify the pp38 / p38 ratio in each case by image analysis of the two bands in the Western blot. Both NHE1 knockdowns markedly reduce the pp38 / p38 ratio. As a negative control, we observed nearly equal amounts of pp38 to p38 in the absence of PMA stimulation (PMA –). As a positive control, PMA stimulation with a scrambled siRNA message induced the majority of p38 towards the phosphorylated state (Scramble, PMA +). (F) Using a luciferase reporter fused to the AP-1 gene, we tested the ability of various siRNAs to reduce AP-1 activation when stimulated with PMA [7]. As a negative control, we observed that scrambled siRNAs resulted in maximal AP-1 transcription ("Scramble"). As positive controls, siRNAs targeted directly to luciferase showed a large reduction of luminescence ("Luciferase"), and siRNAs directed directly to p38, upstream of AP-1, reduced signal intensity by 10-50%. Error bars represent standard errors over six replicate assays.

Kinase subnetworks identify NHE1 as a novel kinase scaffold

We next analyzed the network for evidence of novel MAPK scaffold proteins, which form a signaling apparatus through the simultaneous binding of kinases and their substrates[14]. To identify such scaffolds, we selected proteins that interact with multiple MAPK levels and have at least 40% of their interactions with kinases, yielding a total of 10 candidate scaffolds (Figure 2.2A-D, Supplementary Table S2). As a positive control, this strategy detected a well-established human MAPK scaffold, the actin-binding protein FLNA, which has been postulated to organize kinase signaling between the membrane and cytoskeleton and to regulate transcription factors such as AP-1[15]. Of the 11 interactions involving FLNA, seven are with kinases or transcription factors (Figure 2.2A).

We found that the plasma membrane Na-H exchanger NHE1 interacted with a total of seven MAPK family proteins, spanning all four levels of the MAPK hierarchy including MAP4K4, two MAP3Ks (MAP3K7 and RAF1), MAP2K2, and three MAPKs (ERK1, ERK2, and JNK3) (Figure 2.2B). NHE1 also interacted with the Rho GTPase Rac1, which can NHE1 activity[16]. In addition to its known role in ion exchange, we postulated that NHE1 may function as a plasma-membrane scaffold for the assembly of signaling complexes[17]. Using tagged NHE1 we confirmed the interaction between the C-terminal cytoplasmic domain of NHE1 and MAP3K7 (TAK1) in *vitro* (Figure 2.2E). We also identified two independent siRNAs targeted to NHE1 that were able to significantly reduce phosphorylation levels of p38 in response to phorbol-12-myristate-13-acetate (PMA), an established assay of MAPK pathway function (Figure 2.2F).

Thus, based on the new protein interactions and functional siRNA screening data, it is likely that the cytoplasmic tail of NHE1 promotes the assembly of a GTPase and MAP-family kinases into (possibly multiple) signaling complexes. Because we previously showed that NHE1 binds to and is phosphorylated by MAP4K4 in response to external growth factors[18], a similar mechanism might govern the activity of the other interactions identified here. Since MAP3K7 is activated by the transforming growth factor β (TGF-β) receptor[19] and we have previously identified an NHE1-immune complex containing the type II TGF-β receptor, the new interaction data suggests that NHE1 can act as a regulatory molecule for processing various cell stimuli in union with the TGF-β receptor, MAP3K7 and other MAPKs[20].

A number of novel interactions involving RANBP9 were also detected (Figure 2.2C). Of particular interest is an interaction between RANBP9 and the RAPGEF2 guanine exchange factor, which is a member of the Ras subfamily of GTPases. RANBP9 was originally characterized as binding to the Ras GTPase-binding protein RAN and has been shown to activate the Ras signaling pathway[21]. Based on the Y2H evidence, we postulate that RANBP9 functions as a scaffold for RAPGEF2 and various MAPK kinases, including ERK3, and promotes activation of the transcription factors MAX, MEF2C, and JUN (a component of the AP-1 transcription factor). We found that siRNAs directed to RANBP9 significantly reduced AP-1 transcription in response to PMA (Figure 2.2G), confirming that RANBP9 influences MAPK signaling upstream of AP-1.

Identification of a core set of high-confidence interactions and modules

MAPK pathways are well-conserved among eukaryotes[2]. As a means of further validating and enhancing the identified set of MAPK interactions, we searched for overlap between this protein network and that of budding yeast. A total of 140 MAPK Y2H interactions (~6%) were found to be conserved among close homologs in yeast (see Methods). Using the same reference set of literature-curated interactions discussed earlier, we determined that the set of conserved Y2H interactions was approximately 1.6-fold more precise than the entire MAPK Y2H network, at the expense of an approximate one-fourth reduction in coverage (Figure 2.1B; the reduction was expected since the conserved network is a subset of the entire network). We also observed that precision increased nearly 1.5-fold for interactions that were observed more than once by Y2H screening (Figure 2.1B, Supplementary Figure S2). Thus, we combined the 137 conserved Y2H interactions with the 551 multiply-sampled interactions to form a "core MAPK network" of 641 unique high-confidence interactions (including 47 interactions selected by both criteria) (Figure 2.3A, Supplementary Table S3).

To shed light on the structure and function of this core MAPK network, we organized the network into conserved and/or species-specific modules (Methods). We found that the conserved modules formed six connected components, highlighting potential conserved mechanisms of signaling and regulation (Figure 2.3B-G).

One of these modules shows interactions of the MUC12 protein with p38 and Cdc42 (Figure 2.3B), suggesting that MUC12 might function in a CDC42-responsive signaling cascade. This hypothesis is supported through conservation with yeast, in which Cdc42 is found to interact with mucin family member Msb2 whose function in kinase signaling has already been established[22].

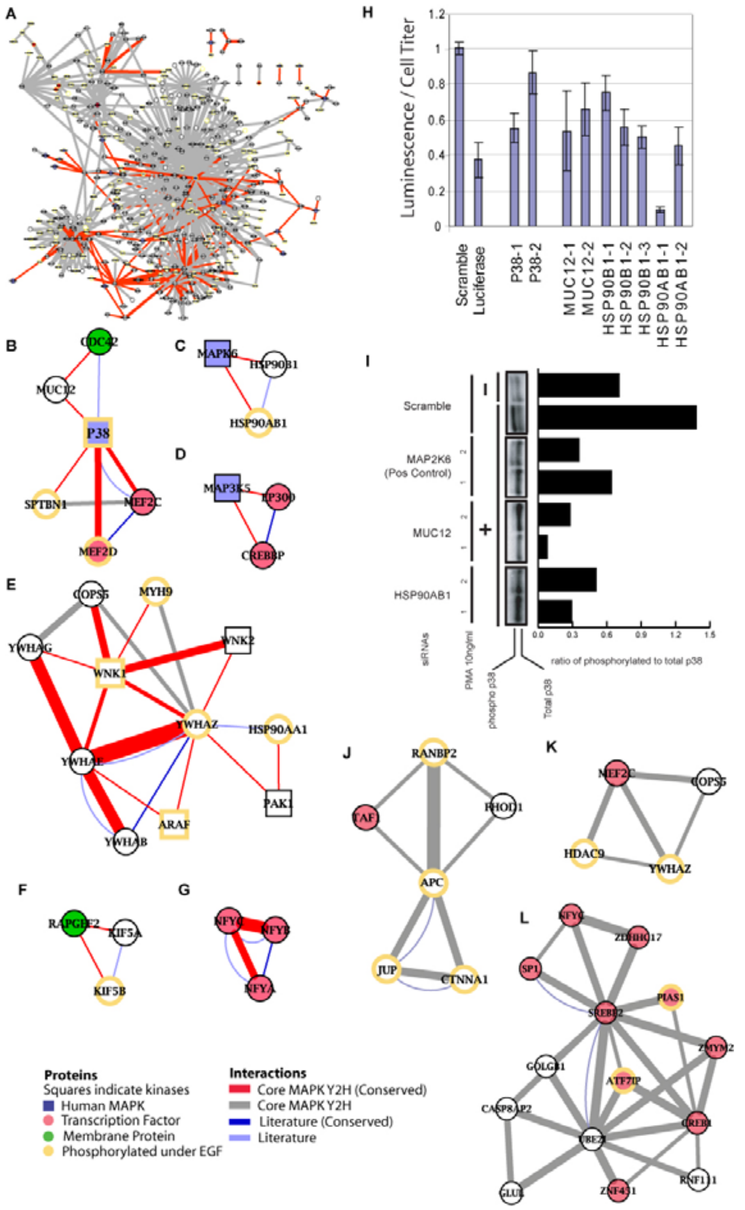


Figure 2.3: Functional Modules in the Core Network.

(A) Bird's eye view of the core MAPK Y2H network. (B-G) High confidence conserved functional modules. Red edges correspond to core MAPK Y2H interactions which were conserved with yeast. Grey edges indicate core interactions not conserved with yeast. Thickness of the edge increases with the number of observations. (H) AP-1 luciferase activation assay for various siRNAs targeting members of conserved modules. (I) p38 phosphorylation levels are decreased with siRNAs targeting members of conserved modules. (J-L) Novel modules not conserved with yeast.

To elucidate the role of MUC12 in human MAPK signaling we found that multiple distinct siRNA constructs against MUC12 significantly reduce both AP-1 activation (Figure 2.3H) and p38 phosphorylation (Figure 2.3I) in response to PMA. Together, the Y2H interaction and siRNA results suggest the existence of a novel human signaling mucin acting in an analogous fashion to the yeast mucin Msb2.

Another conserved module suggests a role for HSP90 members HSP90B1 and HSP90AB1 in the function of MAPK6 (ERK3, Figure 2.3C), perhaps to stabilize it in much the same way as HSP90B1 has been shown to stabilize ERBB2[23]. These interactions are plausible since HSP90 proteins are known to associate with kinases that mediate multiple inputs[24]. To further investigate this hypothesis, we screened multiple siRNA constructs directed to HSP90AB1 and HSP90B1. We observed a reduction in levels of AP-1 transcription (Figure 2.3H) for HSP90AB1 and HSP90B1 as well as phosphorylated p38 (Figure 2.3I) for HSP90AB1, suggesting a critical role for these HSP90 members in MAPK-mediated signaling.

We also identified three distinct network modules (Figure 2.3J-L) which were not conserved with yeast and either indicate missing interactions in yeast or machinery that may be present only in higher eukayotes. The modules highlight evidence for an interaction between RANBP2, the GTP-binding protein at the nuclear pore, and the APC tumor suppressor gene which is known to promote the association of the nuclear pore with microtubules[25]. Furthermore, they reveal a novel high-confidence interaction between APC and the catenin CTNNA1 (Figure 2.3J). Since interaction of APC with other β-catenin pathway members has been shown to be critical for WNT signaling, interactions of CTNNA1 with APC may play a role in this pathway as well. Because

WNT signaling is conserved among metazoans but not yeast, these examples illustrate the power of network mapping on a species-specific basis.

Conclusions

This study demonstrates a combined experimental and computational approach for the systematic expansion and refinement of MAPK signaling based on high-throughput interaction mapping seeded by a core set of known protein components. We have reported the discovery of over 2,000 protein interactions related to MAPK proteins. The quality of this network is comparable to previously-published Y2H datasets and can be substantially increased by cross-species comparative methods. Analysis of core interactions highlight kinase scaffolds and diverse signaling components. Other large-scale technologies such as gene expression profiling, co-affinity purification, and phospho-proteomics may add complementary facets of information to these data, yielding a more complete understanding of MAPK signaling in humans and yeast. Future studies may focus on the activity of the reported kinase interaction pathways in terms of their condition-specificity or crosstalk via shared components.

Materials and Methods

Y2H Screening

Yeast strains and expression vectors for Y2H screening are as previously described[26]. Human protein baits were screened against preys derived from 22 different human cDNA libraries. The original 86 baits were selected based on BioCarta (http://www.biocarta.com/pathfiles/m_p38mapkPathway.asp) as well as[27]. Full experimental and computational details are provided in the Supplementary Methods.

Quality Assessment of Interactions

A reference set of literature-curated interactions was assembled from 23,975 human protein interactions deposited in public databases including known interactions involving MAPKs (Supplementary Table S4). For comparison, a random set of interactions was formed by selecting 100 times as many protein pairs at random (2,397,500 pairs), excluding interactions that were already in the reference. Networks were evaluated with regard to their relative precision and coverage: Precision was defined as the percent of measured interactions confirmed by reference positives (True Positives / [True Positives + False Positives]), and coverage was defined as the percent of all reference positives recovered (True Positives / [True Positives + False Negatives]) by the measured interactions. Note that these figures are relative, not absolute, since the reference set may contain some proportion of false interactions, and the random set may contain a few real protein interactions that have yet to be discovered.

AP-1 luciferase and p38 siRNA assay

The HEK293 cell line stably expresses an AP-1 responsive luciferase reporter as previously described[7]. An aliquot of $8x10^3$ HEK293 cells was plated into 96-well tissue culture plates and each well transfected with 25 ng of indicated siRNA by using Lipofectamine2000 reagent (Invitrogen). After 72 h of transfection, cells were stimulated with 10ng/ml of PMA for 8 hours, and luciferase activity was measured by using Bright Glow (Promega) according to the manufacturer's instructions. Cell titer was measured by the CellTiter-Glo Luminescent Cell Viability Assay (Promega). Both cell titer and luciferase activity counts were normalized by the median of the scramble siRNAs[28]. All siRNA sequences are given in Supplementary Table S7.

Western blotting for phospho-p38 and in-vitro binding assays

HeLa cells were transfected using RNAiMAX according to the manufactured protocol (Invitrogen Life Technologies). After 72 h of transfection, cells were stimulated with 10ng/ml of PMA for 8 hours. Cells were lysed in buffer containing 25 mM Tris-HCl pH 7.6, 150 mM NaCl, 1% NP-40, 1% Sodium deoxycholate, 0.1 % SDS, and phosphatase inhibitors (Sigma Aldrich). A 10 ng amount of cell lysate was resolved by SDS-PAGE, and blots were immunoblotted with the antibodies detecting the phosphorylated form of p38 (Promega). All blots were developed with HRP-conjugated secondary antibodies and ECL (Amersham Biosciences). In-vitro binding assays for N-terminal tagged NHE1 were performed as described previously[4].

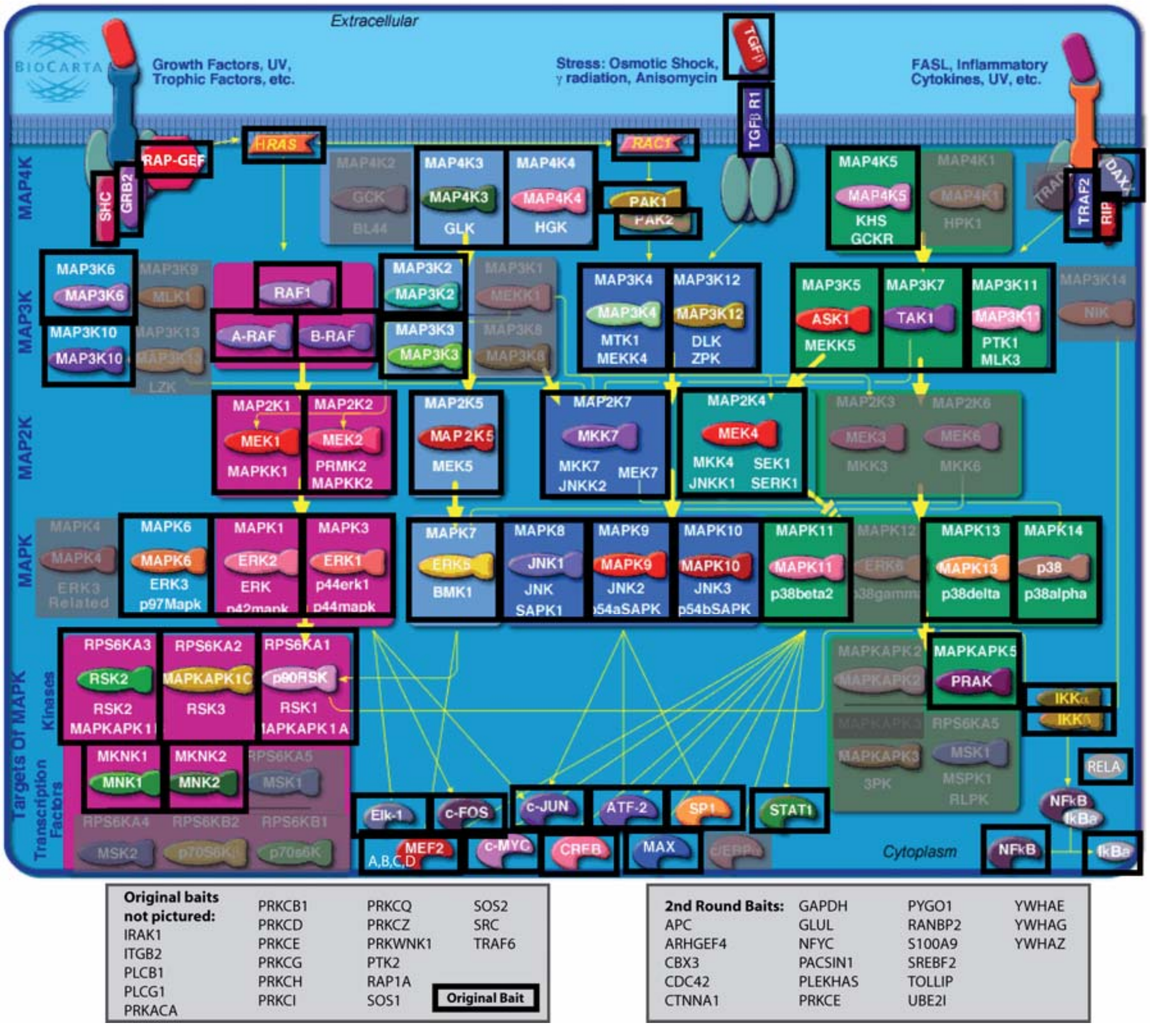
Identification of conserved interactions

A yeast MAPK network was assembled from both literature-curated and experimental sources (listed in Supplementary Methods). A human protein interaction was considered "conserved" if both proteins had homologs that interacted in yeast. Human/yeast homologs were defined using a strict BLAST *E*-value < 10^{-10}, and to avoid spurious matches due to large gene families we allowed no more than 10 yeast homologs for each human protein (10 best *E*-values).

Module finding

We combined core interactions with a reference set of human protein interactions (see 'Quality Assessment of Interactions' above as well as Supplementary Table S4). We identified network modules as all complete interacting triplets of proteins (i.e., triangles) for which at least two interactions were from the core MAPK network. We hypothesized that dense network clusters such as triangles might signify functional modules whose interactions are more reliable owing to their inter-connectivity. In total, 134 triplets were found covering 195 core MAPK Y2H interactions (Supplementary Table S5). Conserved modules were formed by combining triangles for which at least two of the interactions were conserved with yeast. The quality of these interactions were high, over 1.8 fold more precise than the MAPK Y2H network using the reference set described earlier (Figure 2.1B).

Supplemental Figures and Methods

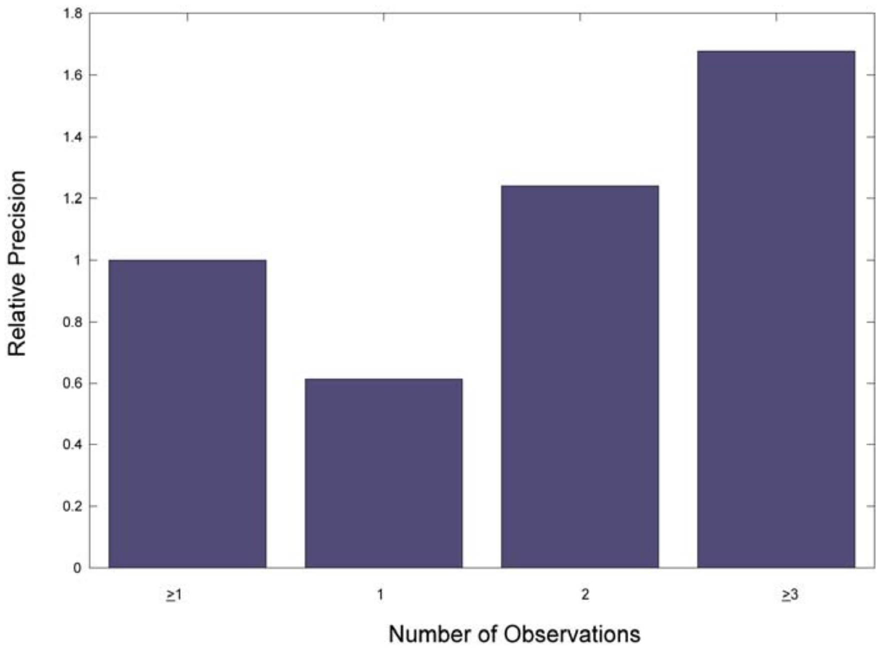


Supplemental Figure 2.1: Simplified signaling diagram of baits used in the MAPK Y2H network.

The 86 original baits and the 20 additional baits are shown. Image adapted from BioCarta "MAPKinase signaling pathway" (http://www.biocarta.com/pathfiles/h_mapkPathway.asp)

Highly sampled interactions are of higher quality

Based on the network quality assessment framework, we investigated the effect of biological sampling on the quality of the derived interaction network. We found a consistent increase in precision (percent of interactions confirmed by reference positives versus negatives) with the number of observations of an interaction within the Y2H screening process. The performance is the best at 3 or more observations with a relative precision nearly three-fold higher than those interactions sampled only once (fig S2). Although the quality of highly sampled interactions is high, their occurrence is low, with only 551 interactions occurring more than once (table S2.1).

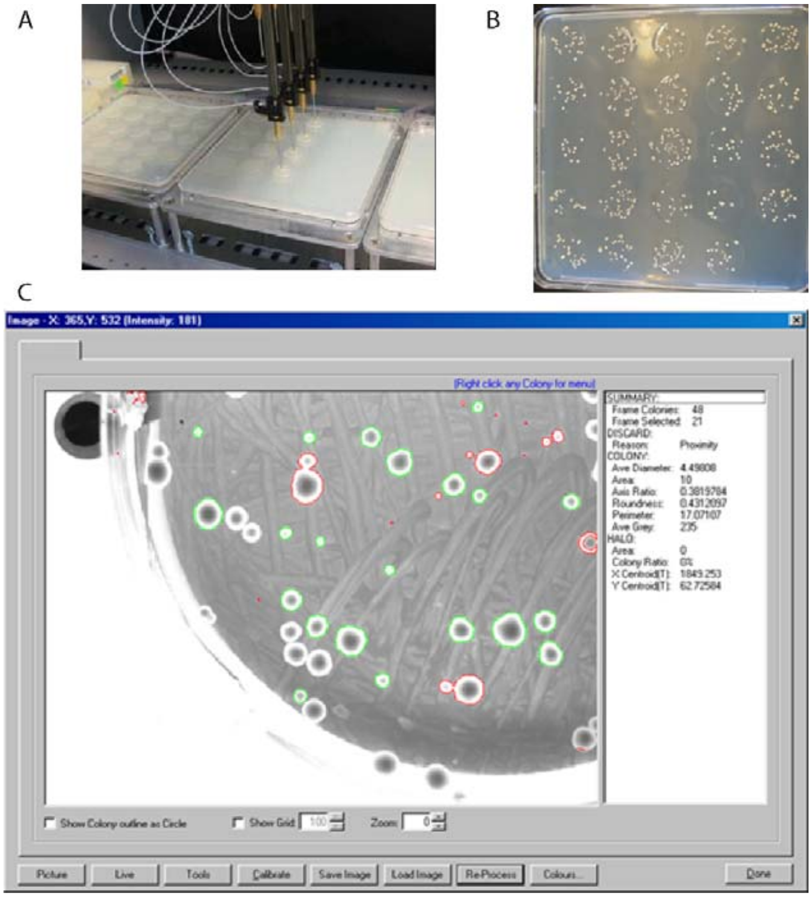


Supplemental Figure 2.2: Yeast two-hybrid interactions observed multiple times are of enhanced quality.

Precision is shown relative to the set of all measured interactions (≥1 on the x-axis).

Y2H network generation

Both bait and prey were cloned as double fusions in plasmids pOBD.111 and pOAD.102 and the Y2H procedure is as described in LaCount et al. [26]. The cDNA was cloned between the 2-hybrid domain on the 5' end of the insert and an in-frame selection marker on the 3' end of the insert. Bait cDNA were cloned between the GAL4 binding domain and the TRP1 coding region and prey between the GAL4 transcriptional activation domain and URA3. Both bait and prey cDNA libraries were prepared by random primed cDNA synthesis from polyA-selected RNA isolated from the human tissues outlined Table S6 followed by the PCR addition of yeast recombination tails. These cDNAs were then cloned into linearized expression vectors by recombination in yeast. Transformed bait yeast were plated on medium lacking tryptophan to select for in-frame TRP1 fusions and prey were selected without uracil for in-frame URA3 fusions. Y2H screens were performed in 96-well plates by mating in each well $5x10^6$ cells of a yeast clone expressing a single bait with $5x10^6$ clonally diverse cells from a prey library. After mating overnight the well contents were plated on medium that selected simultaneously for successful mating, the expression of the ORF-selection markers, and the activity of the metabolic reporter genes, ADE2 and HIS3 (fig S3A,B). Two-hybrid-positive diploids were counted and up to 48 colonies per mating were picked and transferred to liquid medium (fig. S3C). Searches that yielded more than 200 positives were considered to be self activators i.e. resulting from bait plasmids that activated transcription in the absence of specific protein-protein interactions, and were not analyzed further. The liquid cultures were then used as template for separate PCR reactions to amplify insert sequence from bait and prey plasmids for subsequent sequence determination. Sequence information was processed to prepare the protein-protein interaction data set. After vector and adaptor clipping, read assembly, repeat masking and contamination filtering, sequences were BLASTed against RefSeq and the top hit used for identification and Entrez Gene mapping.


A
B
C
Image - X: 365,Y: 532 (Intensity: 181)
(Right click any Colony for menu)
□ Show Colony outline as Circle
□ Show Grid 100
Zoom: 0
Picture
Live
Tools
Calibrate
Save Image
Load Image
Re-Process
Colours...
Done

SUMMARY:
Frame Colonies: 48
Frame Selected: 21
DISCARD:
Reason: Proximity
COLONY:
Ave Diameter: 4.49808
Area: 10
Axis Ratio: 0.3819784
Roundness: 0.4312097
Perimeter: 17.07107
Ave Grey: 235
HALO:
Area: 0
Colony Ratio: 0%
X Centroid[T]: 1849.253
Y Centroid[T]: 62.72584

Supplemental Figure 2.3: High-throughput Y2H screening process.

(A) Plating of mated Y2H clones. (B) View of plated colonies (C) image analysis of Y2H positive colonies for automated picking.

Supplemental Table 2.1: List of cDNA sources used in this study.

	Human Tissue
1	Adipose
2	Brain
3	Brain, caudate nucleus
4	Brain, cerebellum
5	Brain, hippocampus
6	Colon
7	Colorectal adenocarcinoma
8	Fetal Brain
9	Fetal lung
10	Kidney
11	Leukocyte
12	Liver
13	Lung
14	Lung carcinoma
15	Mammary Gland
16	Melanoma
17	Pancreas
18	Placenta
19	Prostate Gland
20	Spinal Cord
21	Spleen
22	Testis

Incorporation of other networks and identification of network modules

Human protein interactions were obtained from BIND[29] and HPRD[30] and curated interactions including those in the BioCarta Database[31] (table S2, November 2006 download). To identify conserved human interactions, yeast literature curated protein-protein interactions were taken from Reguly et al.[32] and combined with Ptacek et al.[33], Gavin et al.[34], Ito et al.[35], Ho et al.[36] and Uetz et al.[37].

Network quality assessment using a reference set

Note that we focus on relative comparisons of precision and coverage between data sets; the absolute values of these measures are less reliable because of their dependence on the particular choice of reference data set and its assumptions. For instance, many studies implicitly assume that the contents of the literature-curated databases are 100% true, that negatives are orders of magnitude in excess of the positives, and that negatives are adequately modeled by random protein pairs[38,39]. The absolute coverage could be higher if some interactions recorded in the public databases were false, or conversely, the absolute precision could be lower if the proportion of positives was less than 1 in 100. For these reasons, we report relative quality measures in the main manuscript.

Acknowledgments

We thank R. Kelley, S. Suthram and G. Warner for assistance on various aspects of this work. This work was generously supported by funds from the National Institutes of Health (R01-GM070743, R01-GM47413 and P30-MH062261) and Unilever PLC.

Chapter 2, in full, is the following manuscript currently under submission,

Bandyopadhyay S, Chiang C, Srivastava J, Gersten M, White S, Bell R, Kurschner C, Martin CH, Smoot M, Sahasrabudhe S, Barber DL, Chanda SK, Ideker T. A *Core Network of Human MAPK Interactions.* Submitted.

The dissertation author is the sole first author on this work, responsible for study design and data analysis.

Bibliography

1. Chang L, Karin M. Mammalian MAP kinase signalling cascades. *Nature* 2001;410:37-40.

2. Widmann C, Gibson S, Jarpe MB, Johnson GL. Mitogen-activated protein kinase: conservation of a three-kinase module from yeast to human. *Physiol Rev* 1999;79:143-80.

3. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M. Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell* 2006;127:635-648.

4. Denker SP, Huang DC, Orlowski J, Furthmayr H, Barber DL. Direct binding of the Na--H exchanger NHE1 to ERM proteins regulates the cortical cytoskeleton and cell shape independently of H(+) translocation. *Mol Cell* 2000;6:1425-36.

5. Kolch W, Calder M, Gilbert D. When kinases meet mathematics: the systems biology of MAPK signalling. *FEBS Lett* 2005;579:1891-5.

6. Johnson GL, Lapadat R. Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science* 2002;298:1911-2.

7. Chanda SK, White S, Orth AP, Reisdorph R, Miraglia L, Thomas RS, DeJesus P, Mason DE, Huang Q, Vega R, Yu DH, Nelson CG, Smith BM, Terry R, Linford AS, Yu Y, Chirn GW, Song C, Labow MA, Cohen D, King FJ, Peters EC, Schultz PG, Vogt PK, Hogenesch JB, Caldwell JS. Genome-scale functional profiling of the mammalian AP-1 signaling pathway. *Proc Natl Acad Sci U S A* 2003;100:12153-8.

8. Kolch W. Coordinating ERK/MAPK signalling through scaffolds and inhibitors. *Nat Rev Mol Cell Biol* 2005;6:827-37.

9. Johnson SA, Hunter T. Kinomics: methods for deciphering the kinome. *Nat Methods* 2005;2:17-25.

10. Reszka AA, Seger R, Diltz CD, Krebs EG, Fischer EH. Association of mitogen-activated protein kinase with the microtubule cytoskeleton. *Proc Natl Acad Sci U S A* 1995;92:8881-5.

11. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;437:1173-8.

12. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005;122:957-68.

13. Friedman A, Perrimon N. A functional RNAi screen for regulators of receptor tyrosine kinase and ERK signalling. *Nature* 2006;444:230-4.

14. Whitmarsh AJ, Davis RJ. Structural organization of MAP-kinase signaling modules by scaffold proteins in yeast and mammals. *Trends Biochem Sci* 1998;23:481-5.

15. Hayashi K, Altman A. Filamin A is required for T cell activation mediated by protein kinase C-theta. *J Immunol* 2006;177:1721-8.

16. Hooley R, Yu CY, Symons M, Barber DL. G alpha 13 stimulates Na+-H+ exchange through distinct Cdc42-dependent and RhoA-dependent pathways. *J Biol Chem* 1996;271:6152-8.

17. Baumgartner M, Patel H, Barber DL. Na(+)/H(+) exchanger NHE1 as plasma membrane scaffold in the assembly of signaling complexes. *Am J Physiol Cell Physiol* 2004;287:C844-50.

18. Yan W, Nehrke K, Choi J, Barber DL. The Nck-interacting kinase (NIK) phosphorylates the Na+-H+ exchanger NHE1 and regulates NHE1 activation by platelet-derived growth factor. *J Biol Chem* 2001;276:31349-56.

19. Yamaguchi K, Shirakabe K, Shibuya H, Irie K, Oishi I, Ueno N, Taniguchi T, Nishida E, Matsumoto K. Identification of a member of the MAPKKK family as a potential mediator of TGF-beta signal transduction. *Science* 1995;270:2008-11.

20. Karydis A, Jimenez-Vidal M, Denker SP, Barber DL. Mislocalized scaffolding by the Na-H exchanger NHE1 dominantly inhibits fibronectin production and TGF-beta activation. *Mol Biol Cell* 2009;20:2327-36.

21. Wang D, Li Z, Messing EM, Wu G. Activation of Ras/Erk pathway by a novel MET-interacting protein RanBPM. *J Biol Chem* 2002;277:36216-22.

22. Cullen PJ, Sabbagh W, Jr., Graham E, Irick MM, van Olden EK, Neal C, Delrow J, Bardwell L, Sprague GF, Jr. A signaling mucin at the head of the Cdc42- and MAPK-dependent filamentous growth pathway in yeast. *Genes Dev* 2004;18:1695-708.

23. Xu W, Mimnaugh EG, Kim JS, Trepel JB, Neckers LM. Hsp90, not Grp94, regulates the intracellular trafficking and stability of nascent ErbB2. *Cell Stress Chaperones* 2002;7:91-6.

24. Citri A, Harari D, Shohat G, Ramakrishnan P, Gan J, Lavi S, Eisenstein M, Kimchi A, Wallach D, Pietrokovski S, Yarden Y. Hsp90 recognizes a common surface on client kinases. *J Biol Chem* 2006;281:14361-9.

25. Collin L, Schlessinger K, Hall A. APC nuclear membrane association and microtubule polarity. *Biol Cell* 2008;100:243-52.

26. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, Fields S, Hughes RE. A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature* 2005;438:103-7.

27. Qi M, Elion EA. MAP kinase pathways. *J Cell Sci* 2005;118:3569-72.

28. Konig R, Chiang CY, Tu BP, Yan SF, DeJesus PD, Romero A, Bergauer T, Orth A, Krueger U, Zhou Y, Chanda SK. A probability-based approach for the analysis of large-scale RNAi screens. *Nat Methods* 2007;4:847-9.

29. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW. BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res* 2001;29:242-5.

30. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker

K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003;13:2363-71.

31. BioCarta. p38 MAPK Signaling Pathway. http://www.biocarta.com/pathfiles/m_p38mapkPathway.asp.

32. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskya OG, Ideker T, Dolinski K, Batada NN, Tyers M. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol* 2006;5:11.

33. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJ, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M. Global analysis of protein phosphorylation in yeast. *Nature* 2005;438:679-84.

34. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;415:141-7.

35. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 2001;98:4569-74.

36. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002;415:180-3.

37. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 2000;403:623-7.

38. Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. *Science* 2004;306:1555-8.

39. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002;417:399-403.

Chapter 3. Systematic identification of functional orthologs based on protein network comparison

Abstract

Annotating protein function across species is an important task which is often complicated by the presence of large paralogous gene families. Here, we report a novel strategy for identifying functionally related proteins that supplements sequence-based comparisons with information on conserved protein-protein interactions. First, the protein interaction networks of two species are aligned by assigning proteins to sequence homology groups using the Inparanoid algorithm. Next, probabilistic inference is performed on the aligned networks to identify pairs of proteins, one from each species, that are likely to retain the same function based on conservation of their interacting partners. Applying this method to *D. melanogaster* and *S. cerevisiae*, we analyze 121 cases for which functional orthology assignment is ambiguous when using sequence similarity alone. In 61 of these cases, the network supports a different protein pair than that favored by sequence comparisons. These results suggest that network analysis can be used to provide a key source of information for refining sequence-based homology searches.

Introduction

The idea that similar protein sequences imply similar protein functions has long been a central concept in molecular biology. With each new completed genome, an increasingly complex array of sequence alignment and comparative modeling tools are used to annotate functions for the typically thousands of encoded proteins, based largely on similarity to proteins that are well characterized in other species[2,3]. Ambiguities in the functional annotation process arise when the protein in question has similarity to not one but many paralogous proteins[4], making it harder to distinguish which of these is the true ortholog—that is, the protein that is directly inherited from a common ancestor. Especially in the genomes of mammals and other higher eukaryotes, large protein families are typically not the exception but the rule.

The difficulty of assigning protein orthology depends largely on the evolutionary history. Protein families for which speciation predates gene duplication are particularly challenging; in these cases, every cross-species protein pair is technically orthologous but it is still necessary to distinguish which protein pairs play functionally equivalent roles, i.e. are *functional orthologs*[5]. Conversely, when gene duplication predates speciation, the family can often be subdivided into orthologous pairs which have higher sequence similarity to each other than to other members. However, evolutionary processes such as gene conversion serve to homogenize paralogous sequences over time, making these cases problematic as well[6]. To complicate matters even further, protein function may be lost between distant organisms or conserved across multiple proteins within a single species.

A variety of sequence-based approaches have been proposed to address these challenges. The COGs approach (Clusters of Orthologous Groups)[7] defines orthologs using sets of proteins that contain reciprocal best BLAST matches [8] across a minimum of three species. Also available are phylogenetic methods that explicitly address an evolutionary tree, as reviewed in[9]. Recent approaches such as Inparanoid[5] and OrthoMCL[6] try to achieve higher sensitivity through sequence clustering techniques which consider a range of BLAST scores beyond the absolute best hits. For Inparanoid, BLAST E-values from the proteins of two species are clustered according to a fixed set of rules which divide proteins into ortholog groups, each of which contains similar-sequence proteins drawn from both species. Within each group, pairs of proteins (cross-species only) can be assigned an overall score reflecting the likelihood they are functional orthologs.

Other than gene and protein sequences, several large-scale data types have recently become available that provide complementary information on functional conservation. For instance, several groups have used correlated patterns of gene expression across species as evidence for functional relatedness[10,11]. Networks of protein-protein interactions are also being generated for a variety of species, through technologies such as the two-hybrid assay[12] or co-immunoprecipitation followed by mass spectrometry[13]. Such networks can be compared to identify "interologs", i.e., interactions that are conserved across species[14]. Beyond comparison of interactions individually, methods such as PathBLAST[15,16] and that of[17] create a global alignment between networks to identify conserved network regions. These approaches can successfully infer conserved components of the cellular machinery and use those components to predict new protein functions and interactions. In addition, interactions that are conserved across species are less likely to represent false positives.

Here, we investigate whether it is possible to use protein network information to predict functionally orthologous proteins across species. While previous tools such as Interolog mapping and PathBLAST have used orthology to identify conserved protein interactions, our approach aims to reverse this logic and use conserved protein interactions to predict functional orthology. It is built on the concept that a protein and its functional ortholog are likely to interact with proteins in their respective networks that are themselves functional orthologs. This type of network-based approach is related to methods for predicting other protein properties based on the interaction network, such as functional annotation of a protein based on the annotations of its neighbors[18-21]. In our case, the orthology relation between each pair of proteins is modeled as a probabilistic function of the orthology relations of their immediate network neighbors, and orthology relationships are inferred using Gibbs sampling. We apply this approach to refine the set of functional orthologs between the budding yeast *Saccharomyces cerevisiae* and the fruit fly *Drosophila melanogaster*: not only are these species among the most important model eukaryotes, they are also associated with the largest numbers of experimentally-measured protein interactions to date.

Results

Motivation: Interaction conservation is related to orthology.

Protein-protein interaction networks for yeast and fly were obtained from the Database of Interacting Proteins[22] (December 2004 download). These contained 14,319 interactions among 4,389 proteins in yeast and 20,720 interactions among 7,038 proteins in fly. First, we applied the Inparanoid[5] algorithm to the complete sets of proteins from *S. cerevisiae* and *D. melanogaster* to define sequence-similar clusters. A total of 2,244 clusters were generated, covering 2,834 yeast and 3,881 fly proteins overall. Of these, 1,552 clusters contained only a single yeast/fly protein pair and were assumed to represent unambiguous or "definite" functional orthologs (orthologs we take to be functionally equivalent because of direct ancestry). The remaining 692 clusters contained multiple proteins from yeast and/or fly, leaving the functional orthologs ambiguous.

To determine the extent to which proteins and their functional orthologs had conserved protein interactions, we examined the network neighborhoods of definite functional orthologs and compared them to the neighborhoods of less related protein pairs (Figure 3.1). As a measure of local network conservation, we computed the *conservation index* of each protein pair as proportional to the fraction of interactions that were conserved across the two species. For example, in Fig. 3.2b the orthologous pairing *B/B′* has a higher conservation index (4/9) than the alternative pairing *B/B″* (2/9).

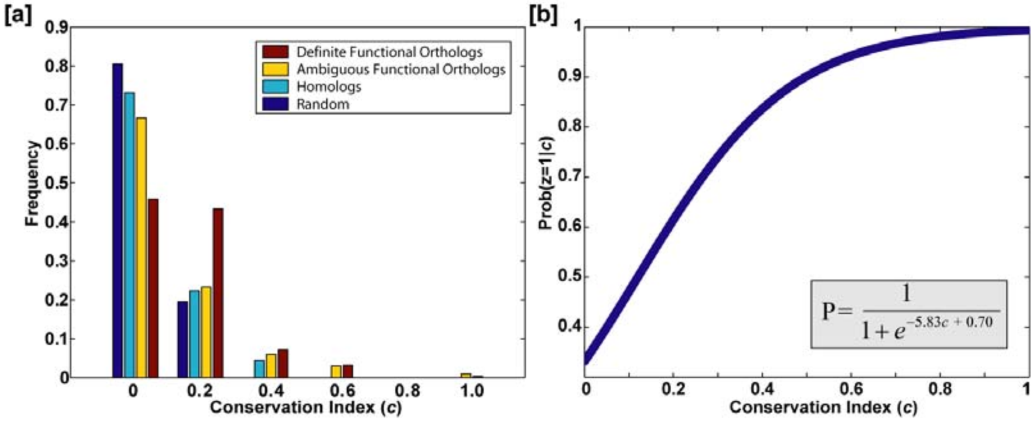


Figure 3.1: Network neighborhood conservation for definite orthologs versus other yeast/fly protein pairs.

[a] The distribution of the conservation index (c) is shown for definite functional orthologs (sole members of an Inparanoid group); ambiguous functional orthologs (in a group with multiple members); homologs (different groups but similar sequences); and random protein pairs. Definite functional orthologs show a shift towards higher conservation of protein interactions between the yeast and fly protein networks. Mean c=0.1512, 0.1171, 0.0870, 0.0615 for definite functional orthologs, ambiguous functional orthologs, homologs, and random pairs, respectively. [b] Logistic function relating conservation index to probability of functional orthology. Logistic regression was performed using the "definite functional ortholog" and "homolog" pairs as positive vs. negative training data, respectively. The resulting function is shown.

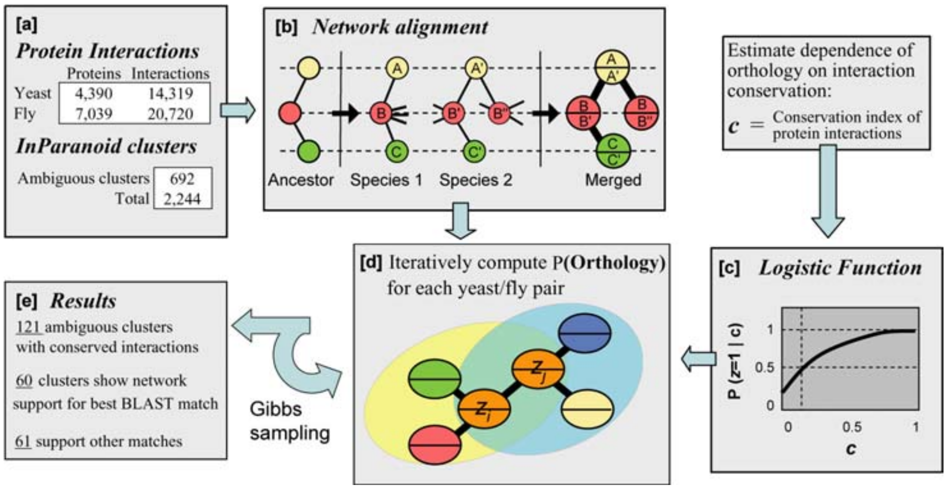


Figure 3.2: Overview of the method.

[a] Protein-protein interaction networks for yeast and fly are combined with clusters of orthologous yeast and fly protein sequences as determined by the Inparanoid algorithm. [b] Networks are aligned into a merged graph representation. In this example, a gene duplication results in two proteins B′ and B″ in species 2 that are orthologous to protein B in species 1. One of these proteins may experience a gain and/or loss of interactions to enable new functional roles 1; however, only conserved interactions are represented in the alignment graph. [c] The logistic function shown in Fig. 1b is used to compute the probability of functional orthology for a protein pair i given the states of functional orthology for its network neighbors. [d] This probability is updated for each pair over successive iterations of Gibbs sampling. [e] The final probabilities confirm 60 of the best BLAST match pairings. The network supports a different hypothesis for 61 pairings.

Fig. 3.1a shows the set of conservation indices for definite functional orthologs versus those of ambiguous functional orthologs, non-orthologous homologs (best cross-species BLAST matches not assigned to the same Inparanoid group), and random pairs of proteins chosen independent of sequence similarity. As expected, the set of definite functional orthologs had the highest occurrence of conserved interactions. Moreover, the mean conservation index was related to the stringency of the pairing: definite functional orthologs tended to have higher conservation indices than ambiguous functional orthologs, ambiguous functional orthologs higher indices than homologs, and homologs higher indices than random protein pairs. Beyond the mean conservation index, there were also significant differences among the four distributions (Supplemental Table 1). These findings confirm that yeast/fly proteins classified as definite functional orthologs are more likely to have equivalent functional roles in the protein network and, conversely, that conserved network context could be used to help discriminate functional orthology from general sequence similarity.

Network-based identification of functional orthologs.

To capture these trends to identify functional orthologs, we formulated a procedure to estimate the likelihood of functional orthology for each ambiguous functional ortholog given its conservation index. By this method, the probability of functional orthology for a pair of proteins is influenced by the probabilities of functional orthology for their network neighbors, which in turn depend on their network neighbors, and so on. This type of probabilistic model is known as a Markov random field[23]. Exact inference in this model is not tractable because of the complex interrelationships between network nodes. Rather, approaches such as clustering, conditioning, and stochastic simulation have been used to derive estimates for the posterior probabilities of node properties. Here, we implemented a method based on Gibbs sampling for its computational tractability and accuracy in densely connected networks[24]. An overview of the approach is given in Figure 3.2, with full details provided in the Methods.

Application to yeast and fly identifies new putative orthologous pairs.

We applied this approach to resolve ambiguous functional orthology relationships in the yeast and fly protein networks. Of the 692 ambiguous Inparanoid clusters, 121 contained protein pairs for which at least one pair had conserved interactions between networks. Application of our Gibbs sampling procedure yielded estimates of probability of functional orthology for each protein pair in these 121 ambiguous clusters. In 60 of these clusters, the highest probability was assigned to the protein pair that was also the most sequence-similar via BLAST. These cases reinforced the intuition that the best sequence matches are also the most functionally similar. The remaining 61 clusters showed the opposite behavior, i.e., the highest probability pair was not the most sequence similar pair. Of these 61 cases, 15 were supported by two or more conserved interactions (Table 1). Because the yeast and fly networks are incomplete (i.e., they contain false negatives), in some of these cases we cannot rule out the possibility that conserved interactions with the best BLAST matches have been missed (see Discussion). A complete listing of the results can be found on the supplemental website (http://bioinf.ucsd.edu/~sbandyop/GR/).

Supplemental Table 3.1: Inparanoid clusters for which the network suggests different functional pairings than BLAST.

Inparanoid Cluster	Yeast / Fly Pairings in Cluster	Total Protein Interactions in Yeast / Fly	BLAST E-value	Number of Conserved Interactions	P(z)
35	Ssa3 / Hsc70-4	3/29	1E-277	0	--
	Ssa2 / Hsc70-4	10/29	7E-275	4	53.22%
	Ssa1 / Hsc70-4	13/29	2E-275	4	48.10%
94	Act1 / Act5c	38/48	9E-201	3	39.56%
	Act1 / Act42a	38/3	3E-200	1	39.24%
	Act1 / Act87e	38/11	1E-199	3	43.53%
	Act1 / CG10067	38/9	1E-198	2	38.20%
	Act1 / Act88f	38/2	9E-198	2	40.17%
126	Vph1 / CG7678	12/0	2.E-174	0	--
	Vph1 / CG18617	12/13	3.00E-170	2	41.87%
	Stv1 / CG18617	11/13	1.00E-148	1	38.44%
246	Kap104 / Trn	47/7	9E-128	2	40.64%
	Kap104 / CG8219	47/20	7E-96	5	46.78%
376	Pda1 / CG7024	8/1	9E-101	0	--
	Pda1 / L(1)g0334	8/13	6E-99	2	57.90%
425	Gpa1 / G-iα65a	14/2	1E-90	0	--
	Gpa1 / G-oα47a	14/12	5E-67	2	41.53%
707	Rpl12b / Rpl12	0/11	2.E-63	0	--
	Rpl12a / Rpl12	6/11	2.00E-63	2	48.39%
917	Cmd1 / Cam	61/19	1E-49	1	35.90%
	Cmd1 / And	61/26	4E-40	6	44.39%
1236	Fkh2 / CG11799	5/14	4E-31	0	--
	Fkh1 / CG11799	29/14	3E-18	2	42.34%
1550	Kel2 / CG12081	3/16	3E-19	0	--
	Kel1 / CG12081	16/16	1E-17	2	45.41%
1562	Egd1 / Bcd	3/16	2E-19	1	47.19%
	Btt1 / Bcd	3/16	2E-13	1	40.86%
	Btt1 / CG11835	3/2	2E-09	2	70.50%
1643	Ngr1 / CG12478	1/1	6E-16	0	--
	Nam8 / Aret	22/10	7E-06	2	45.06%
1687	Tpm2 / Tm1	1/7	3E-15	0	--
	Tpm1 / Tm2	3/17	2E-14	2	43.98%
1740	Mig2 / Opa	0/31	5.E-13	0	--
	Mig3 / Opa	2/31	1.00E-09	2	40.42%
2037	Gid8 / CG18467	3/0	8.E-03	0	--
	Gid8 / CG6617	3/8	0.001	2	76.51%

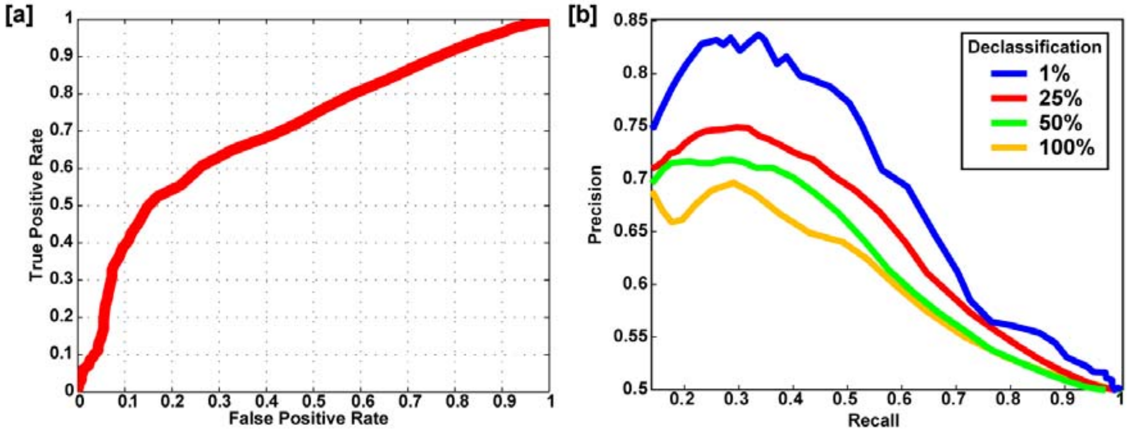


Figure 3.3: Estimated accuracy of the method.

[a] The Receiver Operating Characteristic (ROC) curve shows the true positive rate (percent of true data predicted correctly as positive) vs. the false positive rate (percent of false data predicted incorrectly, i.e. positive) of the method. [b] Dependence of predictions on number of available training examples. Percent recall (true positive rate) vs. precision (percent of positive predictions that were correct) is plotted as the probability cutoff ranges from [0-1]. Different color plots correspond to different percents of declassification of training examples.

Validation.

A straightforward validation of the approach would be to analyze its accuracy in recapitulating a “gold standard” set of protein functional annotations. However, databases of functional annotations are based directly on sequence similarity, and they typically lack the specificity to discriminate among subtle functional differences across large gene families. As an alternative approach, we used the technique of cross validation to test the ability of the approach to reclassify protein pairs in the definite functional ortholog set (positive test data) versus the non-orthologous homolog set (negative test data). In each cross-validation trial, 1% of these assignments were hidden (declassified) and monitored during Gibbs sampling to obtain probabilities of functional orthology for positive and negative examples. Reclassification was judged successful if the probability of functional orthology exceeded a particular cutoff value. These statistics were compiled over 100 trials. Figure 3.3a charts cross-validation performance over a range of probability cutoffs. At a probability cutoff of 0.5, we observed a 50% true positive rate and a 15% false positive rate. This shows marked improvement over a random predictor where we would expect to see the same true positive rate as false positive rate.

Declassifying 1% of the known functional orthologous and non-orthologous pairs reduces the amount of information available to the algorithm and, thus, can reduce its predictive ability. To assess the severity of this effect, we repeated the cross-validation analysis at varying percentages of declassification of positive and negative data (ranging from 1% to 100%) (Fig. 3.3b). For instance, changing the amount of declassification of available training data from 1% to 25% reduced the maximum precision from 83 to 75%. Further declassification yielded more marked reductions in precision and recall.

Discussion

Specific examples of yeast/fly functional orthologs resolved by the network-based approach are shown graphically in Figure 3.4. In Fig. 3.4a, yeast transportin (Kap104) is orthologous to both Trn and CG8219 in fly with highly significant sequence homology (BLAST E-values $9\text{x}10^{-128}$ and $7\text{x}10^{-96}$, respectively). Transportin is a member of a complex responsible for the nuclear import of mRNA binding proteins and is known to be highly conserved among diverse organisms[25]. Drosophila Trn was identified using sequence homology based on human transportin1[26]. Both Trn and CG8219 in fly interact with orthologs of members of the Kap104- associated complex in yeast[27] suggesting that both of these fly proteins may participate in the functionally similar complex in fly. Our analysis suggests that CG8219 retains more of the original functions of Kap104 (probability of functional orthology of 47% versus 41% for the Kap104/Trn pairing) due to their conserved interactions with members of the spliceosome complex (Yju2, Ssa1, and Ssa2 in yeast). This result is a case in which the most sequence-similar protein does not appear to be the most functionally-related protein in an orthologous group given the current network data.

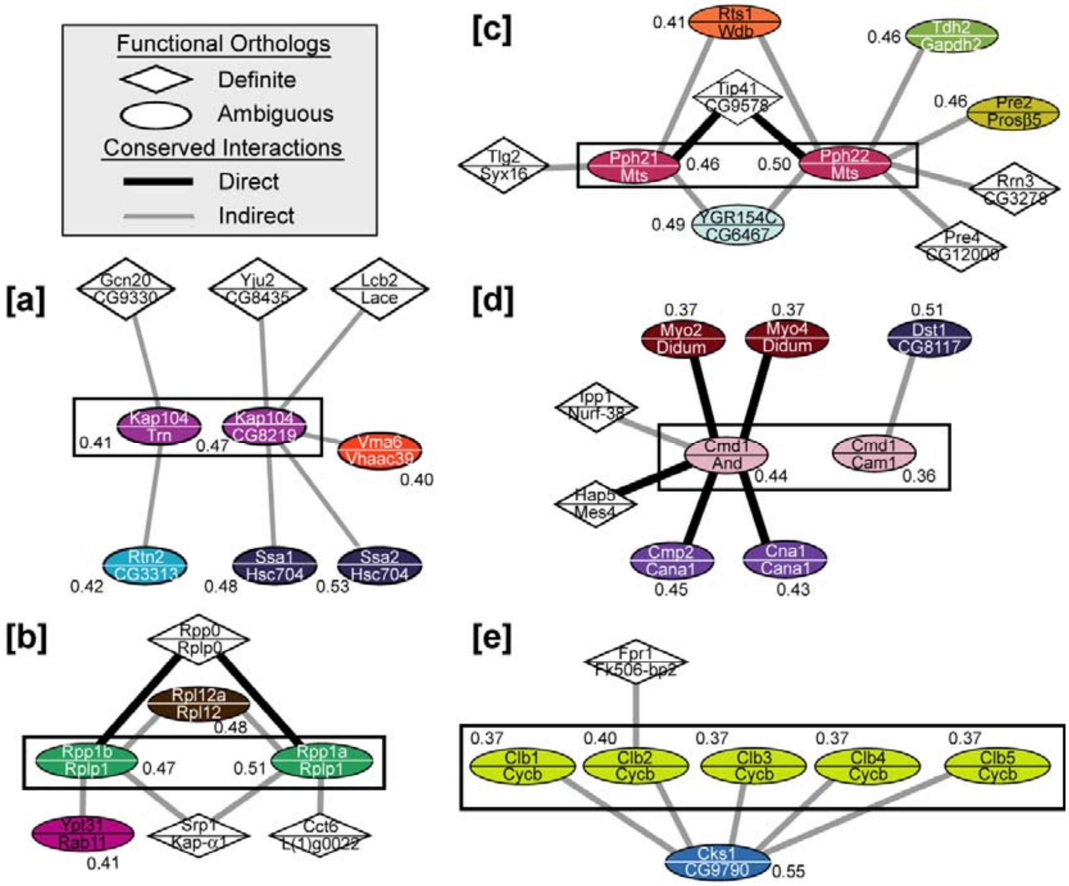


Figure 3.4: Example orthologs resolved by network conservation.

Each node represents a putative functional match between a yeast/fly protein pair (with names shown above/below the line, respectively). Links between nodes denote conserved interactions (thick black, direct interactions in both species; thin gray, indirect interaction in one of the species). Diamond- vs. oval-shaped nodes represent definite vs. ambiguous functional orthologs. Oval nodes of the same color represent ambiguous protein pairs belonging to the same Inparanoid cluster. The mean probability of functional orthology is given next to each ambiguous pair. Cluster 246 [a], 1439 [b], 211 [c], 917 [d], and 1104 [e] show examples of clusters that were disambiguated by conserved network information; the cluster resolved in each panel is outlined by a black rectangle.

The cluster in Fig. 3.4c contains two alternative catalytic alpha subunits of the protein phosphatase type 2A family (yeast Pph21 and Pph22). Both alternatives interact with a member of the beta subunit (Rts1) and have high sequence similarity to the fly Mts protein (75% identity for Pph21, 76% for Pph22). Since Pph21 and Pph22 are at least partially redundant [disruption of both genes in combination is synthetic lethal[28]], it appears that the array of interactions carried out by Mts is conserved across the two yeast orthologs. For instance, based on the available protein interaction data, Pph22 alone has conserved interactions with the proteasome (Pre2/Prosβ5 and Pre4/CG12000), which has been shown to be important for the role of the Pph21/22 complex in degradation of Swe1p[29].

As a final example, Fig. 3.4d shows evidence that the yeast Calmodulin (Cmd1) protein is functionally orthologous to fly Androcam (And) rather than to the more sequence-similar fly Calmodulin (Cam1; 60% identity versus 51% for And). The existence of many conserved interactions for the Cmd1/And pair, compared to only one for Cmd1/Cam1, does not appear to be a result of incomplete coverage: Cmd1 has a total of 61 interactions in the yeast network, and Cam1 and And have 19 and 26 interactions, respectively, in the fly network (most of these do not appear in Fig. 3.4 because the network alignment only shows interactions that are conserved). Furthermore, multiple sequence alignment and phylogenetic analysis of these genes over a larger number of organisms, including worm and mammals, indicates a closer phylogenetic relationship for yeast Cmd1 and fly And, supporting our hypothesis that they are the true functional orthologs (Supplemental Figure 1). This apparent discrepancy between functional and sequence similarity is probably a result of the large amount of sequence variability among the calmodulin family of proteins [30] and would have been difficult to probe without protein network information.

In future work, it is possible that incorporating additional types of conserved linkages, such as transcriptional interactions[31], synthetic-lethal interactions[32], and co-expression relations[10] will allow a more complete and multi-faceted view of protein function. Second, this method would benefit from a more accurate understanding of network evolution. At the core of our approach is a model for measuring the divergence of orthologous proteins by means of a network "conservation index". It encapsulates the notion that shorter evolutionary distances correspond to greater relative numbers of conserved interactions. However, a more sophisticated metric might represent explicit mechanisms of network evolution, such as formation of new interactions through gene mutation or duplication[1]. It should also be noted that comparative methods rely on the conservation of function between evolutionarily related proteins, and that this functional similarity may be lost among orthologs due to large evolutionary distance; thus, network-based methods which search for the absence of a functional ortholog may also be useful. Finally, further work is also needed to analyze the impact of data quality, i.e., numbers of false-positive and false-negative interactions[33]. False positives are largely mitigated by the focus on only those interactions that are conserved across species, because spurious interactions are typically not reproducible [17]. False negatives are a larger concern, because they might cause a functionally orthologous pair to be wrongly rejected due to lack of conserved interactions. Certainly, a preponderance of conserved interactions for one particular pair of proteins versus others provides evidence that these proteins are indeed functional orthologs. Although the expected number of false negative interactions will decrease with forthcoming interaction datasets, future approaches may explicitly incorporate the false negative rate into the probabilistic model.

In summary, we have presented an algorithm that incorporates protein interaction measurements to achieve more specific discrimination of functional orthologs than is possible with sequence-based methods alone. It is built on the concept that conserved proteins typically do not function independently but rely on interactions with other proteins to form conserved pathways, and that the specific patterns of conservation of these pathways are informative for determining which cross-species protein pairs have similar functional roles. As these methods mature and as ever greater numbers of protein interactions become available across species, comparative network analysis will play an increasingly central role as a bridge between protein sequence, evolution, and function.

Methods

Inparanoid clusters generation.

The complete sets of 5,878 yeast and 18,746 fly protein sequences were downloaded from Saccharomyces Genome Database[34] and Flybase[35], respectively. Protein sequences for both species were clustered together into orthology groups using the Inparanoid algorithm[5] with default parameters (overlap threshold=0.5; confidence=0.05). Inparanoid optionally allows a third genome to be used as an outgroup, which can detect missing sequences and thus improve ortholog detection. However, use of *E. coli* as an outgroup had a negligible impact on our analysis.

Network alignment.

A global network alignment between yeast and fly was constructed as described in [16], with the difference that Inparanoid clusters were used instead of BLAST E-values for pairing proteins between the two networks. Briefly, the network alignment is represented as a graph of nodes and links (Fig. 3.2b). Each node denotes a pair of putatively orthologous proteins *a* and *a'*. Each link between a pair of nodes *a*/*a'* and *b*/*b'* denotes a conserved protein interaction, i.e., an interaction observed for both (*a*, *b*) and (*a'*, *b'*). To tolerate a certain amount of missing interaction data, "indirect" links are also defined if a pair of proteins interacts in one species (e.g., *a* and *b* interact) and the other pair of proteins (e.g., *a'* and *b'*) are at most distance two in their corresponding interaction maps. Links involving network distances greater than two, or for which the proteins of both species are at distance two, are not allowed [15]. The yeast/fly network alignment contains 388 nodes (spanning 348 yeast and 256 fly proteins) linked by 308 conserved interactions (110 direct and 198 indirect).

Each node in the alignment graph is associated with a state *z*, indicating whether that protein pair represents true functional orthology (*z=1*) or not (*z=0*). Links between nodes that are each associated with true functional orthology are said to be "strongly conserved". To compute the frequencies shown in Fig. 3.1a, the protein pair in each Inparanoid group having the lowest BLAST E-value is set to z=1; all others to z=0.

Conservation index.

We define the conservation index *c* of node *i* (representing protein pair *a*/*a'*) as twice its number of strongly conserved interactions divided by its total number of interactions over both species:

$$c(i) = \frac{2d(i)}{d(a) + d(a')}$$

where $d(i)$ denotes the number of strongly conserved links involving node i, while $d(a)$ and $d(a')$ denote the degrees (numbers of interactions) of proteins a and a' in their respective single-species networks.

Probabilistic model.

We model the orthology relations for two species using a Markov random field [23]. This model is specified by an undirected graph $G=(V, E)$ corresponding to a network alignment, and conditional probability distributions which relate the event that a given node represents a functionally orthologous pair with those events for its neighbors. A Markov random field model is specified in terms of potential functions on the cliques in the graph:

$$P(\vec{z}) = \frac{1}{Z}\exp\{-U(\vec{z})\}$$

where $\vec{z}$ is some assignment to the states of all nodes in the graph, $U(\cdot)$ is an "energy" function which integrates the potentials over all cliques in the graph, and Z is a normalizing constant. It is not necessary to compute the normalization constant, since all that is required are the conditional probabilities for each node given its neighbors (rather than the joint distribution). For computational efficiency, we used the common auto-logistic model[23] which assigns zero potential to cliques of size > 2. Under this model, the energy takes the form:

$$U(\vec{z}) = -\sum_i \alpha_i z_i - \sum_{(i,j)\in E} \beta_{ij} z_i z_j$$

which, when substituted into the equation for $P(\vec{z})$ above, reduces to a logistic function. Based on our initial observation that the functional orthology of a node is a function of its conservation index (well approximated by a logistic function—see Fig. 3.1a and Results), we set $\alpha_i = \alpha$ and $\beta_{ij} = \beta_i = 2\beta \,/\, [d(a_i) + d(a_i')]$ to obtain the following:

$$P\left(z_i \mid z_{N(i)}\right) \;=\; \frac{1}{1+\exp\left\{-\alpha_i - \sum\limits_{j\in N(i)} \beta_{ij} z_j\right\}} \;=\; \frac{1}{1+\exp\{-\alpha - \beta c(i)\}}$$

where *N(i)* is the set of neighbors of node *i*, and $z_{N(i)}$ denotes the set of all z_j such that $j \in N(i)$. Note that α_i and β_{ij} could be set to accommodate other equations for conservation index, as long as they are linear in the number of strongly conserved neighbors *d*(*i*).

Fitting the logistic function.

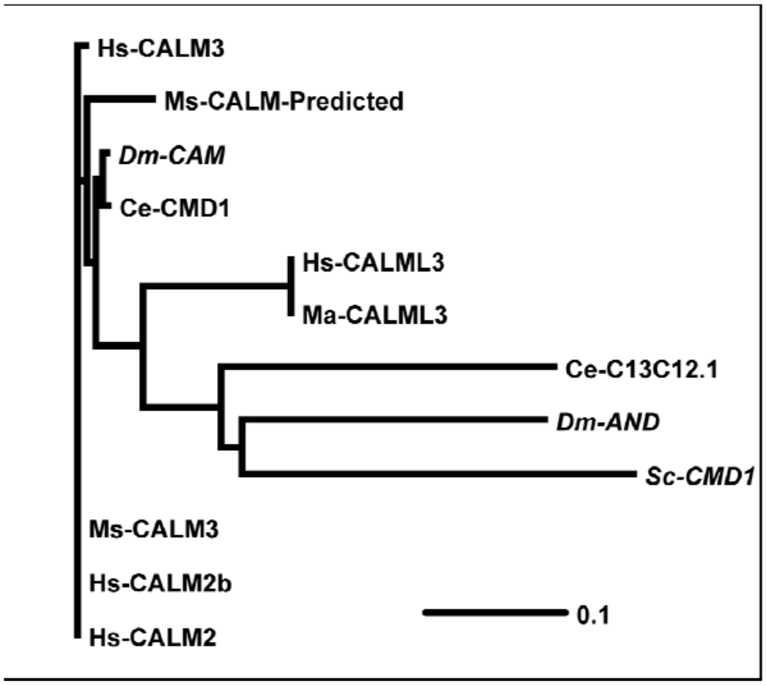
To provide a set of training data for fitting the parameters α and β of the logistic function, 100 of the 212 definite functional orthologs having at least one conserved interaction were randomly chosen as positive examples, and their states set to *z=1*. Negative examples of "non-orthologs" were generated by randomly selecting 100 yeast proteins and pairing each with its best BLAST matching fly protein not in the same cluster; their states were set to *z=0* (ideally, the negative training data would consist of orthologs that are not functional orthologs, but few such examples exist). Parameters α and β were optimized by maximizing the product of $P(z_i \mid z_{N(i)})$ over all positive and $(1 - P(z_i \mid z_{N(i)}))$ over all negative training data using the method of conjugate gradients[36]. The optimal logistic function is shown in Fig. 3.1b. Note that the equal numbers of positive and negative training data assume a prior probability of 0.5 of observing a true functional ortholog. Although the actual prior is unknown and may differ from this value, $P(z_i \mid z_{N(i)})$ remains monotonically related to the true probability of functional orthology.

Orthology inference.

We used the above model to estimate the final posterior probabilities $P(z_i)$ using the method of Gibbs sampling[37]. In this approach, nodes representing ambiguous functional orthologs are each assigned a temporary state $z=0$ or 1, initially at random. At each iteration, a node i is sampled (with replacement) and its value of z_i is updated given the states of its neighbors, $z_{N(i)}$. The new value of z_i is set to 0 or 1 with probability $P(z_i \mid z_{N(i)})$. Over all iterations, the nodes designated as definite functional orthologs and "non-orthologs" are forced to states of 1 and 0, respectively. This process is illustrated in Figs. 3.2c and 3.2d.

The Gibbs sampling procedure was carried out for an initial period of $2 \cdot 10^6$ "burn-in" iterations. From this point onward, $2 \cdot 10^7$ additional iterations were performed and statistics computed on the fraction of iterations in which each node acquires a "functionally orthologous" $z=1$ state. The final probabilities of functional orthology for each node, $P(z_i)$, were estimated as this fraction. The above numbers of iterations were chosen to ensure stabilization of the probability estimates such that results were stable across multiple runs of random initialization configurations (standard deviations for each $P(z_i)$ are available on the website). Compiled results were aggregated over one hundred separate runs of the algorithm and mean probabilities reported.

Supplemental Figures



Supplemental Figure 3.1: Phylogenetic analysis of Calmodulin orthologs

Analysis of yeast, CMD1 (Sc-CMD1), and fly, AND and CAM (Dm-AND and Dm-CAM respectively). Although Sc-CMD1 is more sequence similar to Dm-CAM based on BLAST-similarity, our algorithm predicted Dm-AND to be the more functionally similar ortholog. The tree built on a multiple sequence alignment from related yeast,fly,worm, mouse and human sequences shows that in the context of related sequences, Sc-CMD1 and Dm-AND are more similar.

Acknowledgments

We thank Tomer Shlomi and Ryan Kelley for critical reading of the manuscript as well as Silpa Suthram for providing interaction data. We gratefully acknowledge the following sources of funding for this project: award R01-GM070743-01 from the NIGMS (TI), an Alon fellowship (RS), and a Quantitative Systems Biology grant from the NSF (SB). TI is a fellow of the David and Lucille Packard Foundation.

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

Bandyopadhyay S, Sharan R, Ideker T. *Systematic identification of functional orthologs by protein network comparison.* **Genome Research** 2006; 16(3):428-35.

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

Bibliography

1. Wagner A. How the global structure of protein interaction networks evolves. *Proc R Soc Lond B Biol Sci* 2003;270:457-66.

2. Brenner SE. Errors in genome annotation. *Trends Genet* 1999;15:132-3.

3. Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE. Genome annotation assessment in Drosophila melanogaster. *Genome Res* 2000;10:483-501.

4. Sjolander K. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 2004;20:170-9.

5. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001;314:1041-52.

6. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;13:2178-89.

7. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000;28:33-6.

8. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389-402.

9. Eisen JA, Wu M. Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor Popul Biol* 2002;61:481-7.

10. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003;302:249-55.

11. van Noort V, Snel B, Huynen MA. Predicting gene function by conserved co-expression. *Trends Genet* 2003;19:238-42.

12. Fields S, Song O. A novel genetic system to detect protein-protein interactions. *Nature* 1989;340:245-6.

13. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;422:198-207.

14. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* 2001;11:2120-6.

15. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A* 2003;100:11394-9.

16. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res* 2004;32:W83-8.

17. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* 2005;102:1974-9.

18. Leone M, Pagnani A. Predicting protein functions with message passing algorithms. *Bioinformatics* 2005;21:239-47.

19. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* 2003;19 Suppl 1:i197-204.

20. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* 2003;21:697-700.

21. Espadaler J, Aragues R, Eswar N, Marti-Renom MA, Querol E, Aviles FX, Sali A, Oliva B. Detecting remotely related proteins by their interactions and sequence similarity. *Proc Natl Acad Sci U S A* 2005;102:7151-6.

22. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 2002;30:303-5.

23. Besag J. Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statist. Soc.* 1974;B:192-236.

24. Pearl J. Probabalistic reasoning in intelligent systems : networks of plausible inference. San Mateo, Calif.: Morgan Kaufmann Publishers, 1988:xix, 552 p.

25. Aitchison JD, Blobel G, Rout MP. Kap104p: a karyopherin involved in the nuclear transport of messenger RNA binding proteins. *Science* 1996;274:624-7.

26. Siomi MC, Fromont M, Rain JC, Wan L, Wang F, Legrain P, Dreyfuss G. Functional conservation of the transportin nuclear import pathway in divergent organisms. *Mol Cell Biol* 1998;18:4141-8.

27. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002;415:180-3.

28. Ronne H, Carlberg M, Hu GZ, Nehlin JO. Protein phosphatase 2A in Saccharomyces cerevisiae: effects on cell growth and bud morphogenesis. *Mol Cell Biol* 1991;11:4876-84.

29. Yang H, Jiang W, Gentry M, Hallberg RL. Loss of a protein phosphatase 2A regulatory subunit (Cdc55p) elicits improper regulation of Swe1p degradation. *Mol Cell Biol* 2000;20:8143-56.

30. Tombes RM, Faison MO, Turbeville JM. Organization and evolution of multifunctional Ca(2+)/CaM-dependent protein kinase genes. *Gene* 2003;322:17-31.

31. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;431:99-104.

32. Guarente L. Synthetic enhancement in gene interaction: a genetic tool come of age. *Trends Genet* 1993;9:362-6.

33. Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data? *J Mol Biol* 2003;327:919-23.

34. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res* 2004;32 Database issue:D311-4.

35. Drysdale RA, Crosby MA, Gelbart W, Campbell K, Emmert D, Matthews B, Russo S, Schroeder A, Smutniak F, Zhang P, Zhou P, Zytkovicz M, Ashburner M, de Grey A, Foulger R, Millburn G, Sutherland D, Yamada C, Kaufman T, Matthews K, DeAngelo A, Cook RK, Gilbert D, Goodman J, Grumbling G, Sheth H, Strelets V, Rubin G, Gibson M, Harris N, Lewis S, Misra S, Shu SQ. FlyBase: genes and gene models. *Nucleic Acids Res* 2005;33 Database Issue:D390-5.

36. Press WH. Numerical recipies in FORTRAN : the art of scientific computing. Cambridge: Cambridge University Press, 1992:xxvi, 963 p.

37. Smith A, Roberts G. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods *Journal of the Royal Statistical Society* 1993;B 55:3-23.

Chapter 4. Functional maps of protein complexes from quantitative genetic interaction data

Abstract

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

Introduction

Genetic interactions are logical relationships between genes that occur when mutating two or more genes in combination produces an unexpected phenotype[1-3]. Recently, rapid screening of genetic interactions has become feasible using Synthetic Genetic Arrays (SGA) or diploid Synthetic Lethality Analysis by Microarray (dSLAM)[4,5]. SGA pairs a gene deletion of interest against a deletion to every other gene in the genome (in turn). The growth / no growth phenotype measured over all pairings defines a *genetic interaction profile* for that gene, with no growth indicating a synthetic-lethal genetic interaction. Alternatively, all combinations of double deletions can be analyzed among a functionally-related group of genes[6-8]. A recent variant of SGA termed E-MAP[7] has made it possible to measure continuous rates of growth with varying degrees of epistasis (based on imaging of colony sizes). "Aggravating" interactions are indicated if the growth rate of the double gene deletion is slower than expected, while for "alleviating" interactions the opposite is true[9,10].

One popular method to analyze genetic interaction data has been to hierarchically cluster genes using the distance between their genetic interaction profiles. Clusters of genes with similar profiles are manually searched to identify the known pathways and complexes they contain as well as any genetic interactions between these complexes. This approach has been applied to several large-scale genetic interaction screens in yeast including genes involved in the secretory pathway[8] and chromosome organization[6]. Segré et al.[11] extended basic hierarchical clustering with the concept of "monochromaticity", in which genes were merged into the same cluster based on minimizing the number of interactions with other clusters that do not share the same classification (aggravating or alleviating).

Another set of methods has sought to interpret genetic relationships using physical protein-protein interactions[12]. Among these, Kelley and Ideker[13] used physical interactions to identify both "within-module" and "between-module" explanations for genetic interactions. In both cases, modules were detected as clusters of proteins that physically interact with each other more often than expected by chance. The "within-module" model predicts that these clusters directly overlap with clusters of genetic interactions. The "between-module" model predicts that genetic interactions run between two physical clusters that are functionally related. This approach was improved by Ulitsky *et al.*[14] using a relaxed definition of physical modules. In related work, Zhang et al.[15] screened known complexes annotated by the Munich Information Center for Protein Sequences (MIPS) to identify pairs of complexes with dense genetic interactions between them.

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

Results

Characterization of Genetic and Physical Networks.

We first sought to quantitatively confirm whether, and to what degree, physical and genetic interactions could indicate common membership in a protein complex. To provide genetic data for analysis, we obtained the previously-published results from a large E-MAP of yeast chromosomal biology[6]. This study consisted of genetic interactions measured among 743 genes (including 74 essential genes), yielding quantitative values for 182,669 gene pairs (66% of all possible pair-wise measurements). Each gene pair was assigned an S-score, where positive scores indicate protein pairs for which the double mutant grows better than expected (i.e., an alleviating interaction) and negative scores indicate pairs for which the double mutant grows worse than expected (i.e., a synthetic sick/lethal or aggravating interaction) where the expectation is that the double-deletion of unrelated proteins will have a growth rate equivalent to the multiplicative product of the two individual growth rates[16]. In all, 14,237 gene pairs (8%) showed strong genetic interactions with |S| > 2.5. Physical interactions were taken from a recent computational integration of two large datasets measured by co-immunoprecipitation followed by mass spectrometry[17]. This study assigned to each pairwise interaction a Purification Enrichment (PE) score, with larger values representing a greater likelihood of true binding.

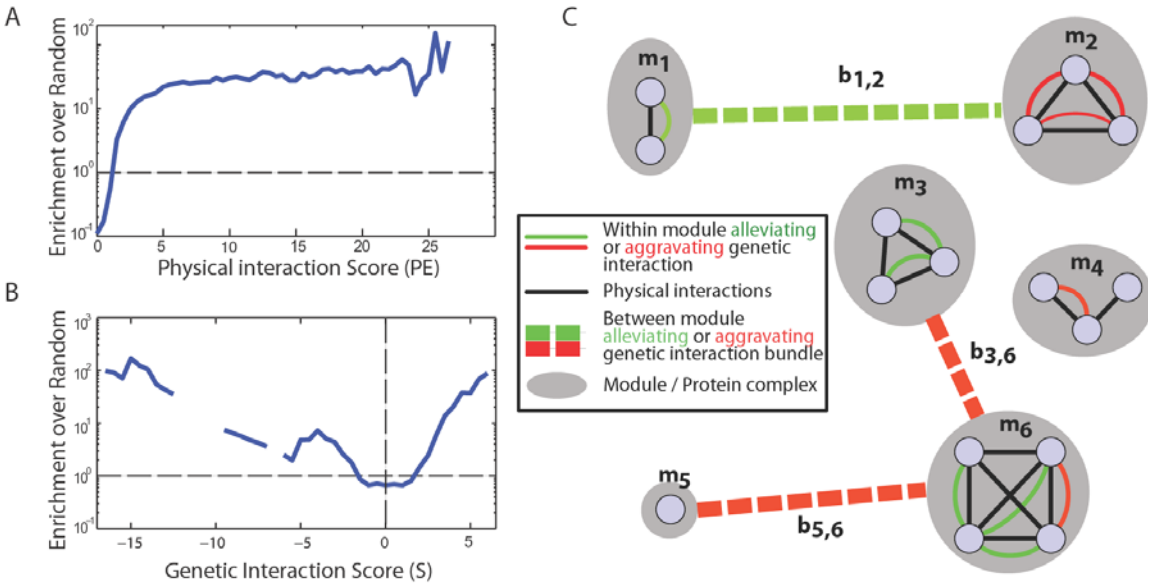


Figure 4.1. Combining physical and genetic interactions to define protein complexes.

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

Figurea confirms that protein pairs with higher PE-scores are more likely to operate in a known small-scale protein complex recorded in the MIPS database[18] (versus protein pairs chosen at random). This result is expected considering that PE-scores were trained based on these complexes[17]. Figureb shows that protein pairs with both positive and negative S-scores are more likely to operate within a known complex. Positive (alleviating) interactions are well-known to occur between subunits of a complex[6]. Negative (aggravating) interactions are to a lesser degree so, although the mechanism has not been as clear as for the alleviating case[19]. By comparing the magnitudes of enrichment between Figurea and b, it is apparent that extreme S-scores are at least as indicative of co-complex membership as strong PE-scores, if not more so (~100-fold enrichment versus ~50-fold enrichment, respectively). Together, these exploratory findings suggest that the physical and genetic information can indeed provide a basis for the identification of protein pairs involved in the same complex.

Functional maps of protein complexes involved in yeast chromosomal biology.

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

Methods.

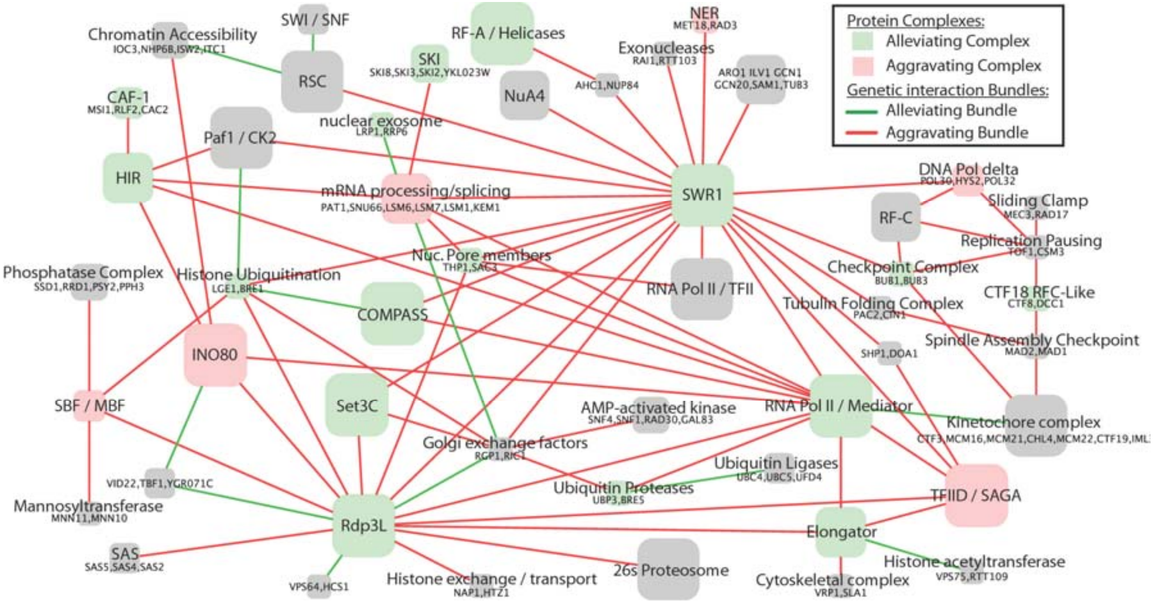


Figure 4.2. Global map of protein complexes involved in yeast chromosome biology.

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

Applying this method, we identified 91 distinct modules with an average size of 4.1 proteins per module. Figure 4.2 gives an overview of a subset of the identified modules and inter-module links. Complete results are catalogued at http://www.cellcircuits.org/Bandyopadhyay2008/html/. Overall, these results suggest ten novel complexes not recorded in either the small-scale or high-throughput MIPS compendium, covering 23 proteins in total. The results also identify 84 new subunits of known complexes (Supplemental Materials). Through permutation testing, 19 versus 9 of the identified modules could be categorized as enriched for alleviating or aggravating genetic interactions, respectively. A total of 313 significant genetic relationships were identified between modules, 94 versus 219 of which were enriched for alleviating or aggravating interactions.

Comparison to related approaches.

The method of choice for interpreting quantitative genetic interactions has been hierarchical clustering (HCL) of genes based on pair-wise distances between their genetic interaction profiles[6,8]. We compared the clusters obtained using HCL to the modules obtained with our present approach (Bandyopadhyay *et al.*) using three gold-standard metrics: gene co-expression (Figure 4.3a), co-functional annotation (Figure 4.3b), or membership in the same previously-identified complex (Figure 4.3c). To ensure a fair comparison between the two approaches, HCL and Bandyopadhyay *et al.* were evaluated across a range of coverages (number of gold-standard gene pairs recovered by the predicted clusters/modules; see Methods). For all three benchmarks, our performance was substantially higher than that of the HCL-based approach at most levels of coverage (and at a level of coverage corresponding to the 91 modules reported above; dotted vertical line in Figure 4.3).

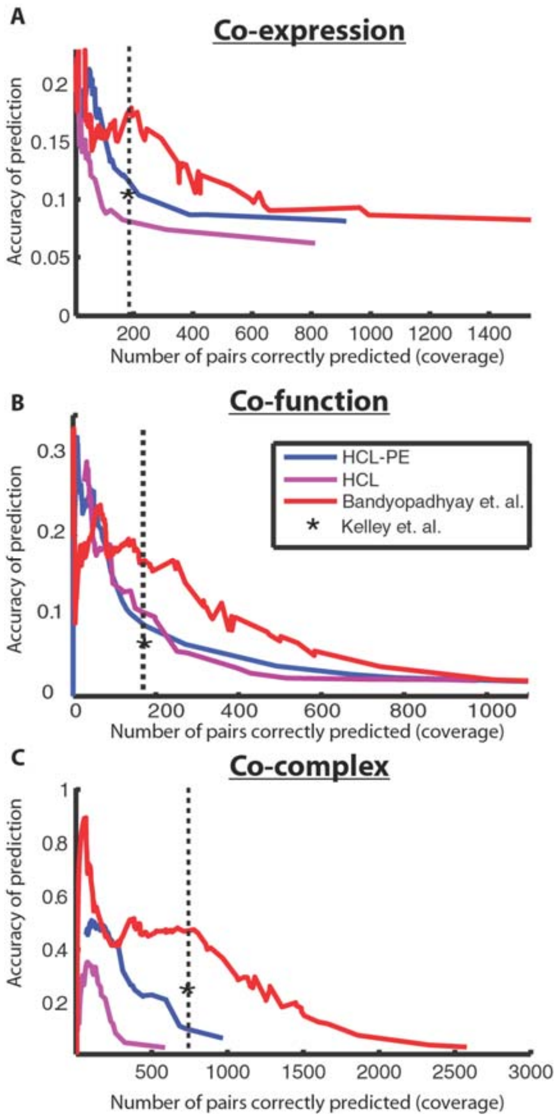


Figure 4.3. Performance of complex identification.

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

We considered that one reason why HCL performed less favorably might be that it was not given access to the same information (i.e., the physical network). This is especially true for the metric based on previously-identified complexes, in which complexes were annotated based on the same high-throughput protein interactions used here. To investigate this possibility, we extended HCL to incorporate physical interactions in a straightforward fashion, by merging only those clusters which share a physical interaction between them (HCL-PE). Although this approach outperformed hierarchical clustering without physical interactions, it was outperformed by the present approach by at least 50% across the three metrics. Finally, our method also shows improvement over the previous approach of Kelley and Ideker[13] for integrating genetic and physical interactions (Figure 4.3).

Aggravating complexes tend to be essential.

Nineteen versus nine of the learned modules had significant enrichment for alleviating versus aggravating genetic interactions, respectively. Identification of “alleviating” modules is expected, since subunits of a complex operate together and the phenotypic effect of removing any pair of proteins in a complex should be no worse than removing any single protein individually. The presence of aggravating interactions within modules was more intriguing. One way in which aggravating interactions could occur among the subunits of a complex is if its function is essential, i.e., the loss of the complex’s function causes a lethal phenotype. In these cases, some protein subunits should be encoded by essential genes, while other subunits might be redundant and thus essential in pairwise combinations[19].

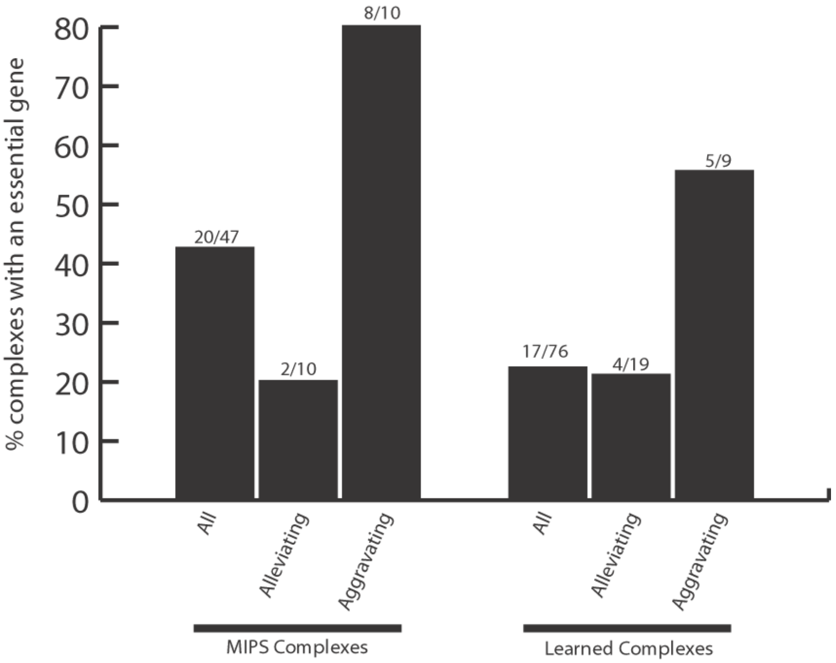


Figure 4.4. Aggravating complexes are more likely to contain essential genes.

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

To test the hypothesis that essential genes are more likely in aggravating modules, we analyzed both MIPS small-scale complexes and our learned modules for the presence of essential genes (Figure 4.4). We found that 80% of aggravating MIPS complexes contained an essential gene, compared to only 20% of alleviating MIPS complexes (a four-fold increase). Similarly, of the aggravating modules determined by our approach, 55% contained an essential gene compared to only 21% of alleviating modules (a 2.6-fold increase). These results are not correlated with module size, as the median size of aggravating learned modules is less than the median size of alleviating learned modules.

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

Discussion

Figure 4.5 presents detailed diagrams of example functional relationships elucidated by our module mapping method. Figure 4.5a shows the alleviating relationship between the RTT109-VPS75 histone acetyltransferase complex[6,20,21] and Elongator, a complex that is associated with RNA Polymerase II and is involved in transcriptional elongation[22]. Since several subunits both of Elongator and RTT109/VPS75 have been shown to be involved in histone acetylation levels[21,23], these two complexes may operate together to effectively clear histones from actively transcribed regions. To identify further mechanisms of their cooperation, future studies may search for specific residues of histone H3 whose acetylation levels are modulated by both complexes. This example highlights the utility of an integrated approach, since although RTT109 and VPS75 are known to form a complex their genetic interaction profiles are not congruent (correlation of profiles of -0.1) and had been missed by hierarchical clustering. Figure 4.5b highlights non-essential components (LRP1 and RRP6) of the exosome, which contributes to the quality-control system that retains and degrades aberrant mRNAs in the nucleus[24]. These components have alleviating interactions with a complex composed of Lsm proteins involved in mRNA decay.

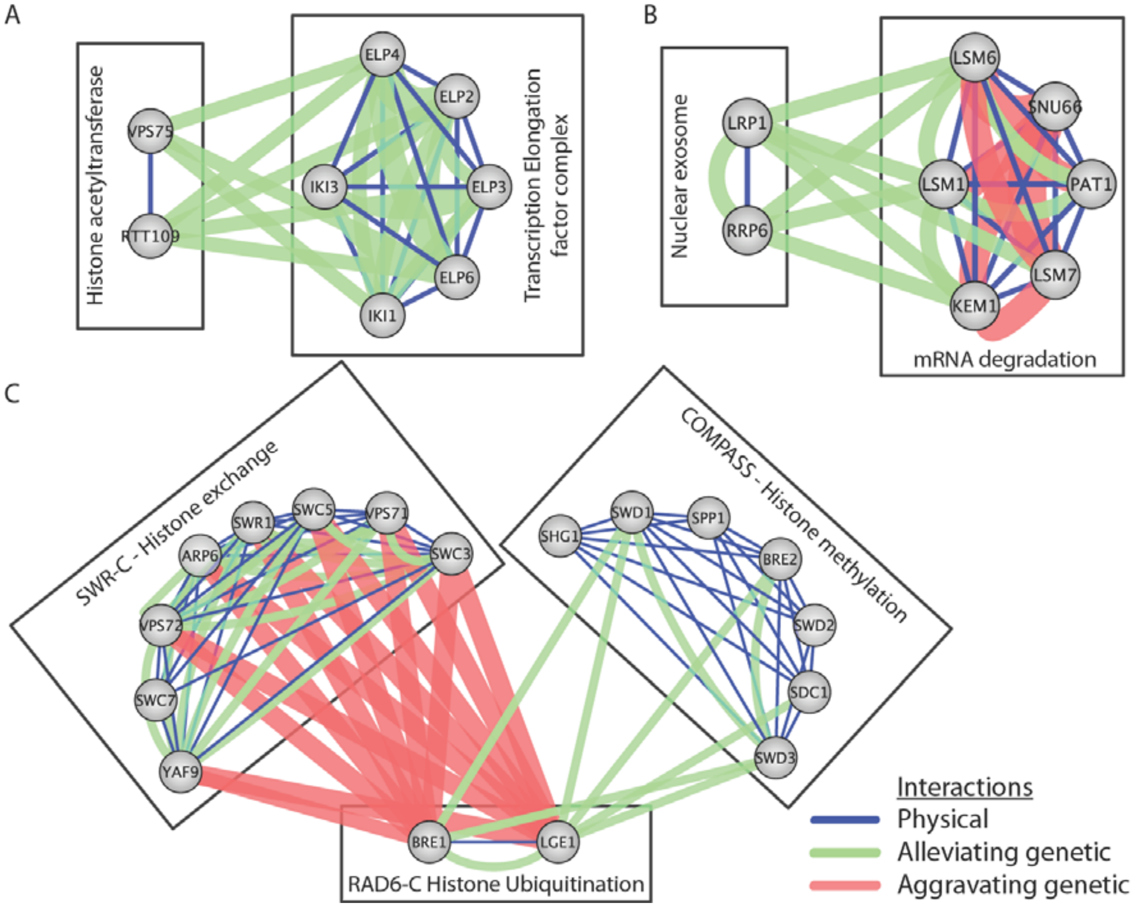


Figure 4.5. Pathway models identify novel functional associations among cellular machinery.

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

Figure 4.5c centers on BRE1/LGE1, subunits of the Rad6 Histone Ubiquitination Complex (RAD6-C; the Rad6 protein itself was not covered by the original E-MAP screen)[25,26]. RAD6-C is functionally connected with two other complexes, SWR-C and COMPASS. SWR-C functions to regulate gene expression through the incorporation of transcriptionally-active histone variant H2AZ[27-29], while COMPASS is involved in mediating transcriptional elongation and silencing at telomeres through methylation of histone H3[30]. Interactions between RAD6-C and SWR are aggravating, suggesting synergy or redundancy towards an essential cellular function. Interactions between RAD6-C and COMPASS are alleviating, suggesting they operate in a potentially serial fashion. Consistent with this analysis, it has been shown that histone H2B ubiquitination by RAD6-C is a prerequisite for histone H3 methylation by COMPASS[31,32].

Several trends emerge from the performance analysis in Figure 4.3. First, genetic interaction data alone can yield substantial information about molecular pathways. Functionally similar proteins often have similar profiles of genetic interaction, a feature we have previously exploited to identify functional interactions between complexes as well as to identify new members of complexes based on a combination of weak physical and genetic data[13]. On the other hand, the ability to detect complexes can be greatly improved by adding information about protein physical interactions. Even the straightforward HCL-PE method was able to greatly improve the accuracy and coverage according to most metrics, while the greatest performance was achieved by the improved probabilistic framework we have presented in this study. This framework has led to the inclusion of YKL023W as a potential new member of the SKI complex and YGR071C in a complex with VID22/TBF1 (Figure 4.2), for a total of 84 novel protein subunit assignments to complexes (Supplemental Data). Both of these examples have both physical and genetic support and would have been missed by an approach based on either type of interaction alone.

Future work may seek to incorporate yet additional types of linkages such as protein-DNA interactions[33,34], kinase-substrate phosphorylations[35], or other genetic perturbation data such as eQTLs[36]. There are also opportunities to refine the modeling framework further. Here, a gold-standard set of complexes was used to explicitly learn the relationship between physical interactions, genetic interactions, and module membership. This supervised approach could be extended to also learn which features best indicate the inter-module functional relationships, perhaps through curation of a gold-standard set of interacting complexes.

Methods

Problem definition.

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

Scoring module co-membership.

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

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

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

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

The probability P_{within} is determined using logistic regression training on 217 complexes curated from small-scale studies in MIPS[18]. $P_{background}$ is the probability of randomly observing the observed value (*PE* or *S*) for the pair (*a*,*b*) in the background of all gene pairs. As shown in Figure and 1b, it is clear that higher values of *PE* are predictive of MIPS complex membership. As both positive and negative values of *S* are predictive, $LLR_S(a,b)$ is trained on the absolute value of *S*. A third predictor based on the correlation of genetic interaction profiles was also evaluated but did not result in any gain in performance (Supplemental Figure).

Scoring inter-module links.

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

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

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

Global optimization of module memberships and links.

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

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

The first term accumulates the within-module scores among gene pairs assigned to the same module. The second term accumulates the inter-module scores for gene pairs spanning any two modules. Gene pairs spanning unlinked modules do not contribute to *L*. The final term is a tunable reward which scales with module size. Larger values of *α* result in fewer, larger complexes. The final module map shown in Figure 4.2 was generated using α=1.6, based on its good coverage and performance across all three metrics in Figure 4.3.

Module search.

Assignment of gene to modules and of inter-module links is performed using a simple variant of UPGMA hierarchical clustering[37]: (a) Initially, each gene is assigned to a separate module; (b) Each pair of modules (m_1, m_2) is evaluated for merging into a single module $m = m_1 \cup m_2$; the pair-wise merging that results in the largest increase in *L* is chosen; (c) Repeat step b until no module merge operation increases *L*.

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

Subsequent to the above procedure, each between-module link is evaluated to assess its significance and whether it represents predominantly aggravating or alleviating genetic interactions. A two-tailed p-value is computed by indexing the sum of *S*-scores for gene pairs falling across the two modules against a distribution of 10^6 sums of equal numbers of *S*-scores drawn from random gene pairs. To account for multiple testing, we use the distribution of between-module p-values to compute a local false discovery rate (FDR)[38]. All reported between-module links have an inferred FDR of <10% with the global map in Figure 4.2 constrained to links with an FDR of <1%. Module maps in Figure 4.2 and Figure 4.5 are visualized using the Cytoscape package[39,40].

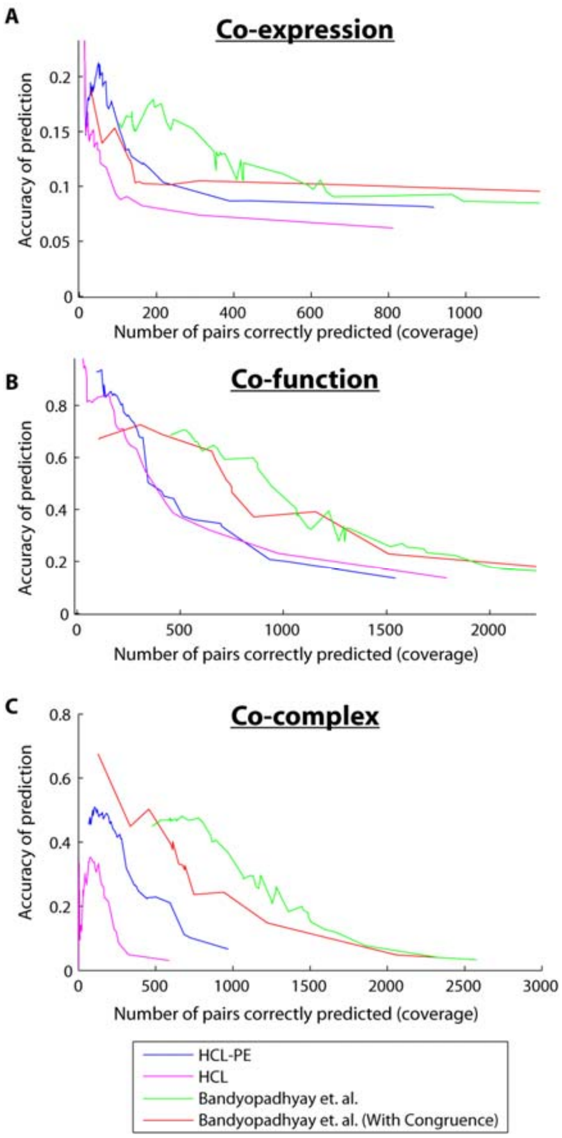
To label modules as "aggravating" or "alleviating" (Figure 4.2), the sum of *S*-scores for gene pairs assigned to the same module is compared to a distribution of sums of equal numbers of randomly drawn *S*-scores. Modules with a two-tailed p-value < 0.05 are labeled as either alleviating (right tail) or aggravating (left tail).

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

Co-expressed gene pairs were defined using gene expression datasets culled from the Stanford Microarray Database covering ~790 conditions[41]. The validation set was taken as the top 5% (13,014) of pairs ranked by Pearson correlation coefficient. The co-function set was based on yeast Gene Ontology annotations from November 2005 which predates the publication of large scale TAP-MS studies that were used to generate the PE-score. This set was taken as the top 5% (13,052) most functionally similar gene pairs covered in the E-MAP. Functional similarity was determined by comparison to the background probability of picking two genes with the same shared functional annotation from the entire yeast genome (via a hypergeometric test). Similar analysis using current Gene Ontology annotation was also performed (Supplemental Figure). The co-complex validation set was defined as gene pairs from 846 MIPS complexes annotated using high-throughput approaches (with interactions also appearing in small-scale studies removed) for a total of 2,885 gold-standard pairs.

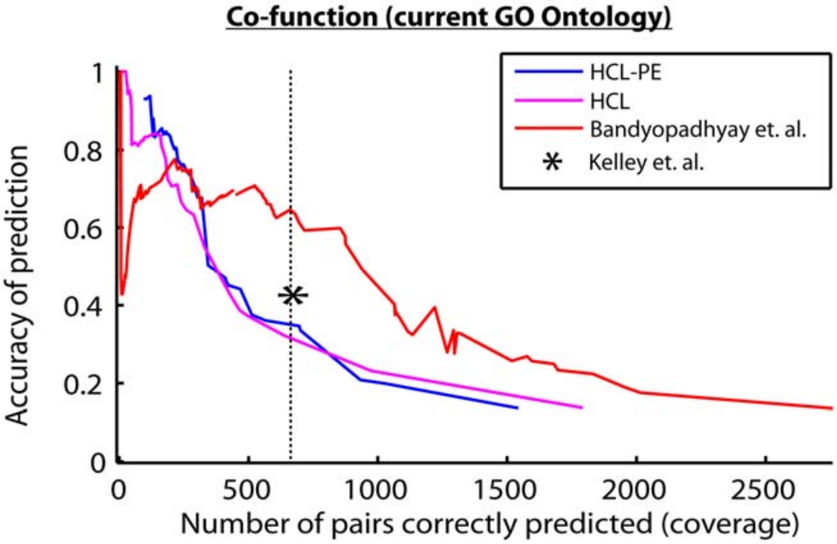
The size and number of final modules was varied by altering the α parameter (see above). To assess performance at low coverage we ran the method with no reward contribution (remove the third term in eq. 4 by setting $\alpha = -\infty$) and plotted the performance of the algorithm at each merge step,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

Supplemental Figures



Supplemental Figure 4.1. Addition of congruence as a predictor of pathway membership.

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



Supplemental Figure 4.2. A current version of the Gene Ontology shows similar performance.

The figure is the same as Figure 4.3 using the current version of the Gene Ontology (March 2007).

Acknowledgements

We thank Tomer Shlomi and Ryan Kelley for critical reading of the manuscript, as well as Silpa Suthram for providing interaction data. We gratefully acknowledge the following sources of funding for this project: award R01-GM070743-01 from the NIGMS (T.I.), an Alon fellowship (R.S.), and a Quantitative Systems Biology grant from the NSF (S.B.). T.I. is a fellow of the David and Lucille Packard Foundation.

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

Bandyopadhyay S, Sharan R, Ideker T. Systematic identification of *functional orthologs by protein network comparison. Genome Research 2006*; **16(3):**428-35.

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

Bibliography

1. Avery L, Wasserman S. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet* 1992;8:312-6.

2. Carter GW, Prinz S, Neou C, Shelby JP, Marzolf B, Thorsson V, Galitski T. Prediction of phenotype and gene expression for combinations of mutations. *Mol Syst Biol* 2007;3:96.

3. Hereford LM, Hartwell LH. Sequential gene function in the initiation of Saccharomyces cerevisiae DNA synthesis. *J Mol Biol* 1974;84:445-61.

4. Ooi SL, Shoemaker DD, Boeke JD. DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nat Genet* 2003;35:277-86.

5. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 2001;294:2364-8.

6. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z, Brown GW, Ingles CJ, Emili A, Allis CD, Toczyski DP, Weissman JS, Greenblatt JF, Krogan NJ. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 2007;446:806-10.

7. Collins SR, Schuldiner M, Krogan NJ, Weissman JS. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol* 2006;7:R63.

8. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 2005;123:507-19.

9. Drees BL, Thorsson V, Carter GW, Rives AW, Raymond MZ, Avila-Campillo I, Shannon P, Galitski T. Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol* 2005;6:R38.

10. St Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* 2007;39:199-206.

11. Segre D, Deluna A, Church GM, Kishony R. Modular epistasis in yeast metabolism. *Nat Genet* 2005;37:77-83.

12. Beyer A, Bandyopadhyay S, Ideker T. Integrating physical and genetic maps: from genomes to interaction networks. *Nat Rev Genet* 2007;8:699-710.

13. Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 2005;23:561-6.

14. Ulitsky I, Shamir R. Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. *Mol Syst Biol* 2007;3:104.

15. Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C, Roth FP. Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. *J Biol* 2005;4:6.

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

17. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* 2007;6:439-50.

18. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res* 2006;34:D436-41.

19. Boone C, Bussey H, Andrews BJ. Exploring genetic interactions and networks with yeast. *Nat Rev Genet* 2007;8:437-49.

20. Driscoll R, Hudson A, Jackson SP. Yeast Rtt109 promotes genome stability by acetylating histone H3 on lysine 56. *Science* 2007;315:649-52.

21. Han J, Zhou H, Horazdovsky B, Zhang K, Xu RM, Zhang Z. Rtt109 acetylates histone H3 lysine 56 and functions in DNA replication. *Science* 2007;315:653-5.

22. Otero G, Fellows J, Li Y, de Bizemont T, Dirac AM, Gustafsson CM, Erdjument-Bromage H, Tempst P, Svejstrup JQ. Elongator, a multisubunit component of a novel RNA polymerase II holoenzyme for transcriptional elongation. *Mol Cell* 1999;3:109-18.

23. Winkler GS, Kristjuhan A, Erdjument-Bromage H, Tempst P, Svejstrup JQ. Elongator is a histone H3 and H4 acetyltransferase important for normal histone acetylation levels in vivo. *Proc Natl Acad Sci U S A* 2002;99:3517-22.

24. Mitchell P, Petfalski E, Shevchenko A, Mann M, Tollervey D. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. *Cell* 1997;91:457-66.

25. Hwang WW, Venkatasubrahmanyam S, Ianculescu AG, Tong A, Boone C, Madhani HD. A conserved RING finger protein required for histone H2B monoubiquitination and cell size control. *Mol Cell* 2003;11:261-6.

26. Wood A, Krogan NJ, Dover J, Schneider J, Heidt J, Boateng MA, Dean K, Golshani A, Zhang Y, Greenblatt JF, Johnston M, Shilatifard A. Bre1, an E3 ubiquitin ligase required for recruitment and substrate selection of Rad6 at a promoter. *Mol Cell* 2003;11:267-74.

27. Kobor MS, Venkatasubrahmanyam S, Meneghini MD, Gin JW, Jennings JL, Link AJ, Madhani HD, Rine J. A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. *PLoS Biol* 2004;2:E131.

28. Krogan NJ, Dover J, Wood A, Schneider J, Heidt J, Boateng MA, Dean K, Ryan OW, Golshani A, Johnston M, Greenblatt JF, Shilatifard A. The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. *Mol Cell* 2003;11:721-9.

29. Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, Wu C. ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* 2004;303:343-8.

30. Li B, Carey M, Workman JL. The role of chromatin during transcription. *Cell* 2007;128:707-19.

31. Dover J, Schneider J, Tawiah-Boateng MA, Wood A, Dean K, Johnston M, Shilatifard A. Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6. *J Biol Chem* 2002;277:28368-71.

32. Sun ZW, Allis CD. Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature* 2002;418:104-8.

33. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 2006;24:1429-35.

34. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;431:99-104.

35. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJ, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M. Global analysis of protein phosphorylation in yeast. *Nature* 2005;438:679-84.

36. Brem RB, Storey JD, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 2005;436:701-3.

37. Sokal RR, Michener C. D. A statistical method for evaluating systematic relationships. *University of Kansas Sci. Bull.* 1958;28:1409-1438.

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

39. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2007;2:2366-82.

40. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498-504.

41. Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, Maier D, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G, Ball CA. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res* 2007;35:D766-70.

Chapter 5. Functional maps of protein complexes from quantitative genetic interaction data Conservation and Rewiring of Functional Modules Revealed by an Epistasis Map (E-MAP) in Fission Yeast

Abstract

An epistasis map (E-MAP) was constructed in the fission yeast, *Schizosaccharomyces pombe*, by systematically measuring the phenotypes associated with pairs of mutations. This high-density, quantitative genetic interaction map, focused on various aspects of chromosome function, including transcription regulation and DNA repair/replication. The E-MAP uncovered a novel component of the RNAi machinery (*rsh1*) and linked the RNAi pathway to several other biological processes. Comparison of the *S. pombe* E-MAP to an analogous genetic map from the budding yeast revealed that while negative interactions were conserved between genes involved in similar biological processes, positive interactions and overall genetic profiles between pairs of genes coding for physically associated proteins were even more conserved. Hence, conservation occurs at the level of the functional module (i.e. protein complex), but the genetic cross-talk between modules can differ significantly.

Introduction

Genetic interactions report on the extent to which the function of one gene depends on the presence of a second. This phenomenon, known as epistasis, can be used for defining functional relationships between genes and the pathways in which the corresponding proteins function. Two main categories of genetic interactions exist: negative (e.g. synthetic sickness / lethality) and positive (e.g. suppression). We have developed a quantitative approach, termed E-MAP, allowing us to measure the whole spectrum of genetic interactions, both positive and negative[1,2]. In budding yeast, *S. cerevisiae*, it has been demonstrated that positive genetic interactions can identify pairs of genes whose products are physically associated and/or function in the same pathway [1,2] while negative interactions exist between genes acting on parallel pathways[4],[5].

We developed the Pombe Epistasis Mapper (PEM) approach[6] that allows high-throughput generation of double mutants in the fission yeast, *S. pombe*. Fission yeast is more similar to metazoans than is *S. cerevisiae*, in its large complex centromere structure, the restriction of spindle construction to mitotic entry, gene regulation by histone methylation and chromodomain heterochromatin proteins, gene and transposon regulation by the RNAi pathway, and the widespread presence of introns in genes. To further study these processes and to try to understand how genetic interaction networks have evolved[7], we generated an E-MAP in *S. pombe* that focuses on nuclear function, designed to be analogous to one we created in budding yeast[2].

Results and Discussion

An E-MAP in *S. pombe*

Using our PEM system[6], we generated a quantitative genetic interaction map in *S. pombe*, comprising approximately 118,000 distinct double mutant combinations among 550 genes involved in various aspects of chromosome function (Figure 5.1a, Tables S1, S4). The genes on the map were chosen based on a previous budding yeast E-MAP [1,2] and also included factors present in human (but not in *S. cerevisiae*), including the RNAi machinery. Colony size measured from high-density arrays was used as a quantitative phenotypic read-out to compute a genetic interaction score (S-score) and previously-described quality control measures were utilized to ensure a high quality dataset[8] (Fig. S1A).

We have previously observed two prominent general trends between genetic interactions and protein-protein interactions: a propensity for positive genetic interactions and strong correlations of genetic interaction profiles between genes coding for proteins participating in protein-protein interactions[8]. Using a high-confidence set of 151 protein-protein interaction pairs from *S. pombe*[9] (Table S2), we observed the same trends in this organism (Figure 5.1B, C). Thus, it appears these relationships are evolutionarily conserved and may represent a general feature of biological networks.

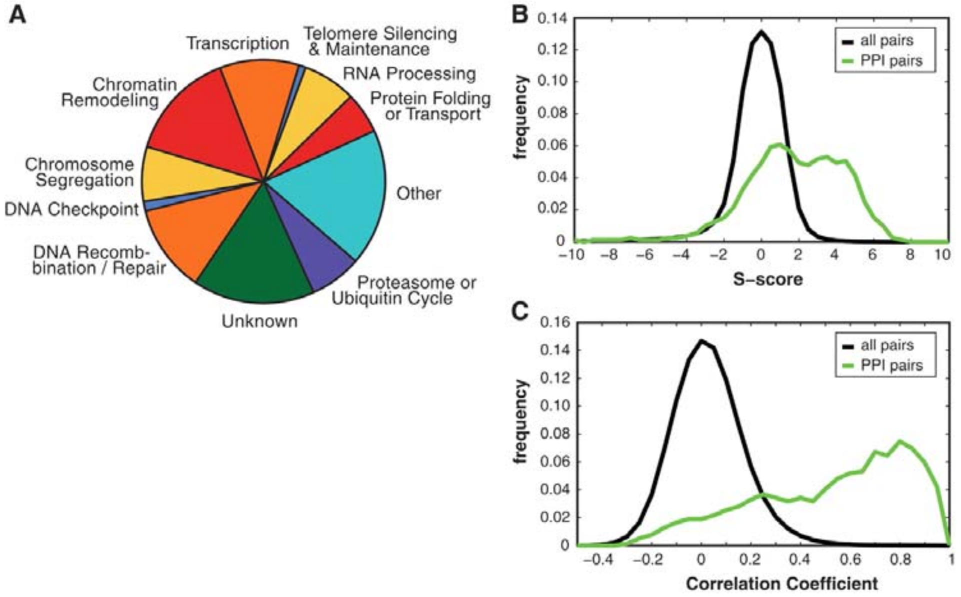


Figure 5.1: Data set overview.

(**A**) Functional classification of the genes contained within the *S. pombe* E-MAP. The map contains 550 genes that were classified into 11 functional categories (table S4). (**B**) Distribution of interaction scores for pairs of genes corresponding to physically interacting proteins (green) and noninteracting proteins (black). (**C**) Distribution of Pearson correlation coefficients of the genetic interaction profiles for the same set of genes used in (B). For a complete list of PPIs used in this analysis, see table S2.

Exploring Nuclear Function in Fission Yeast

A highly structured representation of the genetic map was generated by subjecting the data to hierarchical clustering (Figure 5.2). By scrutinizing several interaction-rich regions, we were able to recapitulate known and identify novel functional relationships.

Genes required for DNA repair / recombination and various checkpoint functions form clusters enriched in negative interactions (Figure 5.2, region 1). The *rad9-hus1-rad1* (9-1-1) checkpoint complex[10] clusters together with rad17 (the homolog of budding yeast *RAD24*) which loads it onto DNA[11]. We find two genes linked to tRNA biogenesis, *sen1* and *trm1*, within the DNA repair cluster. tRNA regulation has been linked to the DNA damage response pathway in *S. cereviaise*[12], and these genetic patterns suggest a similar mechanism may exist in fission yeast. To genetically interrogate the function of essential genes, we used the DAmP strategy for generating hypomorphic alleles[1] (Table S1) and found that the DAmP allele of *mcl1*, involved in DNA replication control and repair, is highly correlated with components of the replication checkpoint (*mrc1* and *csm3*).

The fission yeast homologs of the components of the SWR complex (SWR-C), which in budding yeast incorporates the histone H2A variant Htz1 (Pht1 in fission yeast) into chromatin[13-15], form a highly correlated group (Figure 5.2, region 2). A jumonji domain containing protein, Msc1, whose *S. cerevisiae* ortholog *ECM5* is not part of the budding yeast's SWR-C, is found within the fission yeast SWR-C cluster, consistent with the demonstration that Msc1 acts through Pht1 to promote chromosome stability[16].

The E-MAP reveals functional specialization of the fission yeast Set1 histone H3 lysine 4 methyltransferase complex (SET1-C, COMPASS)[17-19] [20]. In *S. pombe*, five of its subunits (core SET1-C: *set1*, *spp1*, *swd1*, *swd21*, *swd3*) are indispensable for H3-K4 methylation [19] and form a highly correlated cluster on the E-MAP (Figure 5.2, region 3). In budding yeast, another component of COMPASS, Swd2, is essential and part of two distinct complexes: SET1-C and the CPF (Cleavage and Polyadenylation Factor)[17 21]. *S. pombe* contains two non-essential paralogs of *SWD2* (*swd21* and *swd22*), previously shown to act independently in the *S. pombe* SE T1-C and CPF, respectively[22]. Consistent with this, on our map, *swd21* is part of the core SET1-C while *swd22* is strongly correlated with the *SSU72* ortholog, a part of CPF[21,23] (Figure 5.2, region 3). The Ash2-Sdc1 heterodimer within the SET1-C also behaves differently. In *S. cerevisiae*, its orthologous pair (Bre2p-Sdc1p) is exclusively found in the SET1-C[17], while in fission yeast it is shared between the SET1-C and LID2-C[19]. Consistent with this, the dimer does not cluster next to core Set1-C, which is what is observed in budding yeast[2], but is more similar to *snt2*, a member of LID2-C (Figure 5.2, region 3).

Genetic Dissection of the RNAi pathway

The RNAi pathway in S. pombe comprises several components, including CLR4-C, RDR-C, RITS, dicer (Dcr1) and the HP1 homolog, Swi6[24]. All known components of the RNAi machinery that were analyzed cluster next to each other and primarily display positive genetic interaction with one another (Figure 5.3A). Within this cluster are subclusters corresponding to the different protein complexes. Consistent with previous reports, we find positive genetic interactions between the RNAi machinery and epe1, an anti-silencing factor involved in the transcription of heterochromatic regions by RNAPII[25] and required for RNAi-mediated heterochromatin assembly[24]. Conversely, we find negative interactions between RNAi components, involved in posttranscriptional silencing (PTGS), and factors implicated in transcriptional silencing (TGS) of repeat sequences and other loci. In particular, clr3, a histone deacetylase and catalytic subunit of the SHREC complex[26], involved in TGS at centromeric repeats[24] and Tf2 retrotransposons[27,28], shows negative interactions with RNAi components (Figure 5.3A).

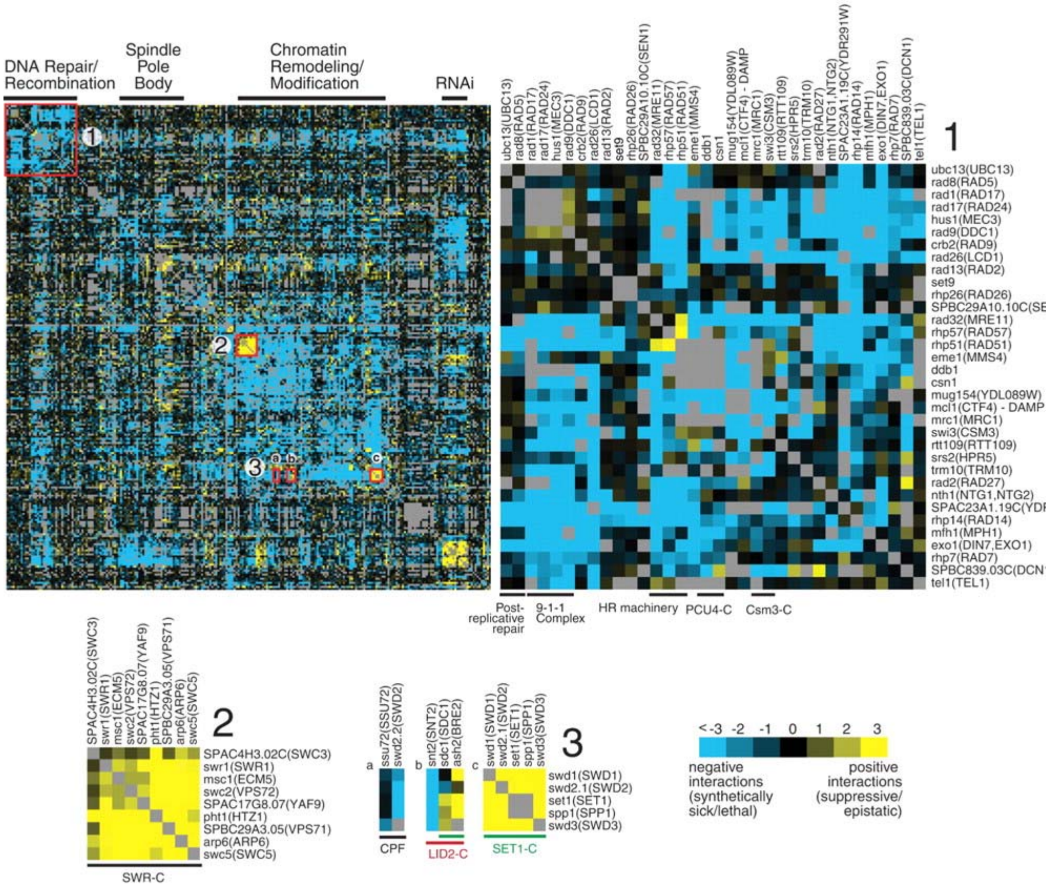


Figure 5.2. The *S. pombe* chromosome function E-MAP.

A section of the E-MAP with specific regions of interest annotated. Further highlighted are the factors involved in DNA repair/recombination (1), as well as two complexes contained within the chromatin remodeling/modification region: the SWR-C chromatin remodeling complex (2) and the Set1, Lid2, and CPF complexes (3). The names of the budding yeast orthologs are shown in parentheses (table S3). The final data set consists of 118,575 measurements and contains 5772 negative (S score ≤–2.5) and 1812 positive (S score ≥2) interactions.

Within the RNAi cluster, we also found a previously unknown component of the RNAi pathway, SPCC1393.05, which we named rsh1 (involved in RNAi silencing and heterochromatin formation) (Figure 5.3A). The gene encodes a 110 kDa protein with no obvious homologs or apparent sequence motifs. Chromatin immunoprecipitation determined that Rsh1 is localized to heterochromatic centromeic regions and its absence causes a significant reduction of silencing at these loci and loss of siRNAs expressed from the centromeric dg/dh repeats (Figure 5.3B-F). Additionally, rsh1Δ leads to a marked reduction of H3-K9 di-methylation and Swi6/HP1 binding

that correlates with lowered levels (>6 fold decrease) of Ago1, component of RITS, recruitment to the outer (otr) centromeric repeat region (Figure 5.3G, H).

We also observe positive interactions between the RNAi machinery and homologs of factors involved in the transition between transcriptional initiation and elongation, including rpb9 and iwr1, components of RNA polymerase II[21,29], and the Mediator complex (pmc2, rox3, pmc5, med2)[30,31]. Indeed, deletions of rpb9, rox3, pmc5 or pmc2 lead to moderate loss of silencing at the centromere (Figure 5.3I, J).

Numerous negative genetic interactions between the RNAi machinery and other cellular complexes and processes were observed (Figure 5.3A) including the spindle-checkpoint pathway (mad1, mad2, bub3, alp14)[32], components of the DASH complex[33] (dad1, dad2, ask1, spc34), and mal3, tub1 and alp31 involved in microtubule stability[34], consistent with the involvement of RNAi / heterochromatin apparatus in proper chromosome segregation[34]. The acetyltrasferase complex, Elongator[35], interacts negatively with the RNAi machinery and clusters next to factors regulating spindle function consistent with the observation that Elongator may be responsible for tubulin acetylation, required for microtubule-based protein trafficking[36]. Finally, components of the DNA repair, checkpoint and recombination apparatus display negative genetic interactions with the RNAi machinery, suggesting the RNAi pathway is also involved in maintaining genomic stability.

Conservation of modular organization of genetic interaction networks

The large evolutionary distance between *S. cerevisiae* and *S. pombe* (ca. 400 million years[37]) allowed us to study the evolution of genetic interactomes. We directly compared the data from this *S. pombe* E-MAP to an analogous dataset from *S. cerevisiae*[2]. The overlap of one to one annotated orthologs[38] between the two E-MAPs encompasses 239 genes (Table S3). First, we analyzed individual negative pair-wise interactions in the two organisms. Recently, it has been suggested[7] that negative interactions between yeast and *C. elegans* were not conserved. Although not strong, we did find a conservation of negative interactions (17.3% for S-score ≤ -2.5), which became more pronounced (33%) when the analysis was restricted to genes that shared the same functional annotations (Figure 5.4A, S2B). To confirm this observation we used an independent dataset from BioGRID[9] and observed similar conservation rates (18% for all and 31% among functionally related genes). Part of the discrepancy seen in *C. elegans* could be due to functional redundancy, multicellularity, or incomplete knock-downs by RNAi. Furthermore, this comparison was not restricted to functionally related genes[7]. In our analysis, we also found a very strong conservation (> 50%) of positive interactions (S-score ≥ 2.0) (which were not considered by[7]) between pairs of genes whose corresponding proteins are physically associated (Figure 5.4A, Fig. S2A-D).

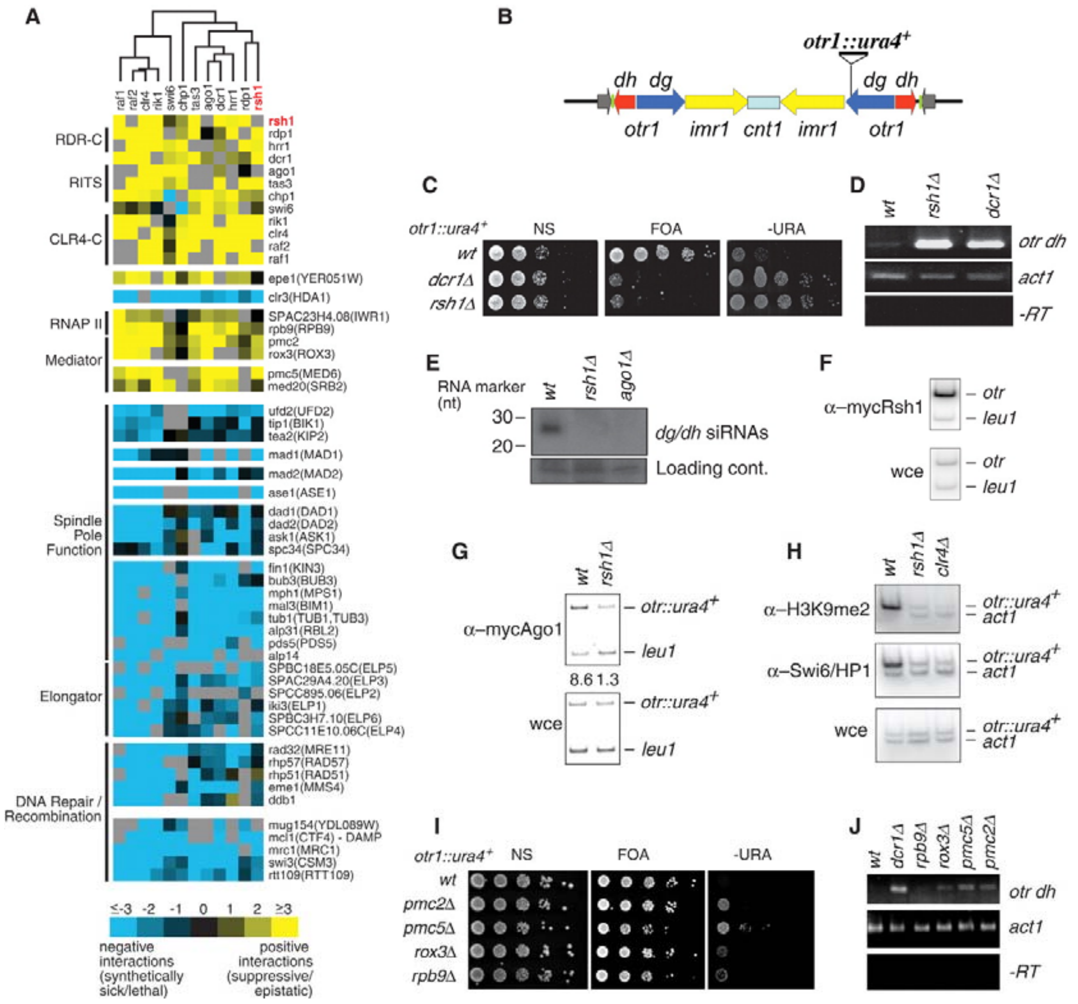


Figure 5.3: Characterization of genes involved in the RNAi pathway.

(**A**) Genetic profiles for genes involved in RNAi with individual protein complexes or processes annotated. (**B**) Schematic of the centromeric region of chromosome 1 with the position of the *ura4*$^+$ reporter gene within the *otr1* region. (**C**) Loss of Rsh1p abolishes heterochromatic silencing of the *ura4*$^+$ reporter gene inserted at the outer repeat region of centromere 1 (otr1*::ura4*$^+$). NS, nonselective; FOA, counterselective; -URA, uracil-deficient media. (**D**) Levels of *dh* transcripts analyzed by reverse transcription polymerase chain reaction (RT-PCR) using RNA prepared from indicated strains. (**E**) Loss of siRNAs derived from *dg/dh* repeats in *rsh1* **Δ** detected by Northern blotting. nt, nucleotides. (**F**) Rsh1 localizes to outer (*otr*) centromeric repeats. An epitope-tagged version of Rsh1 (mycRsh1) was used to perform chromatin immunoprecipitation (ChIP). wce, whole-cell extract. (**G**) Rsh1 is required for localization of Ago1. Localization of mycAgo1 at *otr1::ura4*$^+$ in wild-type and *rsh1* **Δ** cells was assayed using ChIP. leu1 is an internal loading control for ChIP experiments. (**H**) Effect of *rsh1* **Δ** on heterochromatin assembly at centromeric repeats. Levels of histone H3 lysine 9 dimethylation (H3K9me2) and Swi6/HP1 at *otr1::ura4*$^+$ were assayed using ChIPs. (**I** and **J**) Loss of Mediator and RNAPII subunits affects centromeric silencing. The levels of transcripts corresponding to *dh* centromeric repeats were analyzed by RT-PCR. leu1 and act1 are used as internal loading controls.

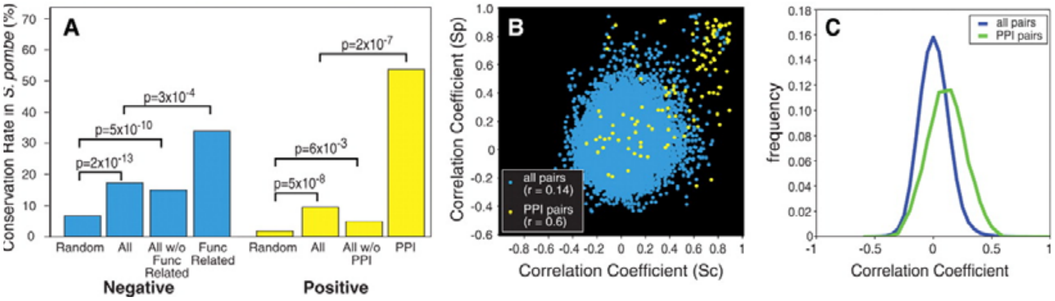


Figure 5.4: Modular conservation of genetic interaction patterns.

A set of 239 one to one orthologs was used for the analysis. (**A**) Conservation of positive and negative genetic interactions based on comparison with *S. cerevisiae*. Conservation rates are higher for the subset of negative interactions between genes with the same functional annotation and the subset of positive interactions corresponding to known protein-protein interactions in S. *cerevisiae.* P-values were determined using a two-sided Student's t-test. (**B**) Scatter plot of Pearson correlation coefficients of genetic interaction profiles. (**C**) Distribution of the cross-species Pearson correlation coefficient of genetic profiles.

The set of genetic interactions for a given gene provides a sensitive phenotypic signature or profile. Although global comparison of all correlations of genetic profiles between orthologous pairs in each species (Table S3) revealed a weak overall conservation (r=0.14) (Figure 5.4B), pairs corresponding to PPIs were much more highly correlated (r=0.60) (Figure 5.4B). An aggregate measure for the likelihood of two proteins to carry out a common function, many of which correspond to PPI pairs, is the COP score[8], which integrates the individual genetic interaction score and correlation coefficient of genetic interaction profiles. Pairs of genes displaying high COP scores in both organisms almost exclusively correspond to PPIs (Fig. S2E).

To further explore the extent of conservation of genetic networks, the profiles of each of the 239 orthologs in both species were compared to all profiles from the other organism (Fig. 4C). We found some conservation between direct orthologs ($p=8x10^{-20}$)

suggesting that genetic interaction profiles of orthologs across species tend to be similar (Fig. S2F). There is, however, a stronger conservation of genetic profiles between a gene and the ortholog of its interacting partner when only co-complex members were considered (Figure 5.4C) ($p=9x10^{-51}$). Thus, genetic profiles of members of protein-protein interaction pairs tend to correlate better not only to their interaction partners within the same species but also to the orthologs of their interaction partner in an evolutionary distant organism.

Collectively, these data demonstrate that genetic interactions between specific subsets of genes are conserved between *S. cerevisiae* and *S. pombe*. Specifically, we find conservation of negative interactions when genes involved in the same cellular process are considered. Better conserved are positive interactions and genetic profiles of genes whose products are physically associated. Therefore, we argue that conservation primarily exists at the level of the functional module (i.e. protein complex), and perhaps protein-protein interactions pose a constraint on functional divergence in evolution.

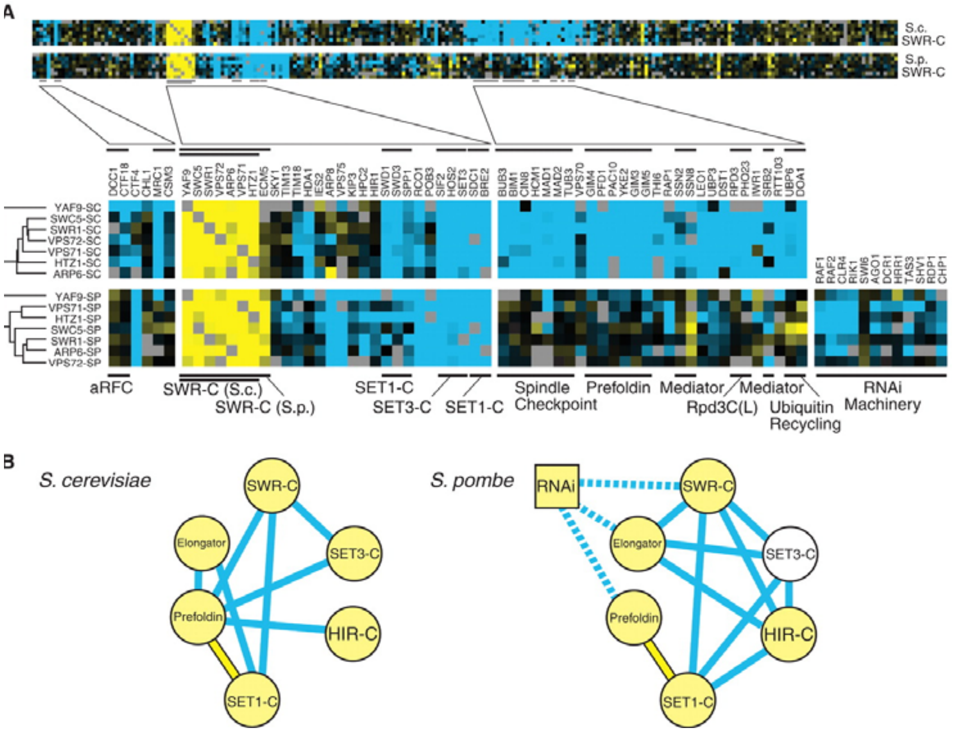


Figure 5.5: Re-wiring of the conserved functional modules

(A) Comparison of genetic interaction profiles of the SWR-C in *S. cerevisiae* and *S. pombe*. Analogous sets of genetic interactions from the two organisms are shown (**Dataset S2**). (**B)** Genetic cross-talk between functional modules. Modules are represented as circles or boxes (in yellow if the interactions within the module are primarily positive). Negative and positive interactions between modules are represented as blue and yellow lines, respectively. The diagram was generated using the method described in [3].

Re-Wiring of Conserved Functional Modules

Biological modules can be defined as highly interconnected groups of physically or functionally associated factors and often correspond to protein complexes. In addition to identifying functional modules, high-density genetic interaction data reports on the functional relationships between modules (i.e. the wiring of the network).

To compare the genetic cross-talk between modules in the two organisms, we merged and clustered the genetic interaction matrix of *S. pombe* with that of *S. cerevisiae* for the 239 1:1 orthologs (Dataset S2). Inspection of this dataset revealed a partial overlap of negative interactions between protein complexes (Figure 5.5A). For example, in both organisms SWR-C display negative genetic interactions with the SET1-C and the histone deacetylatase (HDAC) complex, SET3-C. However, substantial differences were found as well. For instance, in only budding yeast are there negative interactions between SWR-C and components of the spindle checkpoint, the chaperone complex Prefoldin, the HDAC complex, Rpd3C(L) and Mediator (Figure 5.5A).

Several possible explanations can be offered. First, the additional subunit unique to the fission yeast SWR-C, Msc1, may alter the function of the complex. Also, species-specific post-translational modifications may result in different genetic behavior. Msc1 has been shown to harbor ubiquitin ligase activity[39] and may be involved in ubiquitinating proteins related to the function of SWR-C. Another reason could be to the presence or absence of particular cellular machinery. For example, the re-wiring of the genetic space surrounding the SWR-C in *S. pombe* may be due to the presence of the RNAi machinery, which shows negative interactions with the complex (Figure 5.5B). Consequently, dramatic alterations in the network topology of budding yeast may have been necessary to compensate for the absence of the RNAi pathway. We cannot rule out the possibility that many of the interactions do exist under different environmental conditions. Nonetheless, a significant re-wiring of other complexes (e.g. the HIR chromatin assembly complex and Prefoldin, Fig. S3) was also observed under the conditions used.

The modularity of biological networks is believed to be one of the main contributors to their robustness, as it implies enhanced functional flexibility. Much like an electronic circuit, such modular architecture allows different tasks to be accomplished with the same minimal set of components by changing the wiring (or flow of information) between them. Re-wiring due to addition or removal of modules allows for economical design of sophisticated networks that are able to adapt to different conditions and environmental niches at low cost. We observe this behavior derived from high-density genetic interaction data from two evolutionary distant species. Our data strongly support the idea that functional modules are highly conserved, but the wiring between them can differ significantly. Thus, using model systems to make inferences about biological network topology may be more successful for describing modules than for describing the cross-talk between them.

Acknowledgements

We thank P. Beltrao and G. Cagney for critical reading of the manuscript, M. Wiren and S. Forsburg for discussion, P. Kemmeren for setting up the web database, S. Wang, C. Wen, D. Avdic for technical help and F. Stewart for sharing unpublished data. This work was supported by NIH (NIGMS grant# GM084279 (T.I. and N. J.K.)), the Sandler Family Foundation (N.J.K), the Howard Hughes Medical Institute (J.S.W.), National Cancer Institute (S.I.G.), Center for Cancer Research (S.I.G.) and the California Institute of Quantitative Biology (QB3) (N.J.K.).

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

Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins SR, Qu H, Shales M, Park H, Hayles J, Hoe K, Kim D, Ideker T, Grewal SI, Weissman JS, Krogan NJ. *Conservation and Rewiring of Functional Modules Revealed by an Epistasis Map in Fission Yeast.* **Science** 2008;322(5900):405-10.

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

Bibliography

1. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 2005;123:507-19.

2. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z, Brown GW, Ingles CJ, Emili A, Allis CD, Toczyski DP, Weissman JS, Greenblatt JF, Krogan NJ. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 2007;446:806-10.

3. Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T. Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol* 2008;4:e1000065.

4. Pan X, Yuan DS, Ooi SL, Wang X, Sookhai-Mahadeo S, Meluh P, Boeke JD. dSLAM analysis of genome-wide genetic interactions in Saccharomyces cerevisiae. *Methods* 2007;41:206-21.

5. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Global mapping of the yeast genetic interaction network. *Science* 2004;303:808-13.

6. Roguev A, Wiren M, Weissman JS, Krogan NJ. High-throughput genetic interaction mapping in the fission yeast Schizosaccharomyces pombe. *Nat Methods* 2007;4:861-6.

7. Tischler J, Lehner B, Fraser AG. Evolutionary plasticity of genetic interaction networks. *Nat Genet* 2008;40:390-1.

8. Collins SR, Schuldiner M, Krogan NJ, Weissman JS. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol* 2006;7:R63.

9. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;34:D535-9.

10. Kaur R, Kostrub CF, Enoch T. Structure-function analysis of fission yeast Hus1-Rad1-Rad9 checkpoint complex. *Mol Biol Cell* 2001;12:3744-58.

11. Majka J, Burgers PM. Yeast Rad17/Mec3/Ddc1: a sliding clamp for the DNA damage checkpoint. *Proc Natl Acad Sci U S A* 2003;100:2249-54.

12. Ghavidel A, Kislinger T, Pogoutse O, Sopko R, Jurisica I, Emili A. Impaired tRNA nuclear export links DNA damage and cell-cycle checkpoint. *Cell* 2007;131:915-26.

13. Krogan NJ, Keogh MC, Datta N, Sawa C, Ryan OW, Ding H, Haw RA, Pootoolal J, Tong A, Canadien V, Richards DP, Wu X, Emili A, Hughes TR, Buratowski S, Greenblatt JF. A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol Cell* 2003;12:1565-76.

14. Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, Wu C. ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* 2004;303:343-8.

15. Kobor MS, Venkatasubrahmanyam S, Meneghini MD, Gin JW, Jennings JL, Link AJ, Madhani HD, Rine J. A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. *PLoS Biol* 2004;2:E131.

16. Ahmed S, Dul B, Qiu X, Walworth NC. Msc1 acts through histone H2A.Z to promote chromosome stability in Schizosaccharomyces pombe. *Genetics* 2007;177:1487-97.

17. Roguev A, Schaft D, Shevchenko A, Pijnappel WW, Wilm M, Aasland R, Stewart AF. The Saccharomyces cerevisiae Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *Embo J* 2001;20:7137-48.

18. Krogan NJ, Dover J, Khorrami S, Greenblatt JF, Schneider J, Johnston M, Shilatifard A. COMPASS, a histone H3 (Lysine 4) methyltransferase required for telomeric silencing of gene expression. *J Biol Chem* 2002;277:10753-5.

19. Roguev A, Schaft D, Shevchenko A, Aasland R, Shevchenko A, Stewart AF. High conservation of the Set1/Rad6 axis of histone 3 lysine 4 methylation in budding and fission yeasts. *J Biol Chem* 2003;278:8487-93.

20. Nagy PL, Griesenbeck J, Kornberg RD, Cleary ML. A trithorax-group complex purified from Saccharomyces cerevisiae is required for methylation of histone H3. *Proc Natl Acad Sci U S A* 2002;99:90-4.

21. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;415:141-7.

22. Roguev A, Shevchenko A, Schaft D, Thomas H, Stewart AF, Shevchenko A. A comparative analysis of an orthologous proteomic environment in the yeasts Saccharomyces cerevisiae and Schizosaccharomyces pombe. *Mol Cell Proteomics* 2004;3:125-32.

23. Dichtl B, Blank D, Ohnacker M, Friedlein A, Roeder D, Langen H, Keller W. A role for SSU72 in balancing RNA polymerase II transcription elongation and termination. *Mol Cell* 2002;10:1139-50.

24. Grewal SI, Jia S. Heterochromatin revisited. *Nat Rev Genet* 2007;8:35-46.

25. Zofall M, Grewal SI. Swi6/HP1 recruits a JmjC domain protein to facilitate transcription of heterochromatic repeats. *Mol Cell* 2006;22:681-92.

26. Sugiyama T, Cam HP, Sugiyama R, Noma K, Zofall M, Kobayashi R, Grewal SI. SHREC, an effector complex for heterochromatic transcriptional silencing. *Cell* 2007;128:491-504.

27. Hansen KR, Burns G, Mata J, Volpe TA, Martienssen RA, Bahler J, Thon G. Global effects on gene expression in fission yeast by silencing and RNA interference machineries. *Mol Cell Biol* 2005;25:590-601.

28. Cam HP, Noma K, Ebina H, Levin HL, Grewal SI. Host genome surveillance for retrotransposons by transposon-derived proteins. *Nature* 2008;451:431-6.

29. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 2006;440:637-43.

30. Spahr H, Beve J, Larsson T, Bergstrom J, Karlsson KA, Gustafsson CM. Purification and characterization of RNA polymerase II holoenzyme from Schizosaccharomyces pombe. *J Biol Chem* 2000;275:1351-6.

31. Sakurai H, Kimura M, Ishihama A. Identification of the gene and the protein of RNA polymerase II subunit 9 (Rpb9) from the fission yeast Schizosacharomyces pombe. *Gene* 1998;221:11-6.

32. Millband DN, Hardwick KG. Fission yeast Mad3p is required for Mad2p to inhibit the anaphase-promoting complex and localizes to kinetochores in a Bub1p-, Bub3p-, and Mph1p-dependent manner. *Mol Cell Biol* 2002;22:2728-42.

33. Liu X, McLeod I, Anderson S, Yates JR, 3rd, He X. Molecular analysis of kinetochore architecture in fission yeast. *Embo J* 2005;24:2919-30.

34. Asakawa K, Kume K, Kanai M, Goshima T, Miyahara K, Dhut S, Tee WW, Hirata D, Toda T. The V260I mutation in fission yeast alpha-tubulin Atb2 affects microtubule dynamics and EB1-Mal3 localization and activates the Bub1 branch of the spindle checkpoint. *Mol Biol Cell* 2006;17:1421-35.

35. Otero G, Fellows J, Li Y, de Bizemont T, Dirac AM, Gustafsson CM, Erdjument-Bromage H, Tempst P, Svejstrup JQ. Elongator, a multisubunit component of a novel RNA polymerase II holoenzyme for transcriptional elongation. *Mol Cell* 1999;3:109-18.

36. Gardiner J, Barton D, Marc J, Overall R. Potential role of tubulin acetylation and microtubule-based protein trafficking in familial dysautonomia. *Traffic* 2007;8:1145-9.

37. Sipiczki M. Where does fission yeast sit on the tree of life? *Genome Biol* 2000;1:REVIEWS1011.

38. Penkett CJ, Morris JA, Wood V, Bahler J. YOGY: a web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms. *Nucleic Acids Res* 2006;34:W330-4.

39. Dul BE, Walworth NC. The plant homeodomain fingers of fission yeast Msc1 exhibit E3 ubiquitin ligase activity. *J Biol Chem* 2007;282:18397-406.

Chapter 6. DNA-damage induced rewiring of protein signaling revealed by a conditional epistatic interaction map (cE-MAP)

Abstract

Damage to DNA triggers major cellular responses including cell-cycle arrest, chromatin remodeling, and DNA repair. However, how signaling pathways orchestrate this response remains unclear. To uncover these pathways, we developed a conditional epistasis mapping approach, termed cE-MAP, which we use to examine double gene knockouts among a set of 418 signaling genes in yeast including most kinases, phosphatases, and transcription factors with and without the DNA damaging agent, MMS. Analysis of the difference between the two static maps revealed 1,161 conditional interactions which are extremely effective at identifying DNA-damage response genes and pathways. The cE-MAP identifies roles for MAPKs in DNA-damage signaling, for Cbf1 regulation by the DNA-damage phosphatase Pph3, and a functional connection between the checkpoint kinase Mec1 and histone variant Htz1. Thus, cE-MAPS are a valuable tool for mapping pathways that are stimulated under a specific condition and for probing a previously unexplored space of the genetic interactome.

Introduction

Detection and repair of DNA damage is critical for the proper replication and function of every organism. DNA damage is sensed by a highly conserved mechanism involving the two protein kinases: Ataxia-Telangiectasia-Mutated (ATM) and Ataxia-Telangiectasia-and-Rad3-related (ATR) corresponding to yeast Mec1 and Tel1, respectively[2]. These aggregate at DNA lesions and activate signal transduction cascades that include the CHK protein kinases (yeast Chk1, Rad53, and Dun1) which trigger a variety of transcriptional and transcription-independent responses, including activation of DNA repair machinery, cell-cycle arrest, chromatin and cytoskeletal remodeling, RNA and protein turnover, and in some cases apoptosis[4]. However, a global view of the interrelationships among the many response pathways is still lacking and many new processes and pathways involved in the DNA Damage Response (DDR) remain to be identified. Here, we describe a systems approach to mapping DNA-damage signaling pathways based on the generation of a quantitative genetic interaction map in yeast induced by the DNA alkylating agent methyl-methanesulfonate (MMS).

Given such a wide and complex cellular response, a number of genome-scale methodologies have been applied to uncovering the components of DDR pathways. Yeast has been the proving ground for these technologies, which include single gene deletion profiling to identify genes required for response to various damaging agents[3,5,6], expression profiling to identify the transcriptional programs associated with the DDR[7], chIP-chip to identify targets of active transcription factors[8], and phosphoproteomic screening to identify post-translational modifications governed by major checkpoint kinases[9]. Genome-scale technologies have also been applied in humans, including siRNA screening to identify genes whose knockdown confers sensitivity to MMS[10], factors impinging on the activation of γH2AX[11] and phosphoproteomic profiling to identify substrates of ATM and ATR[12]. Despite these technological advances, our understanding of the mechanisms which govern DNA repair processes remains limited.

Genetic interactions report the extent to which two genes have an effect on the same phenotype and can indicate genes functioning in the same or similar pathway[13]. Positive interactions can occur in cases where the double mutant is either healthier (suppressive) or no sicker (epistatic) than the sickest single mutant[14-16]. Such interactions commonly indicate genes functioning in the same pathway or complex[17,18]. Negative genetic interactions (synthetic sick/lethal interactions) indicate genes whose individual mutants are viable whereas their mutation in combination results in a stronger growth defect than expected by either mutation alone. Such interactions are thought to identify pairs of genes which are functionally related but have a parallel or redundant function[18].

Genetic interaction screening has been predominately performed through two approaches, synthetic genetic array (SGA) technology[19] and diploid synthetic lethality analysis on microarrays (dSLAM)[20], the latter has been used to define a genetic interaction network underlying DNA integrity under nominal conditions[21]. Recently, these methods have been complemented by a method termed Epistatic Miniarray Profiles (E-MAP), which is able to quantitatively sample the full spectrum of positive and negative genetic interactions[15,22,23].

Because the pathways which are most essential for growth under DNA damage differ from those that might be required for growth on other conditions, we hypothesized that genetic interaction mapping under an alternative condition might illuminate pathways which are most essential for growth under that condition. In this regard, conditional genetic interactions among genes involved in DNA repair have been shown to exist[24]. However, the widespread prevalence of such interactions and whether such interactions can identify new and novel pathways have yet to be determined. To further elucidate the role of signaling, transcription and DNA maintenance in DDR, we used the E-MAP approach to create all possible pairs of deletion mutations among 418 of these genes and profiled their growth in both rich media as well as in the presence of the DNA alkylating agent, MMS. This systematic approach reveals that tracking the way that these genetic interaction networks change can indicate pathways which are formed dynamically in response to perturbation and are critical for growth, many of which have no known role in DNA repair. The resulting conditional genetic interaction map, derived from the difference of the two static maps, reveals hundreds of compensatory and serial pathways as well as detail single proteins and multi-protein modules which undergo significant rewiring in response to MMS. Our results indicate that conditional genetic interaction maps probe a radically different space of the genetic interactome and provide a new model for the exploration of the mode of action of drugs and cellular stimuli.

Results and Discussion

A conditional epistasis map centered on DNA damage signaling and transcription

We assembled a DNA damage cE-MAP based on a core set of 418 yeast genes, designed to provide near global coverage of the signaling and transcriptional apparatus in the damage response. This core set included 122 kinases, 40 phosphatases, 194 DNA-binding transcription factors representing the vast majority of all genes annotated with these functions. In addition we included 35 and 31 genes involved in chromatin maintenance and DNA repair, to further map functional relationships in relation to these processes (Figure 6.1A, Table S1). Finally, hypomorphic alleles were generated for essential kinases, including *CDC28*, *MEC1* and *RAD53* [16]. To screen for genetic interactions, approximately 80,000 double gene deletion strains were generated representing all pairwise combinations of these core genes. Double mutant combinations were evaluated under two growth conditions: growth in rich media and growth in rich media under exposure to DNA damage by the alkylating agent methane-methanosulfonate (MMS at 0.02%). Analysis of each condition independently revealed two static genetic interaction maps: a network of 367 positive ($S \geq 2$) and 1,562 negative interactions ($S \leq -2.5$) for the untreated map, and 595 positive and 1,744 negative interactions under MMS (Table S2).

Analysis of the untreated interaction map showed strong association with physical interaction networks of various kinds. For example, as has been previously described[22,25], known protein-protein interactions were highly enriched for gene pairs with both positive and negative genetic interactions; and known kinase-substrate and phosphatase-substrate relationships were enriched for positive but not negative interactions[26]. Interestingly, known transcription factor / DNA interactions were enriched for negative but not positive genetic interactions, suggesting significant redundancy in the organization of the transcriptional apparatus (Figure S1).

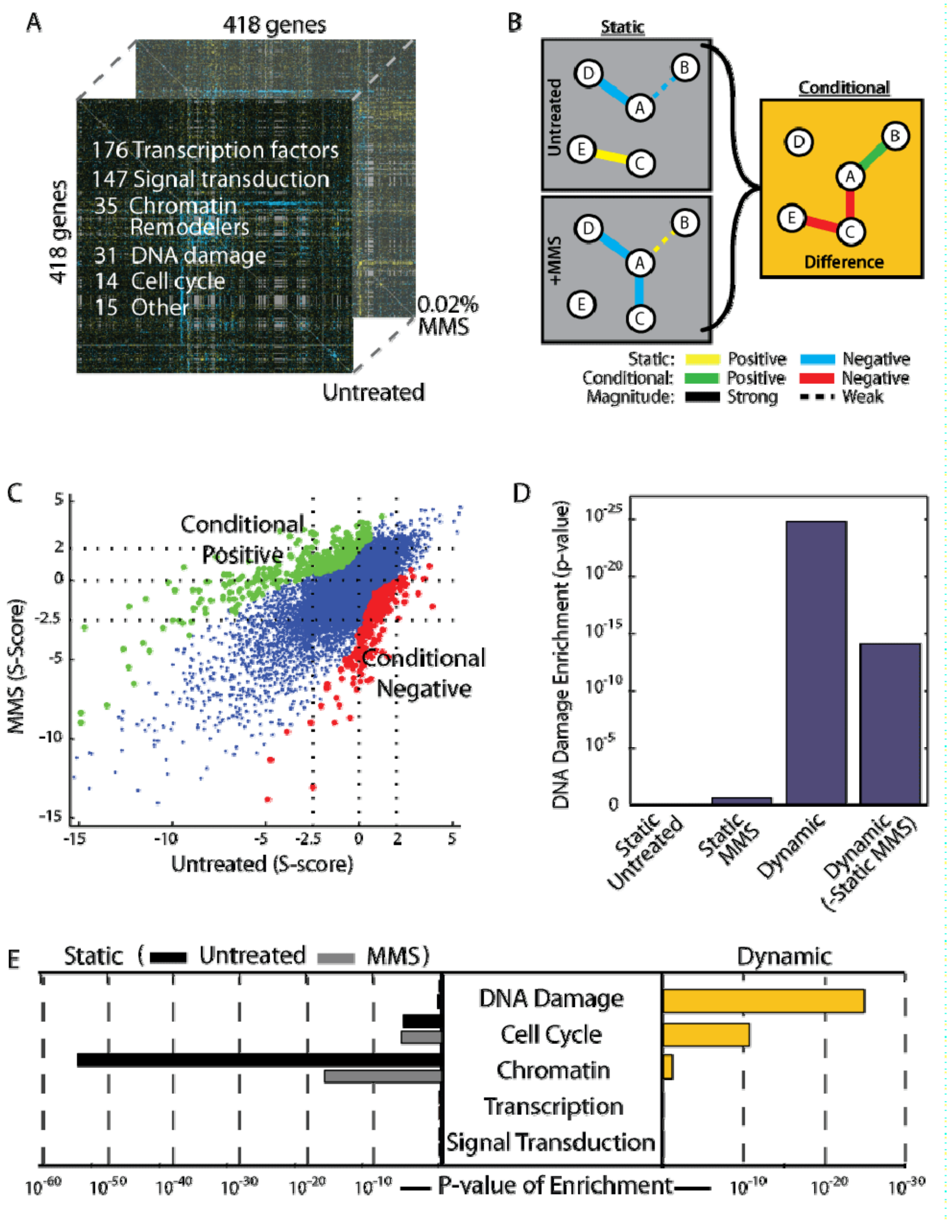


Figure 6.1: Characterization of the DNA damage signaling cE-MAP.

(A) Composition of the genes in the cE-MAP organized by functional categories. (B) Difference between untreated and MMS static maps were determined to identify conditional interactions. (C) Scatter of S-scores between untreated and MMS maps and identification of conditional positive and negative interactions (cS≥3 and cS≤−3, respecitvely). (D) Enrichment of interactions involving genes functioning in DNA damage repair among various static and conditional networks. Static networks consist of positive (S≥2) and negative (S≤−2.5) interactions. For the last bar, all dynamic interactions overlapping with the Static MMS dataset were removed. (E) Enrichment of interactions involving genes in various functions among static and conditinal networks. For each function, enrichment was determined via hypergeometric test based on comparing the proportion of identified versus total possible interactions to the same proportion for all 418 genes in the E-MAP.

We next developed a sensitive metric to detect conditional genetic interactions, i.e., those which are dynamically altered between conditions (Figure 6.1B). For this purpose, we established a null model for the expected difference in genetic interaction score when comparing two independent E-MAP screens performed under identical conditions. Based on departures from this null model, we assigned a *p*-value of significance for the change in interaction between untreated and MMS conditions (Figure 6.1C, Experimental Procedures). This method identified 1,161 conditional interactions ($p < 0.001$) (Table S2). Of these, 522 were 'conditional negative' (cS-score ≤ -3), indicating double mutant hyper-sensitivity to MMS resulting in conditional lethality or sickness. The remaining 639 interactions were 'conditional positive' (cS-score $\geq +3$), indicating cells were less sensitive to MMS than expected from the sensitivity of the single mutants alone.

Remarkably, we found that only 38% of conditional interactions were called positive or negative in either condition individually. Thus, the network of conditional interactions is largely distinct from either of the two static networks from which it is derived. This difference occurs since many genetic interactions are too weak to detect in one condition alone but become very clear considering the change in interaction strength between two conditions (e.g., interactions that are weakly negative in untreated conditions but become weakly positive after MMS exposure).

Next, we investigated which interpretation of genetic interactions, static or conditional, was more effective at identifying components of the DNA damage response. Strikingly, genes involved in the DNA damage response were no more likely than random sets of genes to appear in either the static untreated or the static MMS networks

(Figure 6.1D). In contrast, conditional interactions were highly enriched for DNA damage response genes, even when the interactions held in common with the static MMS map were removed. A more general survey of functional groups indicated that interactions within both of the static networks were highly enriched for genes involved in chromatin organization (Figure 6.1E). This strong genetic signal from chromatin components has been greatly exploited in the past to create rich maps of chromosome function [22,27]. In the conditional network, however, the chromatin contribution has effectively "cancelled out" allowing only differentially-represented pathways to surface such as DNA repair. Beyond DNA repair pathways, the conditional network was highly enriched for genes involved in cell cycle progression, a major component of the DNA damage response that is halted to repair damaged DNA before replication (Figure 6.1E)[4]. Thus, conditional genetic interaction networks explore a fundamentally different landscape of genetic interactions.

Conditional genetic interaction 'hubs' identify mediators of the DNA damage response

'Hubs' in a genetic network are genes with very large numbers of interactions—i.e., for which mutation enhances the phenotypic consequences of mutations in many other diverse pathways. Similarly, in a conditional genetic network, hubs might indicate critical components whose presence is required to modulate many dynamic events in response to the stimulus. In support of this hypothesis, we found that the number of interactions per gene in the DNA damage conditional network was correlated with the severity of growth defect of the gene deletion on MMS (r=0.35, p=10^{-5}, Figure S3)[3]. We also observed that most genes with large numbers of positive or negative conditional interactions had previously defined roles in the DNA damage response (Figure 6.2). However, some genes with many conditional interactions had not been previously linked to DNA damage suggesting novel functions worthy of further investigation. For example, the iron-sensing transcription factor *RCS1* had among the largest number of conditional negative interactions, suggesting a novel role in the transcriptional response to MMS that is corroborated by its reported role in chromosome segregation and stability[28] . Critically, deletion of *RCS1* alone does not confer sensitivity to MMS—hence, the conditional genetic network highlights novel pathway components which would otherwise be missed through more traditional genetic screening techniques.

A global module map of DNA damage signaling

A preliminary examination of the conditional genetic network suggested substantial modular structure, with strong clustering of genetic interactions around functionally-related groups of genes. Three well known DNA-damage response modules are the damage signaling cascade, the alternative Replication Factor C (aRFC) complex required for sister chromatid cohesion, and the RAD52 epistasis group involved in DNA repair via homologous

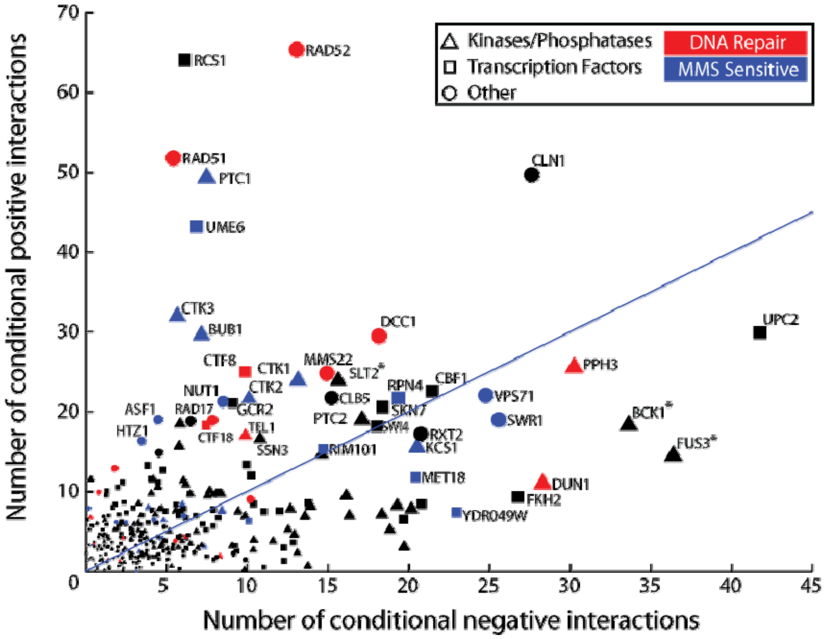


Figure 6.2: Identification of conditional Interaction 'hubs'.

Distribution of the number of conditional positive and negative interactions associated with genes in the cE-MAP. The 30 most sensitive MMS sensitive genes in the cE-MAP are indicated, excluding those which function in DNA repair (**Table S1**) [3]. Starred genes indicate MAPKs. Node positions were perturbed randomly to avoid overlap.

recombination (HR). In untreated conditions, genetic interactions among these three modules were largely unremarkable. However after MMS, we observed a shift to overwhelmingly positive genetic interactions among nearly all components of the three modules (Figure 6.3A), suggesting the alignment of these modules under a common pathway in response to DNA-damage stress. Consistent with these observations, previous reports have linked members of the aRFC with proper activation of the DNA-damage checkpoint kinase Rad53 in MMS[21] as well as demonstrated that the Mec1 kinase regulates the recruitment of Rad52 to sites of DNA damage[29]. One interpretation of these data is that following detection of DNA damage, aRFC establishes a replication block through activation of the checkpoint cascade which ultimately activates a cadre of

DNA repair machinery including the establishment of Rad51/52 at sites of DNA damage (Figure 6.3B).

To form a global picture of the functional modules revealed by the cE-MAP, conditional genetic interactions were integrated with databases of known protein-protein physical interactions, protein complexes, and pathways (Table S3). For this purpose, we employed a previously-published method[30] which identifies 'modules' as clusters of genes defined by both genetic and physical interaction data. We identified pairs of functionally-related modules as those which are interconnected by bundles of many strong conditional genetic interactions (see Experimental Procedures). Using this method we identified 56 multigenic modules and 66 significant module-module genetic interactions conditioned on DNA damage (Figure 6.3C, Table S4). An example of connections between modules includes the previously observed conditional interactions between the damage signaling, RAD52 epistasis group and aRFC modules (Figure 6.3A). Many conditional positive interactions were observed involving the CTK-C complex (Ctk1/2/3), which has been previously shown to phosphorylate RNA polymerase II (Mediator) to regulate transcription[31] and catalyze DNA-damage-induced transcription[32]. The module map suggests that CTK-C plays many additional roles in the DDR together with chromatin regulatory complexes such as SWR-C, RPD3, and RSC, which is consistent with reports that it can regulate the positioning of COMPASS-mediated histone methylation boundaries along genes[33,34]. Conditional negative interactions in the module map also suggest a role for the Elm1/Hsl1 septin checkpoint kinases in regulation of INO80, an ATP-dependent chromatin remodeling complex [35]. Such regulation may regulate bud morphogenesis, potentially through phosphorylation of INO80 subunit

Nhp10 by Hsl1 as has been shown to occur *in vitro*[36]. Critically, permutation analysis revealed that even finding one significant link between two modules rarely happens by chance alone ($p < 0.01$, Figure S4), confirming the modular organization of the network. Thus, at the level of protein modules and complexes, conditional genetic interactions highlight functional connections between modules and complexes which occur in a condition-dependent context.

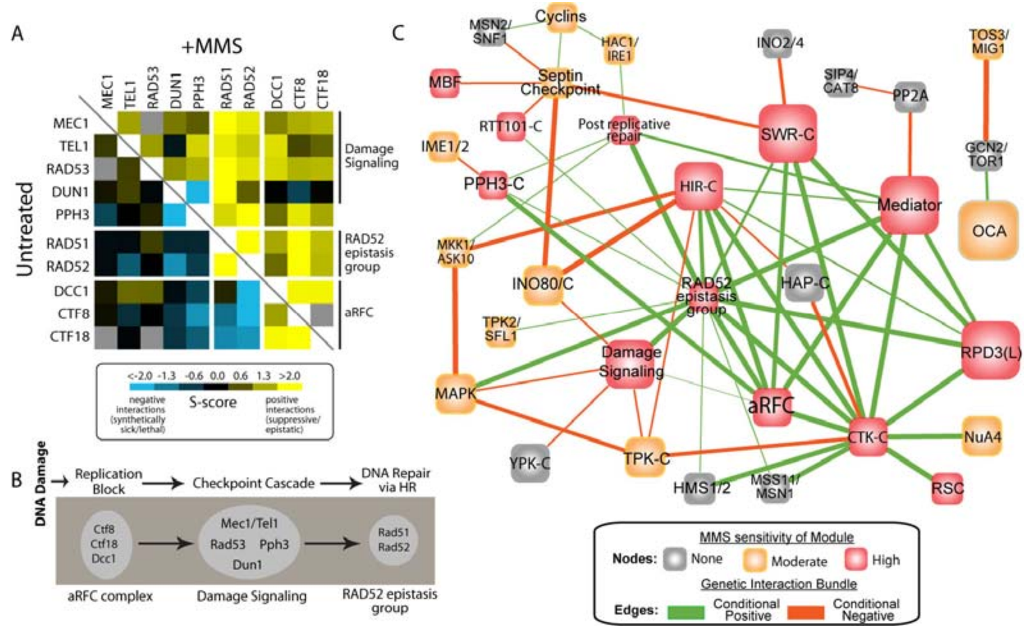


Figure 6.3: Module-based interpretation of the conditional genetic interactions.

(A) Static genetic interactions between module members. (B) Pathway interpretation of conditional positive genetic interactions. (C) Module map of selected protein complexes and pathways connected by bundles of condition-specific genetic interactions. Node colors represent the most severe phenotype among members of a module.

MAPK pathways play a vital role in signaling DNA damage

The Mitogen-Activated Protein Kinases (MAPK) form a series of signal transduction pathways that mediate the response of cells to a variety of extracellular stimuli and stress. We observed three MAPK proteins in particular— Bck1, Fus3, and Slt2— that were implicated as hubs of conditional genetic interactions (Figure 6.2) and also appeared as a prominent component of the module map (Figure 6.3**C**, blue labels). Two of these proteins, Bck1 and Slt2, are members of the PKC1-mediated cell integrity pathway required for cell wall remodeling and budding in response to stress[37]. We observed conditional negative genetic interactions between these MAPKs and members of the canonical DNA damage checkpoint kinase cascade including *TEL1* (ortholog of human ATM) and *DUN1* (ortholog of human CHK2), suggesting significant crosstalk between this pathway and MAPK pathways (Figure 6.4A). Although MAPK regulation of the DNA damage response has not yet been reported in yeast, it is in agreement with preliminary reports in human in which PKC proteins may regulate poly(ADP-ribose) polymerase-1 (PARP-1)[38] and activate apoptotic caspases in response to cisplatin[39]. Furthermore, human p38/MAPK pathways have been suggested to regulate the cell cycle in response to UV damage in a parallel pathway to CHK1/CHK2[40].

To further study the role of MAPK pathways in signaling DNA damage, we examined both the expression and the cellular localization of the Slt2 MAPK in increasing concentrations of MMS. These experiments indicated that Slt2 is dramatically induced upon exposure to MMS (Figure 6.4B-C) and that it also trans-locates to the nucleus (Figure 6.4C-D). Further analysis revealed a role for Slt2 in the recovery from DNA damage revealed by monitoring the phosphorylation status of the checkpoint kinase

Rad53 (Figure 6.4E). Since MAPK pathways often regulate transcriptional responses[41], we next measured the effect of Slt2 on expression of the ribonucleotide reductase (RNR) complex subunits in response to MMS. Induction of RNR is a useful marker for the DNA damage response, as it is a key component which catalyzes production of nucleotide pools needed for DNA replication and repair and is under both positive and negative control by a number of DNA repair proteins[42,43]. We found that the expression levels of all four RNR subunits, which are strongly induced by MMS, were hyper-induced by 4- to 8-fold in a *slt2*Δ gene deletion mutant (Figure 6.4F).

One interpretation of these results was that Slt2 directly regulates the transcription of RNR subunits, either alone or through interactions with transcription factors[44]. An alternative explanation was that the link between Slt2 and RNR was indirect, with lack of Slt2 increasing sensitivity to DNA damage which, in turn, causes an increase in RNRs. To distinguish between these two possibilities, we used the technique of chromatin immunoprecipitation followed by microchip analysis (ChIP-chip) to determine whether Slt2 binds genomic DNA in the neighborhood of RNR genes and, if so, its precise binding sites at these loci. This technique has been used previously to show that some MAPKs precipitate with genomic DNA through their occupancy at defined target genes[45]. The ChIP experiment showed that the 5' genomic regions of the *RNR1* and *RNR2* loci were among the most heavily bound by Slt2 (occurring in the top 99.5% percentile) (Figure 6.4G, Figure S5). Interestingly, we observed binding of Slt2 to the *RNR1* and *RNR2* loci, the two essential genes in the regulon, but not to loci encoding non-essential genes *RNR3* and *RNR4* (Figure S6). Taken together, our data suggest that MAPK pathways play a vital role in the DNA damage response which may function in parallel to the canonical DNA damage checkpoint kinase pathway converging at the level of *RNR* transcription (Figure 6.4H).

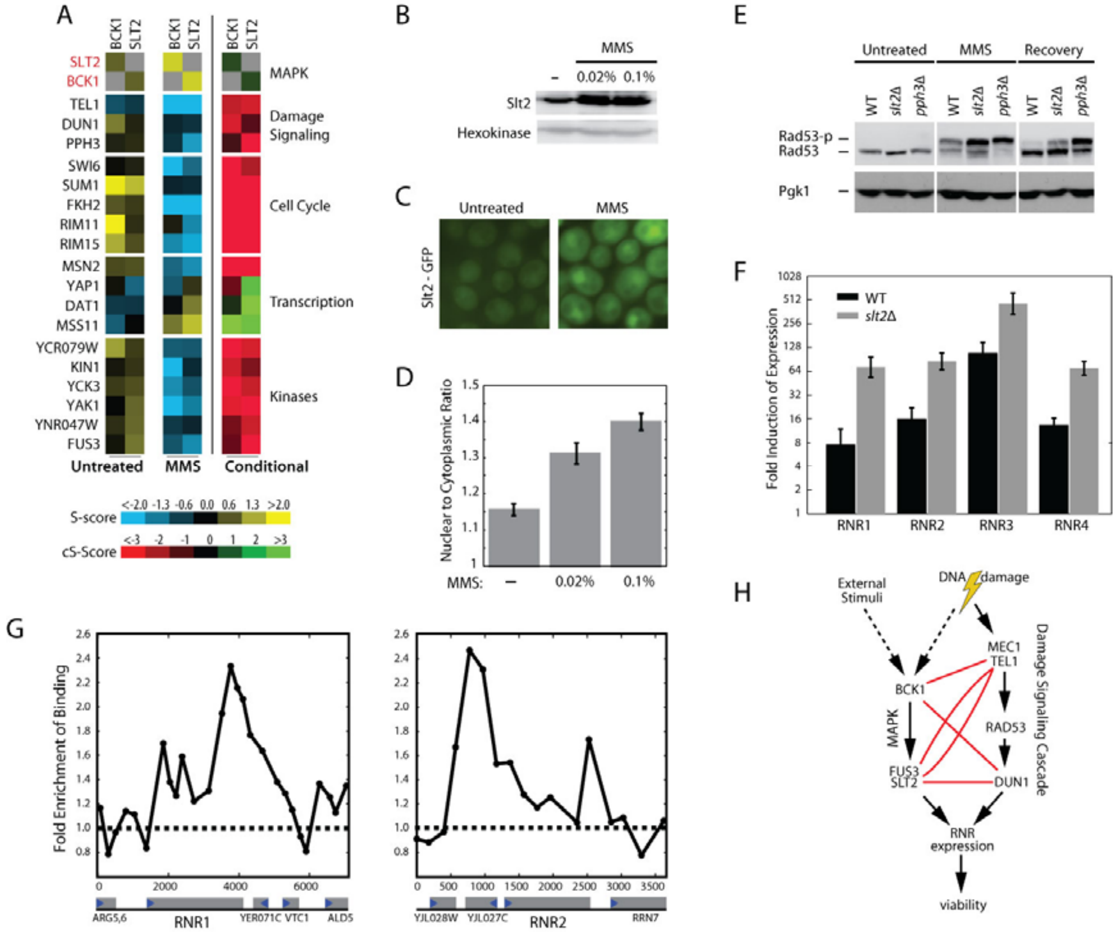


Figure 6.4: Crosstalk between MAPK and DNA repair pathways.

(A) Representative conditional genetic interactions with *BCK1* and *SLT2*. (B) Immunoblot of GFP tagged Slt2 indicated an increase in abundance in response to MMS treatment for 1 hour. Hexokinase is used as a loading control. (C) Fluorescence microscopy of Slt2-GFP cells before and after exposure to 0.1% MMS. (D) Nuclear accumulation of Slt2 in response to MMS. Cytoplasmic versus nuclear fluorescent intensity was quantified for random cells in each condition. Error bars are s.e.m. (E) Phosphorylation status of checkpoint kinase Rad53 in response to MMS and after 1 hour recovery in YPD. *pph3*Δ is used as control and is defective in recovery from MMS. (F) Induction of RNR subunit expression in response to MMS using real-time PCR in wild-type and a *slt2*Δ mutant. (G) Occupancy of the RNR1 and RNR2 genes by Slt2p after 1 hour exposure to 0.03% MMS based on ChIP-chip analyses. The genomic positions of probe regions and their enrichment ratios are displayed on the x and y axes, respectively. Open reading frames are depicted as gray rectangles, and arrows indicate the direction of transcription (H) Schematic illustration of the uncovered role of MAPK pathways in the response to DNA damage based on conditional negative interactions.

The DNA damage phosphatase Pph3 regulates Cbf1, a component of the kinetochore

Pph3 is the catalytic subunit of a conserved phosphatase complex required for dephosphorylation of the DNA damage checkpoint kinase Rad53 and resumption of replication during damage recovery[46,47]. *PPH3* was identified as a major hub of conditional genetic interactions in our DNA damage cE-MAP (Figure 6.2). The spectrum of genetic interactions observed with *pph3Δ* changed substantially under MMS (Figure 6.5**A**), with the strongest positive interactions occurring with DNA repair proteins *RAD17* and *RAD52*, the G1/S cyclin *CLN1*, and a transcription factor and component of the inner kinetochore *CBF1*. We also examined the entire genetic interaction profile of *PPH3*, i.e., the vector of genetic interaction scores pairing *PPH3* with each of the genes on the cE-MAP. Similarity between the genetic interaction profiles of two genes has been termed 'genetic congruence' and suggests a close functional relationship [23,48,49]. As expected, *PPH3* was highly congruent with other members of its phosphatase complex (Psy2, Psy4)[47] in both untreated and MMS-treated conditions (Figure 6.5B). In untreated conditions only, *PPH3* was highly congruent and showed a strong negative interaction with the DNA checkpoint gene *RAD17*. Rad17 is an early DNA damage sensor which binds chromatin before damage occurs in order to efficiently recruit repair machinery[50]. Thus, Pph3 may also have a role in recruitment of repair machinery prior to damage, potentially through Rad17. In MMS only, *PPH3* became highly congruent with the *TEL1* kinase (Pearson correlation of −0.10 to +0.36). This observation is consistent with the role of Tel1 and Pph3 in the phosphorylation and de-phosphorylation of Rad53 and the histone variant γH2AX during DNA damage response and recovery[47,51] and suggests that these factors may have more targets in common.

A gene with an unknown role in DNA repair which became highly congruent with *PPH3* after DNA damage was *CBF1*(Figure 6.5B). This congruence, along with the strong conditional positive interaction between *PPH3* and *CBF1* (Figure 6.5A), suggested a role for Pph3 in the regulation of Cbf1 during the DNA damage response. Like *PPH3*, *CBF1* itself was also identified as a conditional interaction hub (Figure 6.2). Further investigation revealed that Cbf1 is phosphorylated in response to MMS (Figure 6.5C).

Moreover, we found that Cbf1 was hyperphosphorylated in a *pph3*Δ strain independent of the treatment condition, suggesting that damage-dependent phosphorylation of Cbf1 is counteracted by the Pph3 phosphatase (Figure 6.5C) . In addition, using a quantitative mass spectroscopy approach based on phospho-proteome profiling we found that Cbf1 is the most hyperphosphorylated protein detected in a *pph3*Δ strain (Figure 6.5D-E, Table S5). Furthermore, the observed hyper-phosphorylation occurred at Serine-145 followed by Glutamine— this is the canonical SQ phosphorylation motif of Mec1 and Tel1 DNA damage kinases which have also been shown to target Cbf1[9]. To further validate the relationship between *PPH3* and *CBF1*, we compared the expression profiles of *pph3Δ* and *cbf1Δ* mutants using whole yeast genome DNA microarrays. This revealed a very significant overlap in the set of differentially expressed genes, both in untreated conditions (p=10^{-28}) as well as after treatment with MMS (p=10^{-62}) providing further evidence that the two genes function together *in vivo* (Figure S7, Table S6).

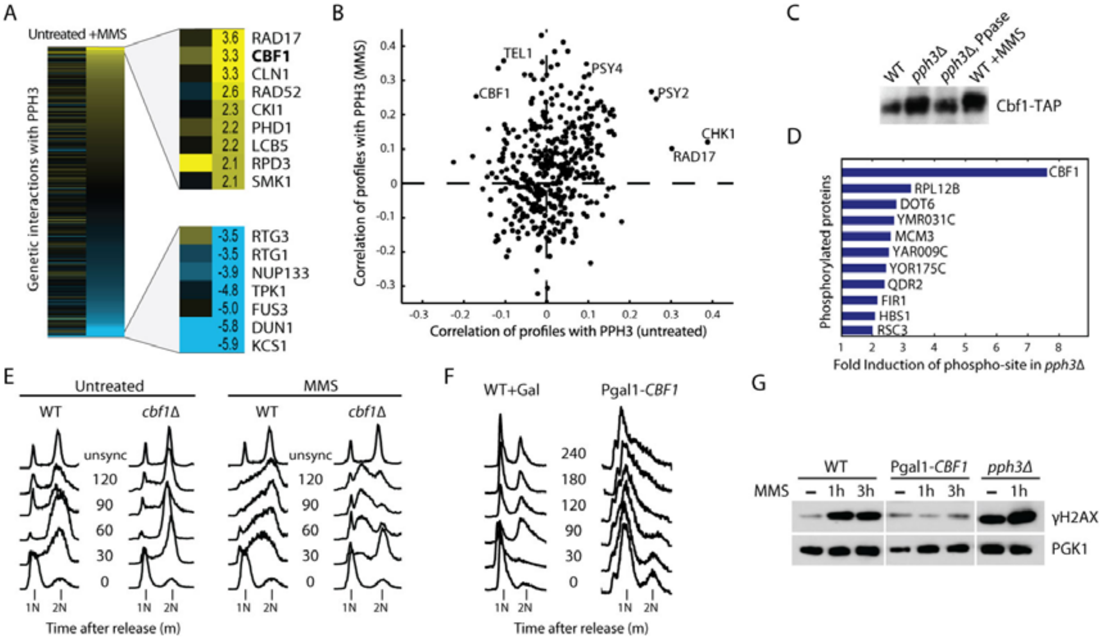


Figure 6.5: Conditional Genetic interactions with Pph3 identify a novel substrate Cbf1 which functions in the DNA damage checkpoint.

(A) Full spectrum of *PPH3* genetic interactions. Genes with the strongest positive and negative genetic interactions in MMS are highlighted. (B) Scatter plot of correlation coefficients for each mutant compared to the profile of *PPH3.*For *PPH3* in each condition, the correlation of profile is shown for each mutant in the same condition. (C) Phos-tag gel shift analysis of TAP tagged Cbf1. PPase indicates phosphatase treatment to remove protein phosphorylation. MMS treated cells were exposed to 0.03% MMS for 1 hour. (D) Comparison of abundance of protein phosphorylation sites in *pph3*Δ cells versus wild-type by phospho-proteomic profiling. (E) Effect of deletion of *CBF1* on cell-cycle progression in arrested cells released into media with or without 0.02% MMS. Facs was performed as described [1]. (F) Effect of over-expression of *CBF1* on cell cycle progression. Cells were shifted to galactose media for three hours before alpha-arrest and released into media containing galactose as carbon source. (G) γH2AX-activation levels for wild-type, *CBF1* over-expression and *pph3*Δ cells. PGK1 is used as a loading control.

In addition to its role as a transcription factor, Cbf1 is a component of the inner kinetochore indicating that it may have a role in cell division. Thus, we hypothesized that it might participate in one of the several damage-dependent cell-cycle checkpoints, including the G1/S, G2/M and intra-S checkpoints which are largely controlled by Mec1 and Tel1 kinases[4]. We found that after arresting cells in G1, *cbf1*Δ cells progressed through the cell cycle similar to wild-type cells in untreated conditions (Figure 6.5E). For cells released into MMS-containing media, the progression of wild-type cells was almost completely halted by checkpoint activation; however, *cbf1*Δ mutant cells progressed freely through the cell cycle, indicating a defect in DNA damage dependent checkpoints. Cells ultimately accumulated in S-phase suggesting failure to complete replication of damaged DNA. Furthermore, overexpression of *CBF1* resulted in cell cycle arrest in G1. (Figure 6.5F). Hence, the regulation of Cbf1 by Pph3 likely plays a key role in the activation and de-activation of cell-cycle checkpoints in response to MMS. In support of this hypothesis, the two strongest positive interactions in untreated conditions with *CBF1* were the cell cycle transcription factor *SWI4* (S=+4.6) and cyclin *PHO80* (S=+3.8).

In response to DNA double-strand breaks, histone H2A is rapidly phosphorylated by the checkpoint kinases Mec1 and Tel1 to form γH2AX, which signals the recruitment and accumulation of DNA repair proteins[52]. In both humans and yeast, Pph3-containing complexes are directly associated with and responsible for the dephosphorylation of γH2AX which is a signal for recovery from the DNA damage checkpoint[47]. Due to the similar mechanisms of regulation, we tested for a relationship between the processes which activate Cbf1 and γH2AX. Intriguingly, over-expression of *CBF1* resulted in nearly complete failure of cells to activate γH2AX in response to MMS (Figure 6.5G). This suggests functional cross-talk between pathways leading to the activation of γH2AX and cell cycle arrest mediated through Cbf1 regulation by Mec1/Tel1 and Pph3. Further work will be required to reveal the mechanistic details of this connection between these two substrates in common between Mec1/Tel1 and Pph3.

Histone variant H2A.Z is regulated by MMS and the checkpoint kinase Mec1

Another means of identifying new DNA-damage response factors is to look for genes with genetic interaction profiles that change dramatically between untreated and MMS-treated conditions. For instance, the genetic interaction profile most altered by MMS treatment was that of *RAD52*, a critical factor in repair of damaged DNA via HR[53]. Another highly-altered profile was that of Htz1 (Figure 6.6A), a histone H2A variant known in metazoans as H2A.Z whose role in DNA repair has not been well established. In untreated conditions, we observed extremely strong congruence of the Htz1 profile with members of the SWR-C complex (*SWR1*, *SWC5*, *VPS71*, *VPS72*), which is responsible for incorporating Htz1 into chromatin[54-56]. In MMS, this congruence was almost completely lost, suggesting a functional disassociation with SWR-C (Figure 6.6B). For instance, correlation of the profiles of *HTZ1* and *SWR1*, the catalytic subunit of SWR-C, fell from 0.65 to 0.00 after MMS exposure. Conversely, in MMS *HTZ1* became highly congruent with DNA-damage checkpoint kinases *MEC1* (correlation from -0.04 to 0.34) and *RAD53* (0.04 to 0.25) suggesting active regulation of Htz1 by checkpoint kinases after DNA damage (Figure 6.6B). In MMS, *HTZ1* also acquired positive interactions with members of the RAD52 epistasis group (RAD52-EG) (Figure 6.6C).

We further tested the function of *HTZ1* in combination with the essential kinase *MEC1*. To support the genetic interactions uncovered using the *MEC1* hypomorphic allele in the cE-MAP, we suppressed the lethality of *mec1Δ* by simultaneous deletion of *SML1*, an inhibitor of ribonucleotide reductase[57]. Using a *sml1*Δ background, we observed that the deletion of *htz1Δmec1Δsml1Δ* together was synthetically sick when

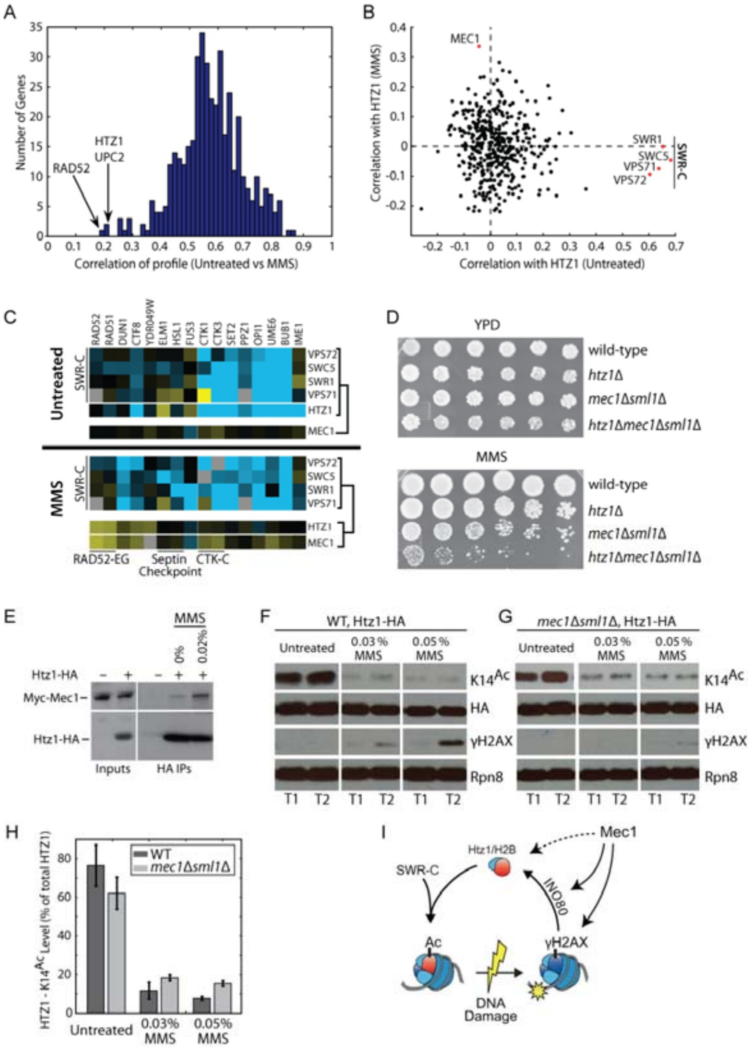


Figure 6.6: The checkpoint kinase, Mec1, associates with and regulates the histone H2A variant Htz1.

(A) Histogram of correlation coefficients of the genetic interaction profile of each mutant compared between untreated and MMS. (B) Scatter plot of correlation coefficients for each mutant compared to the profile of *HTZ1*. For *HTZ1* in each condition, the correlation of profile is shown for each mutant in the same condition. Four members of the SWR-C complex are highlighted. (C) Representative genetic interactions that are in common with *HTZ1* and members of the SWR-C in untreated conditions and are in common with *HTZ1* and *MEC1* in MMS. (D) Synthetic sickness of *htz1*Δ with *mec1*Δ*sml1*Δ in MMS. Plates were grown for 3 days at 30 degrees with 0.02% MMS. (E) Coimmunoprecipitations of Myc-Mec1 with Htz1-3HA. The indicated strains were immunoprecipitated with HA antibody either with or without prior MMS treatment (0.03% for 1 hour), and the blot was cut and probed with a monoclonal MYC antibody to detect Mec1. The input represents 1/200th of the sample for the IPs (F) Htz1 acetylation is dramatically reduced in MMS. Total acetylation of Lysine 14 was monitored in a Htz1-3HA strain for two independent replicates (T1,T2). HA represents total amount of Htz1, Rpn8 serves as loading control. (G) Same as in F but with a *mec1*Δ*sml1*Δ strain, the γH2AX, a Mec1 substrate, serves to verify the deletion of *MEC1*. (H) Quantification of F and G. K14 acetylation channel was compared to the HA channel. Error bars represent standard deviation. (I) Proposed model of Htz1 function in response to DNA damage.

grown on MMS (Figure 6.6D). Previous observations have shown that while two mutations in a nonessential pathway or complex often result in a positive genetic interaction, the introduction of two mutations into an essential pathway is very likely to have a synthetic negative effect on growth[25,30]. This phenomenon could reflect a situation where mutations in essential genes leave the cell in a vulnerable state and that further insults in the pathway mediated by this essential gene lead to a significant negative effect on growth. Hence, both the conditional correlation and negative genetic interaction suggest a direct functional connection between Mec1 and Htz1.

We therefore tested whether the Htz1 and Mec1 proteins could physically interact *in vivo*. Via coimmunoprecipitations of an N-terminal MYC tagged Mec1 (Myc-Mec1) and Htz1-3HA tagged strain, we found a significant physical association between Htz1 and Mec1 which was increased in the presence of MMS (Figure 6.6E). To further understand both the dramatic change in genetic interaction profile (Figure 6.6A) and the condition specific correlation with Mec1 (Figure 6.6B), we monitored Htz1 acetylation levels in response to MMS. After incorporation into chromatin by the SWR-C complex, Htz1 is acetylated on Lys 14 (K14) by the NuA4 histone acetyltransferase complex which has been observed to impact DNA repair processes as well as chromosome transmission and telomeric silencing [58]. We found that while the total amount of Htz1 was nearly unaltered, Htz1-K14 acetylation levels were dramatically reduced in response to a 1 hour treatment with MMS (Figure 6.6F). Consistent with condition specific congruence and genetic interactions with Mec1, we found that the deacetylation of Htz1 in response to MMS was abrogated in a *mec1*Δ mutant (Figure 6.6G-H), suggesting that Htz1 deacetylation in response to MMS is at least partly mediated through the activity of

Mec1.

The reduction in acetylation could be due to the deacetylation of already incorporated Htz1, or a reduction of Htz1-containing nucleosomes. Sub-cellular fractionation indicated that while the total level of acetylated Htz1 in response to MMS was diminished nearly 95%, there was only a 50% loss of acetylation on chromatin-bound Htz1 (Figure S9). In addition, there was nearly a 30% reduction in the total amount of chromatin-bound Htz1, indicating that the observed reduction in acetylation is due to reduction in the amount of chromatin-bound Htz1 (Figure S9). Consistent with this model and our genetic data, previous reports have indicated that Htz1 - containing nucleosomes are cleared from DNA double-strand break regions by the INO80 complex in a manner independent of the SWR-C, while γH2AX accumulates in these regions[59,60]. In addition, both γH2AX and the INO80 subunit, Ies4, are damage-dependent substrates of Mec1[61]. Taken together, our data supports a model where Mec1 physically associates with incorporated Htz1 catalyzing its deacetylation in response to MMS via eviction of htz1-containing nucleosomes from chromatin near sites of damage (Figure 6.6I).

Perspective

By monitoring the dynamics of genetic interactions in response to DNA damage by MMS, we were able to specifically reveal the genetic architecture of signaling underlying the DNA damage response. In contrast to static genetic interactions, we found that a large proportion of conditional genetic interactions specifically mapped components of the damage response. These interactions were able to highlight dynamic functional connections between various pathways and complexes as well as identify new pathways that are specifically activated under DNA damage stress. Critically, we found that the comparison of untreated and MMS maps was much more sensitive in the identification of dynamic genetic interactions than analysis of either condition alone; suggesting a new paradigm for exploring the mechanism of action of drugs and other external stimuli.

Genetic interactions uncovered by exposure to MMS reveal a plethora of new pathways which we begin to describe here. For example, conditional interactions highlighted a role for the Slt2 MAPK in the transcriptional response to DNA damage. We also identified Cbf1 as a component of the damage dependent cell-cycle checkpoint and as a new target of the major DNA-damage dependent phosphatase, Pph3. Lastly, conditional interactions pointed to a novel role for the DNA-damage checkpoint kinase, Mec1, in the regulation of Htz1 in response to MMS. Furthermore, the map of the relationships between multimeric modules (Figure 6.3C) serves as a roadmap for functional inference of the interconnectivity of modules after DNA damage. However, because most kinases and transcription factors do not operate in multimeric modules, other methods might be developed to highlight pathways implicated in the function of single genes. Such methods include the identification of network motifs which are common in the genetic interaction data and can identify shared pathway membership[26].

Synthetic lethality has been exploited for the discovery of parallel pathways required to compensate for oncogenic mutations and is an emerging paradigm for the discovery of new therapautics (Reinhardt et al, 2009). A potential use of conditional negative genetic interactions might be used to suggest pathways working in parallel with genes involved in DNA repair and oncogenesis which might then be targeted with drugs to specifically sensitize cancer cells to chemotherapeutic agents such as MMS. This approach has been exploited using the synthetic lethal relationship between the DNA repair protein Poly (ADP-Ribose) Polymerase (PARP) and genes commonly mutated and defective in breast and ovarian cancer, BRCA1 and BRCA2. PARP and BRCA genes function in a redundant fashion, where PARP inhibition results in the generation of double strand breaks (DSB) which are then repaired via BRCA1/2[62]. For BRCA1/2 mutated cancers, PARP inhibitors are able to specifically sensitize cancer cells to DNA damage (Farmer et al, 2005). In such a manner, the data in yeast suggests a variety of conserved pathways which might be targeted to specifically inhibit replication in cells with mutations or other aberrations in oncogenic pathways.

The cE-MAP in budding yeast quantitatively maps pairwise genetic interactions among nearly all kinases, phosphatases, transcription factors and chromatin and DNA damage machinery. This dataset provides a launching point to further study the genetic architecture between a critical interface in biology, signaling and transcription. For example, we have noted that kinases and their substrates tend to share positive genetic interactions and have correlated genetic interaction profiles. Integrating this information alone or with other high throughput datasets can aid in the identification of novel kinase-substrate relationships including those which are DNA damage dependent. Future work might examine the effects of other drugs on focused sets of genes such as rapamycin on RNA Processing and benomyl on spindle function or probe mechanism of action of uncharacterized compounds using more unbiased gene sets. In addition, comparison with similar data from other species such as S. *pombe* can reveal conserved and diverged interactions between species[63] providing evolutionary insight into the drug response and point to which interactions which might be conserved in humans.

Methods

E-MAP experiments.

Strain construction, plating of mutants, mutant selection, and scoring of genetic interactions in each condition were performed as previously described[15,16]. For the MMS map, double mutants were grown under the appropriate drug selection and 0.02% MMS. This protocol resulted in a quantitative *S*-score assigned to each gene deletion pair, in which $S \leq -2.5$ is considered a significant negative interaction, and $S \geq 2.0$ is considered a significant positive interaction[15].

Conditional interaction scoring system.

Conditional interactions were evaluated by first computing the difference in *S*-score ($S_1 - S_2$) of a double gene deletion strain grown in two conditions 1 and 2. To estimate the null distribution of this difference, replicate *S*-scores were obtained from a set of 8018 double deletions assayed in identical conditions in both of two previously published E-MAPs (Figure S2)[22,26]. We found that this null distribution had a standard deviation that increased non-linearly with the magnitude of *S*-score (Figure S2). This deviation was estimated as a non-parametric function $\sigma(S_1 + S_2)$ of the sum of *S*-scores using a sliding window. The conditional *S*-score was then computed as:

$$cS = sign(S_1 - S_2) \times \log_{10}\left(1 - \Phi\left|\frac{S_1 - S_2}{\sigma}\right|\right)$$

where Φ is the cumulative distribution function of the standard normal distribution. The *cS* score reflects the log p-value of the normalized difference in *S*-score between conditions (i.e., $cS = 3$ reflects a p-value of 0.001) with the sign reflecting the direction of change in *S*-score of condition 1 relative to condition 2. In our study, condition 1 was set to MMS-treated cells and condition 2 was set to untreated cells. An interaction with cS-score ≥ 3 was considered ‘conditional positive’ in MMS, and an interaction with cS-score ≤ −3 was considered ‘conditional negative’ (Table S2).

Assembly of the module map.

The module map in Figure 3C was generated as previously described [30] with the following modifications. A physical protein-protein interaction network was downloaded from the Biogrid database (interactions annotated as genetic were excluded) and filtered to those interactions supported by two or more publications (Table S3). This was combined with interactions based on tandem affinity purification followed by mass spectrometry having PE-score ≥ 2[64]. Members of the same module were required to be connected in this physical network. Modules were initially identified using the maximum absolute value of the *S*-score in either condition using a cluster reward of 5. Within-module scores were calculated using both the genetic interaction strength as well as correlation of genetic interaction profiles. Links between modules were determined based on the hypergeometric enrichment for either conditional positive of negative interactions among all possible pairs of members between modules. Modules links with enrichment *p*-value < 0.001 are shown (Table S4).

Acknowledgements

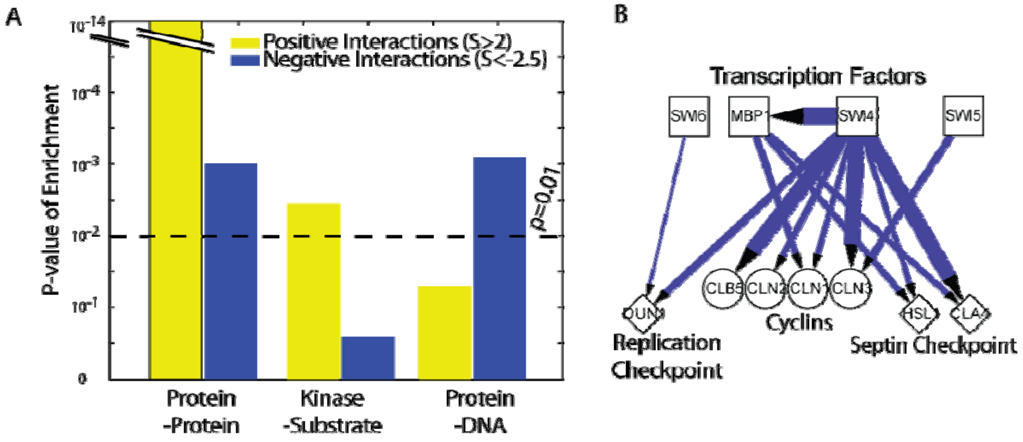
The authors would like to thank Sean Collins, Ryan Kelley, Hans Hombauer, Arshad Desai, Susan Gasser and Xuetong Shen for helpful discussions and strains. This work was funded by a grant from the U. S. National Institutes of Health (R01-ES14811). W.H. and M.S. were funded by the 21C Frontier Functional Proteomics Project (FPR08A1-060), Republic of Korea.

Chapter 6, in full, is the following manuscript currently in preparation,

Bandyopadhyay S, Mehta M, Kuo D, Sung M, Jaehnig E, Chuang R, Bodenmiller B, Licon K, Copeland W, Shales M, Fiedler D, Shokat KM, Kolodner RD, Huh W, Aebersold R, Keogh MC, Krogan NJ, Ideker T. *DNA-damage induced rewiring of protein signaling revealed by a conditional epistatic interaction map (cE-MAP)*. In preparation.

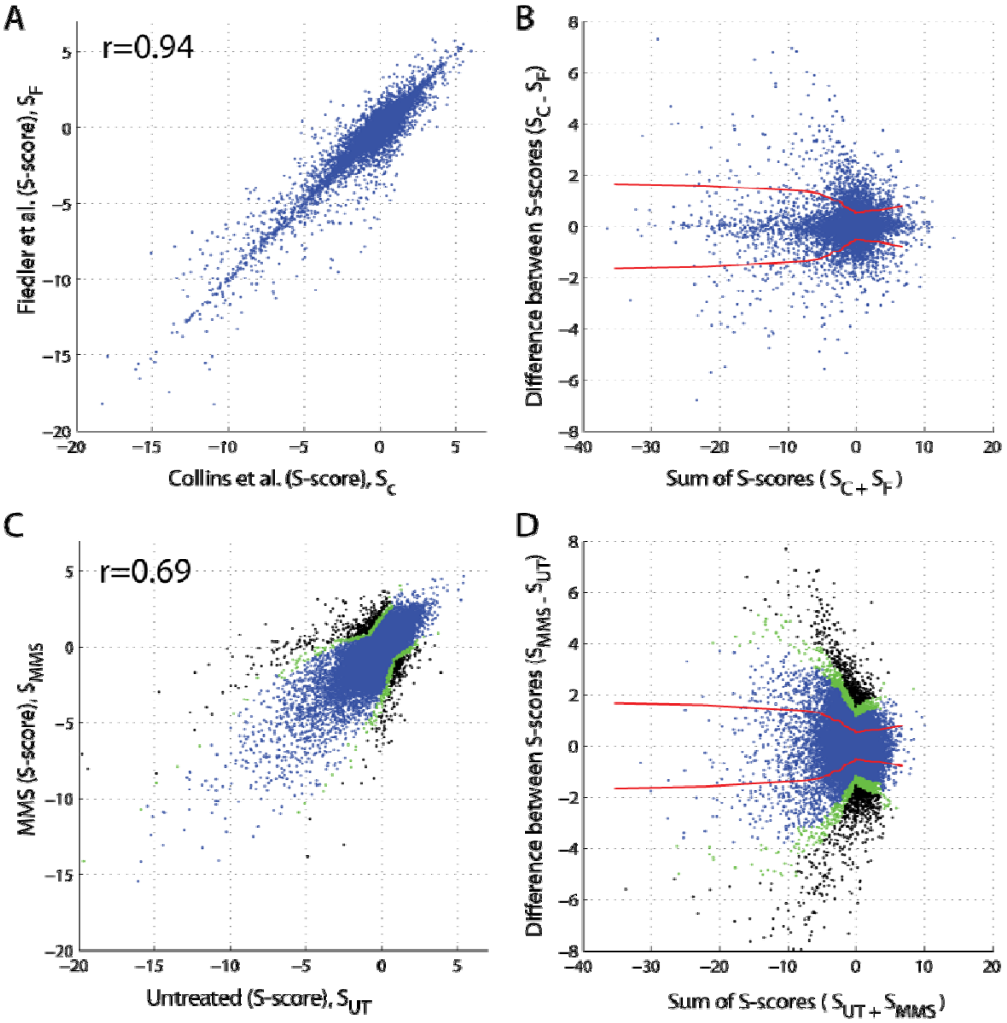
The dissertation author is the sole first author on this work, responsible for study design and data analysis.

Supplemental Figures



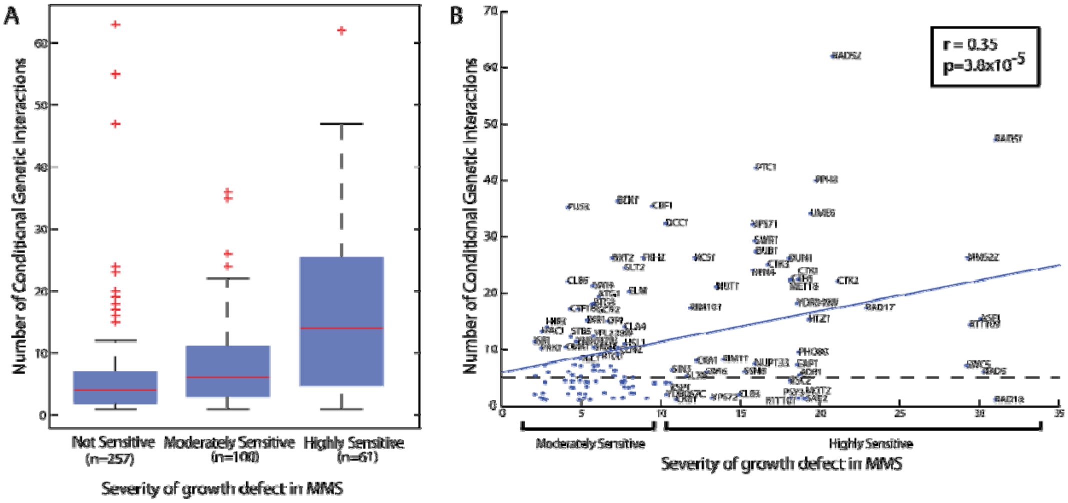
Supplementary Figure 6.1: Analysis of physical interactions overlapping with the untreated E-MAP.

(A) Enrichment of positive and genetic interactions for various physical interaction datasets. Sources: Collins et al *Mol. Cell. Proteomics* 6(3):439-50 (Protein-Protein), Fielder et al. *Cell* 136(5):952-63. (Kinase-substrate), Monteiro et al. *Nuc. Acids. Res.* 36:D132-D136 (Protein-DNA). **(B)** Subnetwork highlighting negative genetic interactions among cell-cycle transcription factors and downstream target genes. Thickness of edge corresponds to strength of negative interactions.



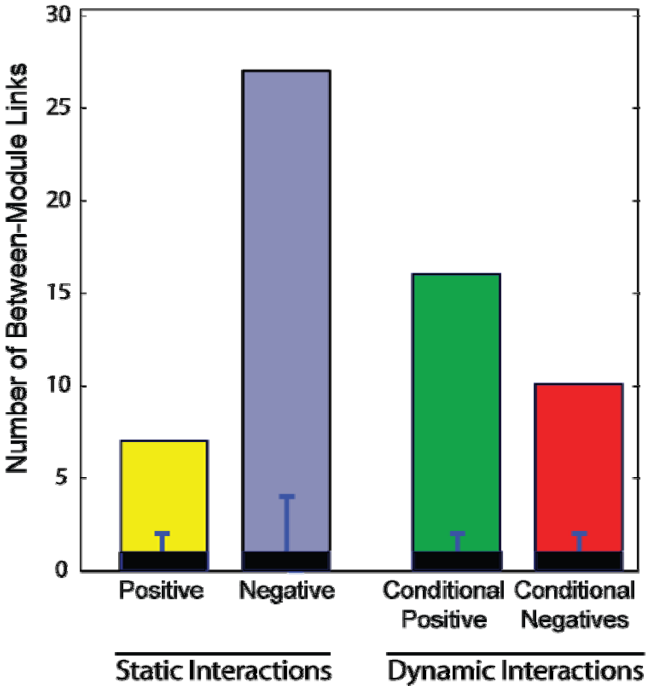
Supplementary Figure 6.2: Determination of condition-specific genetic interactions.

When comparing two untreated E-MAPs, we found that the difference in S-Score over the same gene pair was linked with the magnitude of the genetic interaction score. **(A)** Comparison of genetic interaction S-scores for 8,018 double mutants in common in two E-MAPs. Pearson correlation of scores is shown. **(B)** Same data plotted as the sum of S-scores on the x-axis and the difference between them on the y-axis. Along a sliding window, the observed mean difference was near zero but the variance increased with stronger negative and positive S-scores (red line). **(C)** Scatter of untreated versus MMS S-scores in this cE-MAP. **(D)** Sum versus difference for pairs in the cE-MAP. Red line is the same as in (B). Double deletions whose difference in S-score between conditions was significantly greater than expected based on comparisons in (B) are indicated (green and black, $p<0.01$; black, $p<0.001$).



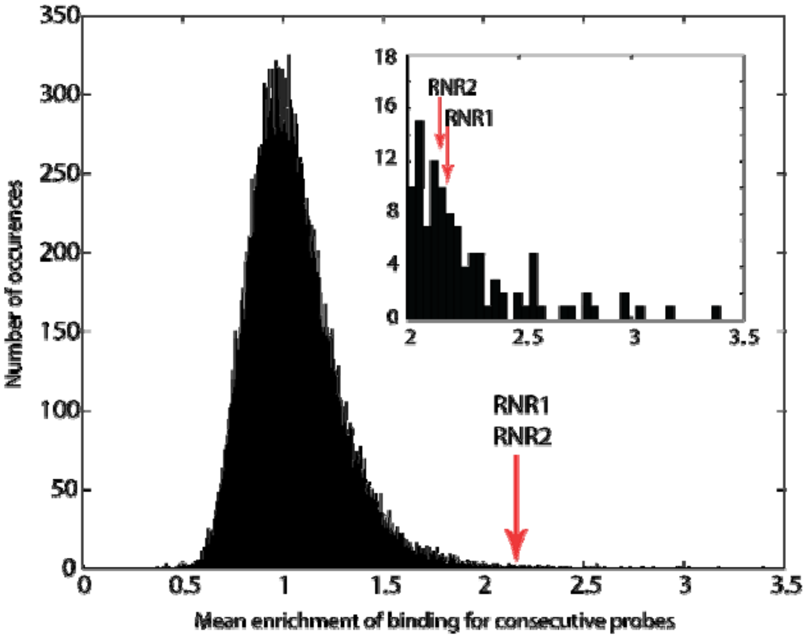
Supplemental Figure 6.3: The number of conditional genetic interactions associated with a gene is linked to the severity of growth defect of the gene–knockout upon MMS exposure.

(A) All genes in the E-MAP were binned based on the severity of their defect (Begley et al, Mol. Cancer Research 2002). **(B)** For genes identified as sensitive to MMS, quantitative estimates of the severity of growth defect in MMS are significantly correlated with the number of conditional genetic interactions for a give gene (solid line). The number of conditional genetic interactions expected at random is shown (dotted line).



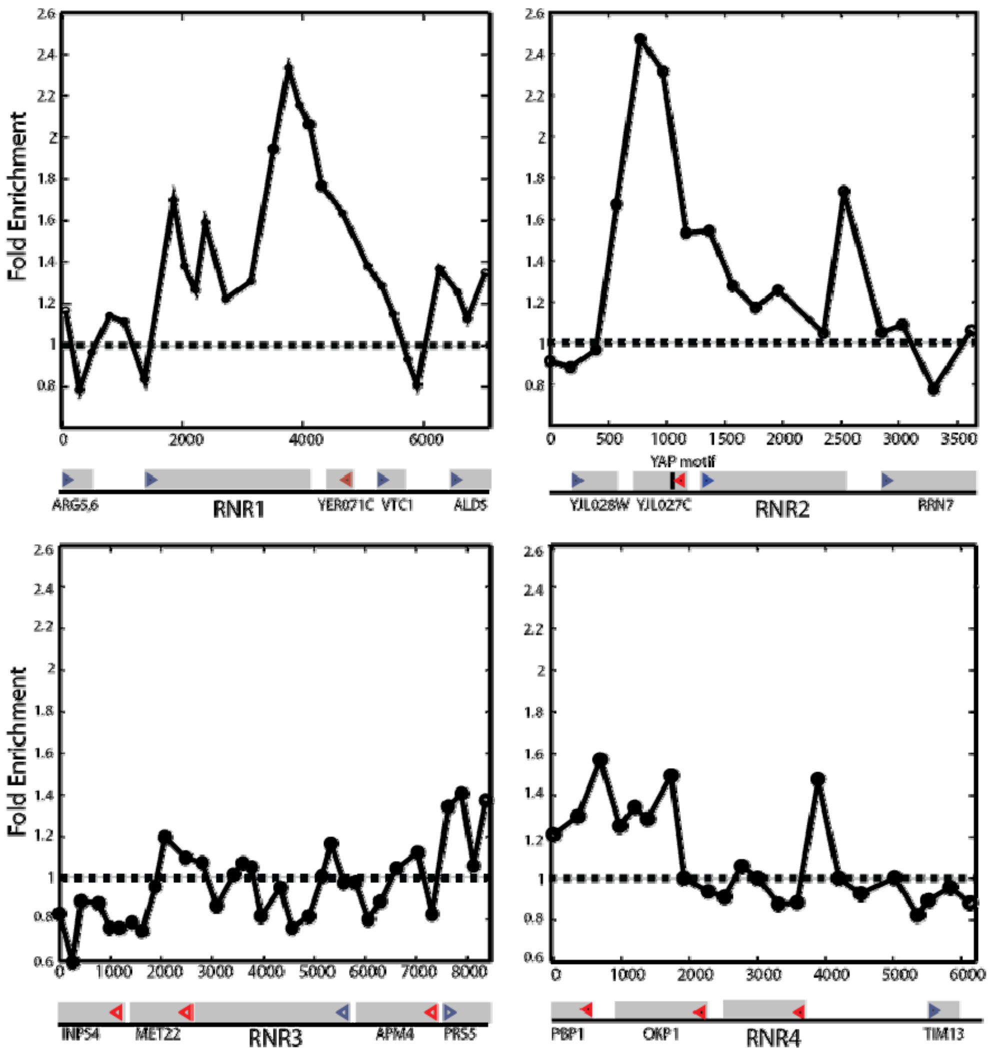
Supplemental Figure 6.4: There are more between-module links than expected at random.

The number of between-module links between a set of well-characterized, curated modules was evaluated (Table S1). The number of between-module links expected at random (black bars) was established by permuting the genetic interaction data and evaluating the number of between module links identified in the permuted dataset over a total of 100 trials. In most cases, the number of between-module links expected by chance was ≤1. Error bars represent the 95th percentile.



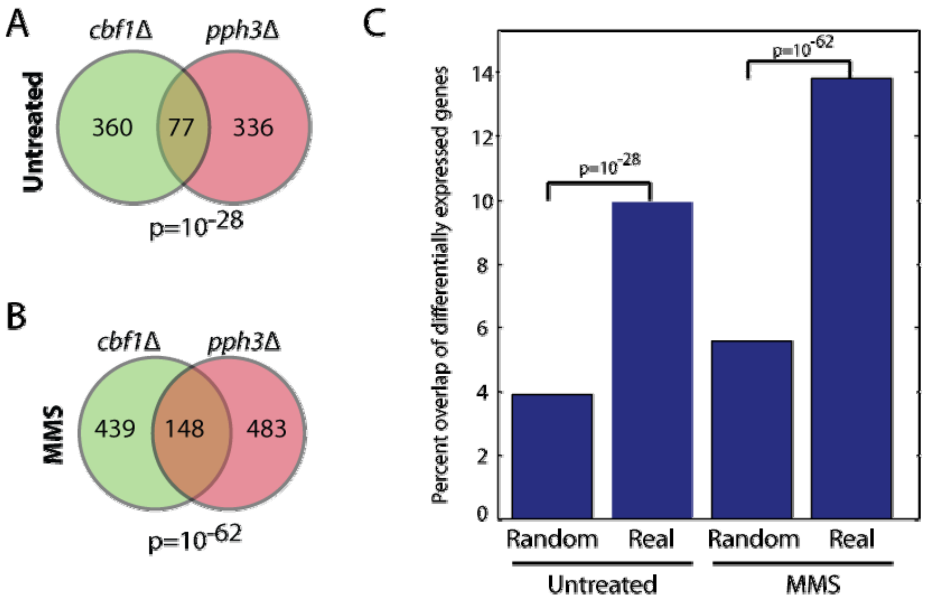
Supplemental Figure 6.5: The 3' region of *RNR1* and the 5' region of *RNR2* are among the most heavily bound genomic regions by Slt2p.

Histogram of the mean enrichment of binding for each consecutive three-probe region in the genome is shown. Both *RNR1* and *RNR2* were in the top 0.5% of bound regions.



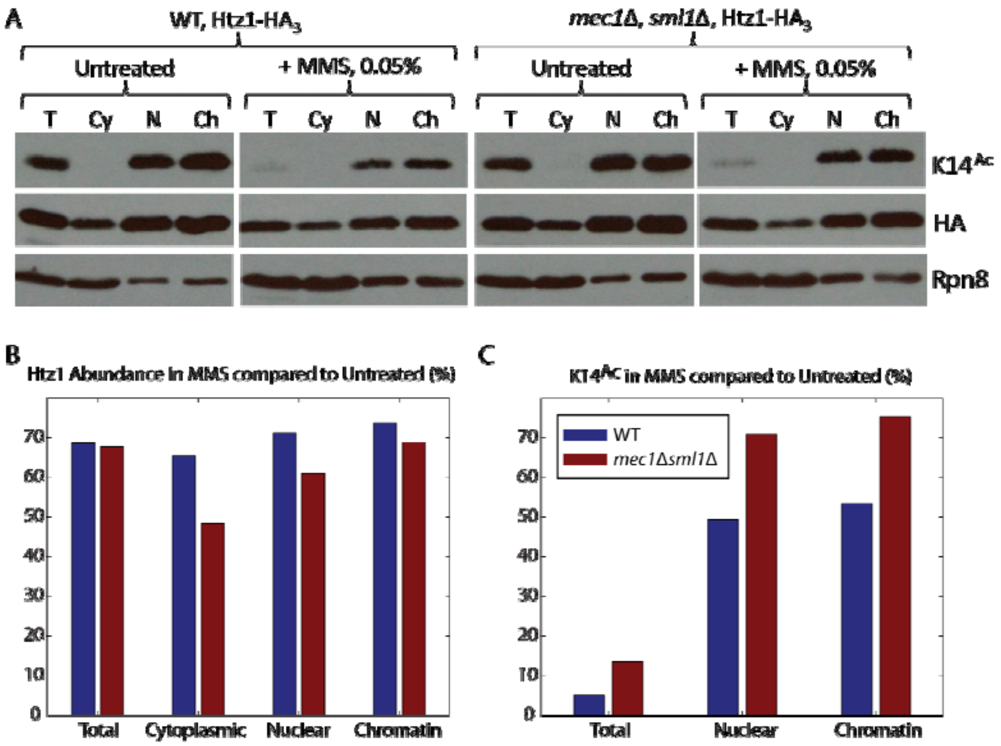
Supplemental Figure 6.6: Slt2 binds the genomic segments proximal to *RNR1* and *RNR2* but not *RNR3* and *RNR4*.

Occupancy of the genes by Slt2p after 1 hour exposure to 0.03% MMS based on genome-wide ChIP-Chip analyses. The genomic positions of probe regions and their enrichment ratios are displayed on the x and y axes, respectively. Open reading frames are depicted as gray rectangles, and arrows indicate the direction of transcription



Supplemental Figure 6.7: Microarray analysis of *PPH3* and *CBF1*

(A) Overlap of genes differentially expressed in a *CBF1* and *PPH3* knockout based on comparison of microarrays measuring *cbf1*Δ vs wt and *pph3*Δ vs wt. P-values represent significant overlap between the sets of differentially expressed genes (Table S6). **(B)** Comparison of genes differentially expressed in a *CBF1* and *PPH3* knockout based on comparison of microarrays as in (A) except all samples were treated with 0.02% MMS for 1 hour. **(C)** Comparison of the percent overlap of differentially expressed genes. Overlap is calculated as the number of genes in common over the total number of unique genes differentially expressed in both samples. The number of overlapping genes expected at random is shown for comparison. P-values were assessed using a hypergeometric test with a background of 6,000 genes.



Supplementary Figure 6.8: Fractionation indicates that chromatin-bound Htz1 is not deacetylated in response to MMS.

(A) Cells containing Htz1-HA3 were separated into total (T), cytoplasmic (C), nuclear (N) and chromatin (Ch) fractions and immunoblotted with the indicated antibodies. **(B)** Quantification of blots in A using Rpn8 as loading control. The total amount of Htz1 is reduced ~30% in total, nuclear and chromatin fractions in response to MMS. **(C)** The total level of acetylated Htz1 is reduced ~90% upon exposure to MMS. In both nuclear and chromatin fractions this reduction is only ~30-50% indicating that the deacetylation of Htz1 is due to decreased levels of Htz1 in chromatin rather than deacetylation of existing chromatin-bound Htz1.

Bibliography

1. Enserink JM, Hombauer H, Huang ME, Kolodner RD. Cdc28/Cdk1 positively and negatively affects genome stability in S. cerevisiae. *J Cell Biol* 2009;185:423-37.

2. Shiloh Y. ATM and related protein kinases: safeguarding genome integrity. *Nat Rev Cancer* 2003;3:155-68.

3. Begley TJ, Rosenbach AS, Ideker T, Samson LD. Damage recovery pathways in Saccharomyces cerevisiae revealed by genomic phenotyping and interactome mapping. *Mol Cancer Res* 2002;1:103-12.

4. Zhou BB, Elledge SJ. The DNA damage response: putting checkpoints in perspective. *Nature* 2000;408:433-9.

5. Chang M, Bellaoui M, Boone C, Brown GW. A genome-wide screen for methyl methanesulfonate-sensitive mutants reveals genes required for S phase progression in the presence of DNA damage. *Proc Natl Acad Sci U S A* 2002;99:16934-9.

6. Lee W, St Onge RP, Proctor M, Flaherty P, Jordan MI, Arkin AP, Davis RW, Nislow C, Giaever G. Genome-wide requirements for resistance to functionally distinct DNA-damaging agents. *PLoS Genet* 2005;1:e24.

7. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol Biol Cell* 2001;12:2987-3003.

8. Workman CT, Mak HC, McCuine S, Tagne JB, Agarwal M, Ozier O, Begley TJ, Samson LD, Ideker T. A systems approach to mapping DNA damage response pathways. *Science* 2006;312:1054-9.

9. Smolka MB, Albuquerque CP, Chen SH, Zhou H. Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases. *Proc Natl Acad Sci U S A* 2007;104:10364-9.

10. Ravi D, Wiles AM, Bhavani S, Ruan J, Leder P, Bishop AJ. A network of conserved damage survival pathways revealed by a genomic RNAi screen. *PLoS Genet* 2009;5:e1000527.

11. Paulsen RD, Soni DV, Wollman R, Hahn AT, Yee MC, Guan A, Hesley JA, Miller SC, Cromwell EF, Solow-Cordero DE, Meyer T, Cimprich KA. A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability. *Mol Cell* 2009;35:228-39.

12. Matsuoka S, Ballif BA, Smogorzewska A, McDonald ER, 3rd, Hurov KE, Luo J, Bakalarski CE, Zhao Z, Solimini N, Lerenthal Y, Shiloh Y, Gygi SP, Elledge SJ. ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* 2007;316:1160-6.

13. Boone CB, H. Andrews, B. Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics* 2007.

14. Beltrao P, Cagney G, Krogan NJ. Quantitative Genetic Interactions Reveal Layers of Biological Modularity. *Cell* 2010;In Press.

15. Collins SR, Schuldiner M, Krogan NJ, Weissman JS. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol* 2006;7:R63.

16. Schuldiner M, Collins SR, Weissman JS, Krogan NJ. Quantitative genetic analysis in Saccharomyces cerevisiae using epistatic miniarray profiles (E-MAPs) and its application to chromatin functions. *Methods* 2006;40:344-52.

17. Beyer A, Bandyopadhyay S, Ideker T. Integrating physical and genetic maps: from genomes to interaction networks. *Nat Rev Genet* 2007;8:699-710.

18. Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 2005;23:561-6.

19. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 2001;294:2364-8.

20. Ooi SL, Shoemaker DD, Boeke JD. DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nat Genet* 2003;35:277-86.

21. Pan X, Ye P, Yuan DS, Wang X, Bader JS, Boeke JD. A DNA integrity network in the yeast Saccharomyces cerevisiae. *Cell* 2006;124:1069-81.

22. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z, Brown GW, Ingles CJ, Emili A, Allis CD, Toczyski DP, Weissman JS, Greenblatt JF, Krogan NJ. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 2007;446:806-10.

23. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 2005;123:507-19.

24. St Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* 2007;39:199-206.

25. Wilmes GM, Bergkessel M, Bandyopadhyay S, Shales M, Braberg H, Cagney G, Collins SR, Whitworth GB, Kress TL, Weissman JS, Ideker T, Guthrie C, Krogan NJ. A genetic interaction map of RNA-processing factors reveals links between Sem1/Dss1-containing complexes and mRNA export and splicing. *Mol Cell* 2008;32:735-46.

26. Fiedler D, Braberg H, Mehta M, Chechik G, Cagney G, Mukherjee P, Silva AC, Shales M, Collins SR, van Wageningen S, Kemmeren P, Holstege FC, Weissman JS, Keogh MC, Koller D, Shokat KM, Krogan NJ. Functional organization of the S. cerevisiae phosphorylation network. *Cell* 2009;136:952-63.

27. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet* 2006;38:896-903.

28. Measday V, Baetz K, Guzzo J, Yuen K, Kwok T, Sheikh B, Ding H, Ueta R, Hoac T, Cheng B, Pot I, Tong A, Yamaguchi-Iwai Y, Boone C, Hieter P, Andrews B. Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. *Proc Natl Acad Sci U S A* 2005;102:13956-61.

29. Barlow JH, Rothstein R. Rad52 recruitment is DNA replication independent and regulated by Cdc28 and the Mec1 kinase. *Embo J* 2009;28:1121-30.

30. Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T. Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol* 2008;4:e1000065.

31. Hampsey M, Kinzy TG. Synchronicity: policing multiple aspects of gene expression by Ctk1. *Genes Dev* 2007;21:1288-91.

32. Ostapenko D, Solomon MJ. Budding yeast CTDK-I is required for DNA damage-induced transcription. *Eukaryot Cell* 2003;2:274-83.

33. Wood A, Shukla A, Schneider J, Lee JS, Stanton JD, Dzuiba T, Swanson SK, Florens L, Washburn MP, Wyrick J, Bhaumik SR, Shilatifard A. Ctk complex-mediated regulation of histone methylation by COMPASS. *Mol Cell Biol* 2007;27:709-20.

34. Xiao T, Shibata Y, Rao B, Laribee RN, O'Rourke R, Buck MJ, Greenblatt JF, Krogan NJ, Lieb JD, Strahl BD. The RNA polymerase II kinase Ctk1 regulates positioning of a 5' histone methylation boundary along genes. *Mol Cell Biol* 2007;27:721-31.

35. Shen X, Mizuguchi G, Hamiche A, Wu C. A chromatin remodelling complex involved in transcription and DNA processing. *Nature* 2000;406:541-4.

36. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJ, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M. Global analysis of protein phosphorylation in yeast. *Nature* 2005;438:679-84.

37. Verna J, Lodder A, Lee K, Vagts A, Ballester R. A family of genes required for maintenance of cell wall integrity and for the stress response in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 1997;94:13804-9.

38. Hegedus C, Lakatos P, Olah G, Toth BI, Gergely S, Szabo E, Biro T, Szabo C, Virag L. Protein kinase C protects from DNA damage-induced necrotic cell death by inhibiting poly(ADP-ribose) polymerase-1. *FEBS Lett* 2008;582:1672-8.

39. Basu A. Involvement of protein kinase C-delta in DNA damage-induced apoptosis. *J Cell Mol Med* 2003;7:341-50.

40. Manke IA, Nguyen A, Lim D, Stewart MQ, Elia AE, Yaffe MB. MAPKAP kinase-2 is a cell cycle checkpoint kinase that regulates the G2/M transition and S phase progression in response to UV irradiation. *Mol Cell* 2005;17:37-48.

41. Karin M. The regulation of AP-1 activity by mitogen-activated protein kinases. *J Biol Chem* 1995;270:16483-6.

42. Elledge SJ, Davis RW. DNA damage induction of ribonucleotide reductase. *Mol Cell Biol* 1989;9:4932-40.

43. Jordan A, Reichard P. Ribonucleotide reductases. *Annu Rev Biochem* 1998;67:71-98.

44. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;431:99-104.

45. Pokholok DK, Zeitlinger J, Hannett NM, Reynolds DB, Young RA. Activated signal transduction kinases frequently occupy target genes. *Science* 2006;313:533-6.

46. Chowdhury D, Xu X, Zhong X, Ahmed F, Zhong J, Liao J, Dykxhoorn DM, Weinstock DM, Pfeifer GP, Lieberman J. A PP4-phosphatase complex dephosphorylates gamma-H2AX generated during DNA replication. *Mol Cell* 2008;31:33-46.

47. Keogh MC, Kim JA, Downey M, Fillingham J, Chowdhury D, Harrison JC, Onishi M, Datta N, Galicia S, Emili A, Lieberman J, Shen X, Buratowski S, Haber JE, Durocher D, Greenblatt JF, Krogan NJ. A phosphatase complex that dephosphorylates gammaH2AX regulates DNA damage checkpoint recovery. *Nature* 2006;439:497-501.

48. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Global mapping of the yeast genetic interaction network. *Science* 2004;303:808-13.

49. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS. Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol Syst Biol* 2005;1:2005 0026.

50. Zou L, Cortez D, Elledge SJ. Regulation of ATR substrate selection by Rad17-dependent loading of Rad9 complexes onto chromatin. *Genes Dev* 2002;16:198-208.

51. Travesa A, Duch A, Quintana DG. Distinct phosphatases mediate the deactivation of the DNA damage checkpoint kinase Rad53. *J Biol Chem* 2008;283:17123-30.

52. Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem* 1998;273:5858-68.

53. Paques F, Haber JE. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev* 1999;63:349-404.

54. Kobor MS, Venkatasubrahmanyam S, Meneghini MD, Gin JW, Jennings JL, Link AJ, Madhani HD, Rine J. A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. *PLoS Biol* 2004;2:E131.

55. Krogan NJ, Keogh MC, Datta N, Sawa C, Ryan OW, Ding H, Haw RA, Pootoolal J, Tong A, Canadien V, Richards DP, Wu X, Emili A, Hughes TR, Buratowski S, Greenblatt JF. A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol Cell* 2003;12:1565-76.

56. Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, Wu C. ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* 2004;303:343-8.

57. Zhao X, Muller EG, Rothstein R. A suppressor of two essential checkpoint genes identifies a novel protein that negatively affects dNTP pools. *Mol Cell* 1998;2:329-40.

58. Keogh MC, Mennella TA, Sawa C, Berthelet S, Krogan NJ, Wolek A, Podolny V, Carpenter LR, Greenblatt JF, Baetz K, Buratowski S. The Saccharomyces cerevisiae histone H2A variant Htz1 is acetylated by NuA4. *Genes Dev* 2006;20:660-5.

59. Morrison AJ, Shen X. Chromatin remodelling beyond transcription: the INO80 and SWR1 complexes. *Nat Rev Mol Cell Biol* 2009;10:373-84.

60. van Attikum H, Fritsch O, Gasser SM. Distinct roles for SWR1 and INO80 chromatin remodeling complexes at chromosomal double-strand breaks. *Embo J* 2007;26:4113-25.

61. Morrison AJ, Kim JA, Person MD, Highland J, Xiao J, Wehr TS, Hensley S, Bao Y, Shen J, Collins SR, Weissman JS, Delrow J, Krogan NJ, Haber JE, Shen X. Mec1/Tel1 phosphorylation of the INO80 chromatin remodeling complex influences DNA damage checkpoint responses. *Cell* 2007;130:499-511.

62. Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, Lopez E, Kyle S, Meuth M, Curtin NJ, Helleday T. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* 2005;434:913-7.

63. Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, Collins SR, Qu H, Shales M, Park HO, Hayles J, Hoe KL, Kim DU, Ideker T, Grewal SI, Weissman JS, Krogan NJ. Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* 2008;322:405-10.

64. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* 2007;6:439-50.

Chapter 7. Conclusion

With the emergence of physical and genetic interactions maps for a variety of species, my work points to novel ways in which these maps can be constructed and used to create biological models which can be used to generate hypotheses which can subsequently be tested in the laboratory. These methods include comparing networks across species, generating maps of pathways and their inter-relationships, and understanding the role of perturbations to these networks for pathway discovery.

In chapter 2 I demonstrated a method for pathway mapping using the yeast two-hybrid (Y2H) system. While Y2H screening is the most popular method for creating interaction maps, other methods such as mammalian two-hybrid[1] and TAP-MS[2] are also emerging methods. Each experimental method can be better suited for different types of interactions. For example, mammalian two-hybrid is reported to uncover molecular interactions that are dependent on the specific cellular environment in mammalian cells and TAP-MS techniques are better suited to uncover stable stoichiometric interactions such as those which exist in protein complexes. With more and more screening technologies better techniques will be required to synthesize the flood of data into coherent pathways.

While genetic interactions I have described are limited to two yeast species, there is currently much interest in adapting these approached to multi-cellular organisms such as C. *elegans* and even humans. These are primarily via high-throughput siRNA screening where pairs of genes can be knocked-down and their effects on cellular phenotype and development can be measured. These approaches will surely use the lessons learned mapping pathways in yeast to develop siRNA screening into a predictive biological tool for identifying genetic interactions.

Finally, the conditional genetic interaction screen in chapter 6 points to a general framework for the discovery of drug-induced pathways in a variety of species. These maps have the potential not only discover numerous biological pathways as I have shown but also point to novel drug targets in fighting diseases such as cancer. The primary treatment of most cancers is to kill proliferating cancer cells through DNA damage. However, the efficacy of DNA damaging agents in humans is limited by their toxicity to normal tissue. Accordingly, there has been significant interest in development of DNA-damage sensitizers which act specifically on cancer cells via synthetic lethal interactions[3]. A number of sensitizers have been or are currently being investigated. Most notably, much attention has been given to a new class of sensitizers known as PARP inhibitors[4]. These drugs target a gene involved in the BER pathway, which is synthetic lethal with HR pathway genes such as BRCA1 and BRCA2 which accrue cancer mutations. In effect, one potential use of the conditional genetic interaction data is to identify and target proteins encoded by genes that are synthetic lethal with cancer-causing mutations.

With the emergence of many high-throughput techniques for mapping protein interactions it is expected that new paradigms for the understanding of cellular function will be established in the future. The utility of these techniques is firmly based on their ability to generate models and hypotheses of biological function which can ultimately be tested experimentally. My work presents what I believe are the initial steps in synthesizing vast amounts of physical and genetic network data to create such models.

Bibliography

1. Luo Y, Batalao A, Zhou H, Zhu L. Mammalian two-hybrid system: a complementary approach to the yeast two-hybrid system. *Biotechniques* 1997;22:350-2.

2. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;415:141-7.

3. Michod D, Widmann C. DNA-damage sensitizers: potential new therapeutical tools to improve chemotherapy. *Crit Rev Oncol Hematol* 2007;63:160-71.

4. Farmer H, McCabe N, Lord CJ, Tutt AN, Johnson DA, Richardson TB, Santarosa M, Dillon KJ, Hickson I, Knights C, Martin NM, Jackson SP, Smith GC, Ashworth A. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* 2005;434:917-21.