UNIVERSITY OF CALIFORNIA, SAN DIEGO

Dissecting Genotype-Phenotype Relationships Through Integration and Analysis of Differential Genetic Interaction Maps

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

in

Bioengineering

by

Rohith Kannappan Srivas

Committee in charge:

Professor Trey Ideker, Chair
Professor Sumit Chanda
Professor Amy Kiger
Professor Richard Kolodner
Professor Bing Ren
Professor Kun Zhang

2012

The Dissertation of Rohith Kannappan Srivas is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2012

DEDICATION

To my parents and family for their unwavering belief in me, I dedicate this thesis.

EPIGRAPH

This having learnt, thou hast attaind the summe
Of wisdom; hope no higher, though all the Starrs
Thou knewst by name, and all th' ethereal Powers,
All secrets of the deep, all Natures works,
Or works of God in Heav'n, Air, Earth, or Sea,
And all the riches of this World enjoydst,
And all the rule, one Empire; onely add
Deeds to thy knowledge answerable, add Faith,
Add Vertue, Patience, Temperance, add Love,
By name to come call'd Charitie, the soul
Of all the rest: then wilt thou not be loath
To leave this Paradise, but shalt possess
A Paradise within thee, happier farr.

Paradise Lost, John Milton

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to acknowledge the support of my thesis advisor, Trey Ideker, without whom none of this work would have been possible. His enthusiasm and his unwavering faith in my abilities guided me through many of the rough patches a Ph.D. student faces. More over, from the first time we met, he always treated me as an equal, giving thorough consideration to all my ideas (even when they, perhaps, did not deserve it!). I sincerely hope this is a trait I will be able to carry with me into my future endeavors.

I would also like to thank many of my fellow members of the Ideker lab for their support, insights, and friendships which have been a source of comfort and inspiration through my graduate studies. In particular I would like to thank Dr. Gregory Hannum for an incredibly rewarding and fruitful collaboration which resulted in two publications. I would like to thank Rob DeConde for his encyclopedic knowledge of statistics and his willingness to share this with everyone around him. I would like to thank Menzies Chen for his offbeat sense of humor, which made coming into lab everyday a little less burdensome, as well as for his sharp insights into yeast experimental biology.

Finally, I would like to thank many of my collaborators without whom, much of this work would not have been possible. In particular, Aude Guénolé and Dr. Haico van Attikum for an incredibly rewarding collaborative effort as well as for deeply expanding my knowledge of the yeast DNA damage response.

Chapter 2, in full, is a reprint of the material as it appears in: “Srivas R*, Hannum G*, Ruscheinski J, Ono K, Wang PL, Smoot M, Ideker T. *Assembling global maps of cellular function through integrative analysis of physical and genetic networks*. Nat. Protoc. 6(9) (2011)”. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of a manuscript currently in revision: “Guénolé A*, Srivas R*, Vreeken K, Licon K, Wang Z.Z, Wang S, Krogan NJ., Ideker T., van Attikum H. *Dissection of DNA Damage Response Pathways using a Multi-Conditional Genetic Interaction Map*. Mol. Cell. In revision”. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in: “Hannum, G*, Srivas, R*, Guenole, A., van Attikum, H., Krogan, N.J., Karp, R.M., Ideker, T. *Genome-wide association data reveal a global map of genetic interactions among protein complexes*. PLoS Genetics 5(12):e1000782 (2009)”. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of a manuscript currently in submission: “Huang Y., Srivas R., Guénolé A, van Attikum H, Krogan NJ, Ideker T, Przytycka TM. *Allele specific compatibility of locus interactions underlying yeast DNA repair phenotypes*. Genome Biology. In submission”. The dissertation author was the second author of this paper, responsible for the generation of all experimental data.

VITA

2006	Bachelor of Science Bioengineering	University of California, San Diego
2010	Master of Science Bioengineering	University of California, San Diego
2012	Doctor of Philosophy Bioengineering	University of California, San Diego

PUBLICATIONS

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, **Srivas R**, Palsson BØ. *Global reconstruction of the human metabolic network based on genomic and bibliomic data*. Proc Natl Acad Sci U S A. 2007 Feb 6;104(6):1777-82.

Hannum G*, **Srivas R***, Guénolé A, van Attikum H, Krogan NJ, Karp RM, Ideker T. *Genome-wide association data reveal a global map of genetic interactions among protein complexes*. PLoS Genet. 2009 Dec;5(12):e1000782. Epub 2009 Dec 24. (*Equal contribution)

Srivas R*, Hannum G*, Ruscheinski J, Ono K, Wang PL, Smoot M, Ideker T. *Assembling global maps of cellular function through integrative analysis of physical and genetic networks*. Nat Protoc. 2011 Aug 11;6(9):1308-23. doi: 10.1038/nprot.2011.368.
(*Equal contribution)

Guénolé A*, **Srivas R***, Vreeken K, Licon K, Wang Z.Z, Wang S, Krogan NJ., Ideker T., van Attikum H. *Dissection of DNA Damage Response Pathways using a Multi Conditional Genetic Interaction Map*. Mol. Cell. (In revision)
(*Equal contribution)

Srivas R, Malta E, Sarkar S, McHugh P, van Attikum H, Ideker T. *A UV-Induced Epistasis Map Links the Chromatin Remodeling RSC Complex to Nucleotide Excision Repair*. (In preparation)

Huang Y., **Srivas R.**, Guénolé A, van Attikum H, Krogan NJ, Ideker T, Przytycka TM. *Allele specific compatibility of locus interactions underlying yeast DNA repair phenotypes*. Genome Biology. (In submission)

Choi S, **Srivas R**, Hood BL, Dost B, Van Houten B, Bandeira N, Conrads TP, Ideker T., Bakkenist CJ. *Quantitative proteomics reveals ATM kinase-dependent exchange in DNA damage response complexes*. J Proteome Res. 2012 Aug 21.

FIELDS OF STUDY

Major Field: Bioengineering

Studies in Bioinformatics and Yeast DNA Damage Response
Professor Trey Ideker

ABSTRACT OF THE DISSERTATION

Dissecting Genotype-Phenotype Relationships Through
Integration and Analysis of Differential Genetic Interaction Maps

by

Rohith Kannappan Srivas

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2012

Professor Trey Ideker, Chair

Epistasis, refers to the phenomenon, in which the phenotypic effect of one gene depends on or is modified by a secondary gene. High-throughput screening of genetic interactions has been made possible through a variety of methods such as Synthetic Genetic Array, combinatorial RNAi and genome-wide association studies.

However, thus far the majority of data has been generated in standard laboratory conditions. Yet in the course of their lives, cells are exposed to a wide-array of environmental stresses. How genetic interaction networks are re-wired in response to such stimuli remains an open question. In this thesis, I describe the generation and analysis of differential genetic interaction data, in response to numerous genotoxic stresses and demonstrate how this data can be used to elucidate cellular pathways required for the response to these stresses.

In Chapter 2, I describe the development of computational and visualization algorithms designed to integrate physical and differential genetic interaction data. This integrative approach enables the automatic assembly of raw interactions into pathway models and maps the higher-order functional relationships between such pathways.

In Chapter 3, I map changes in the cell's genetic network across a panel of mechanistically distinct DNA-damaging agents. This multi-conditional genetic interaction map identifies both agent-specific and general DNA damage response pathways. More over, we anticipate that this data will be an important resource for the study of the DDR and its associated diseases.

In Chapters 4 and 5, I describe our efforts to analyze genetic interactions derived from forward genetic screening approaches, such as genome-wide association studies (GWAS). We develop a novel computational algorithm, which greatly increases our power to detect such interactions and furthermore, through projection of these genetic interactions within and across protein complexes, demonstrate that such

pathway-based interpretations of GWAS data provide novel hypothesis regarding the mechanism through which combinations of polymorphisms may affect a phenotype.

Chapter 1. Introduction

Proteins regulate and mediate most of the processes within a cell. In nearly all cases, they act in concert with other proteins as part of pathways or larger molecular assemblies such as complexes[1]. Studies in model organisms suggest that these networks of physically-interacting proteins have topological and dynamic properties that reflect biological function. Furthermore, a complete catalog of all these physical associations can expand our knowledge of the mechanistic details required for the execution of a particular phenotype. For example knowledge of kinase-substrate interactions can help us to order signaling pathways[2], while transcription factor-DNA interactions can help to elucidate the transcriptional response to various stimuli[3]. Thus, an understanding of biological processes and disease pathogenesis will require a shift from a reductionist paradigm with its emphasis on the investigation of single proteins to a more global analysis of the structure, function, and dynamics of networks of interacting proteins.

Until recently, protein interactions were mainly discovered by small-scale methods such as co-immunoprecipitation, fluorescence correlation spectroscopy, or FRET microscopy. Unfortunately, most of these methods were both time consuming and could be only be used to reveal a small fraction of the interactome. In the past decade, however, numerous technologies such as yeast two-hybrid[4] (Y2H) or tandem affinity purification followed by mass spectrometry (TAP-MS)[5] have enabled the high-throughput mapping of protein-protein interactions (Figure 1.1). As of May

2010, the BioGRID database of protein interactions houses nearly 180,000 interactions spanning 17 different species[6].

In contrast to physical interactions, genetic interactions represent functional relationships between genes, in which the phenotypic effect of one gene is modified by another[7-9]. Genetic interactions are identified by comparing the effect of single knockouts to the joint effect of a double knockout. When the measured phenotype is growth, a genetic interaction is indicated when the growth rate of a double mutant is slower than expected (e.g. synthetic sickness or lethality) or faster than expected (e.g. suppression)[7,8,10]. In yeast, the systematic and high-throughput screening of genetic interactions has been made possible through a variety of methods including Synthetic Genetic Array (SGA) analysis[7] and to a lesser extent diploid Synthetic Lethality Analysis by Microarray (dSLAM)[11]. In its simplest form, SGA 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 (Figure 1.2)[8]. Additionally, essential genes can be screened as well through the use of a method for creating hypomorphic alleles (as opposed to complete null alleles) termed DAmP (decreased abundance by mRNA perturbation)[12]. In this method, the essential gene's 3' untranslated region (UTR) is disrupted with an antibiotic resistance cassette leading to a two to ten-fold reduction in the amount of mRNA (Figure 1.3).

A more recent variant of the SGA methodology, termed E-MAP (epistatic mini-array profiles) has enabled the identification of quantitative genetic interactions. As shown in Figure 1.4 the size of each double mutant colony is measured and

assigned a quantitative score reflecting deviation from an expected colony size, which is defined as the product of each single mutant colony size[10]. This feature enables the classification of both positive (i.e., double mutant grows better than the expected colony size) and negative (i.e., double mutant grows worse than the expected colony size) genetic interactions, as well as the magnitude of the genetic interaction (i.e., how much the double mutant deviates from the expected colony size). The E-MAP platform has been used extensively to study genetic interactions amongst subsets of genes involved in chromosomal biology[9], RNA processing[13], secretory pathways[14], and signaling[15]. In all cases, having the entire quantitative profile of phenotypic changes induced by each double mutant have allowed for numerous new genes and pathways to be implicated in each of these processes.

Recently, our lab has extended this genetic interaction mapping approach to examining how interactions are re-wired in response to an external stimulus. We have dubbed this approach differential epistasis mapping or dE-MAP[16]. In the dE-MAP approach, SGA is used to measure genetic interactions under standard conditions as well as under perturbations of interest and, by comparing the resulting networks, interactions that are altered in response to perturbation can be quantitatively assessed. These 'differential' genetic interactions reveal a unique view of cellular processes and their inter-connections under specific stress conditions[17].

However, with the ever increasing quantity of data being deposited into public databases, the question arises of how to effectively organize these raw interactions into pathways and complexes so as to facilitate biological discovery? Analysis of large-

scale genetic networks generated in *Sacchromyces cervisiae* have revealed a striking orthogonality between physical and genetic networks; less than 1% of gene pairs which exhibit synthetic lethality also interact physically[7], suggesting that the two types of networks provide complementary views of the cell. Accordingly, the interpretation of genetic networks has largely proceeded via integration with physical networks. Two common models that have been used include the "between-pathway" and "within-pathway" explanations. In the, "within-pathway" model genetic interactions are found to be enriched amongst genes encoding for proteins in the same complex or pathway[7,18,19]. Such models typically include positive or alleviating genetic interactions consistent with the hypothesis that deleting pairs of genes in the same or similar biological process leads to a less severe growth defect compared to deletions in two un-related pathways (Figure 1.5A). On the other hand, in the "between-pathway" model genetic interactions are found to span pairs of protein complexes or signaling pathways, which themselves are defined as sets of genes encoding for proteins which interact physically (Figure 1.5B). Such "between-pathway" models typically encompass mostly synthetic lethal or negative genetic interactions, which are consistent with their role in connecting genes belonging to compensatory or redundant pathways[7,18-20].

In Chapter 2, I leverage these previously described network models to design an algorithm which automatically identifies both 'modules', i.e., sets of proteins whose physical and genetic interaction data matches that of known protein complexes and the higher-order functional cooperativity and redundancy between modules. More

over, we have designed a number of intuitive ways of visualizing both the modules and their relationships which should enable new functional relationships implicated in the vast flood of genetic and physical interaction data to be easily seen. In Chapter 3, I use the dE-MAP approach, along with computational methods developed in Chapter 2, to construct a large resource of genetic modules and networks induced by distinct types of DNA damage. We find that the network differences induced by each DNA damaging compound is able to distinguish DNA damage response pathways with high statistical power and can help to elucidate both agent-specific and general DNA damage response pathways.

While the above experimental techniques have been instrumental in model organisms, performing genetic interaction analysis in higher eukaryotes has been less straightforward. First, genetic screens have relied on easy-to-measure cell-based phenotypes, such as fitness in rich growth conditions. However, genetic interactions governing complex traits in humans (such as body weight, blood pressure, or incidence of disease) are difficult to study using cell-based assays and are highly condition-dependent. Second, systematically engineering a series of double gene disruptions in mammals remains technically difficult, although combinatorial RNAi knockdowns show promise in this regard[21,22].

As an alternative to engineered genetic perturbations, high-throughput genotyping and sequencing platforms have made it possible to characterize the millions of polymorphic genetic markers present in the genome. Genome-wide linkage or genome-wide association studies (GWAS) attempt to identify polymorphic markers

that have associations, ideally causal associations, with a phenotype of interest[23]. Numerous technologies are currently available for measuring upwards of 10^5 Single Nucleotide Polymorphisms (SNPs) in the human genome[24]. In addition, full genome sequencing is becoming increasingly cheaper thus promising nearly complete interrogation of all sequence variations on an unprecedented scale.

In theory, mapping epistatic interactions from forward genetic approaches, such as GWAS data is relatively simple and proceeds in a manner similar to reverse genetic approaches. The goal is to identify a pair of sequence perturbations (e.g. a pair of SNPs) which are *jointly* correlated with a particular phenotype than either perturbation alone. Figure 1.6 provides an example of this situation: two SNPs B and C are used to segregate a particular population under study, but neither SNP alone can partition the variance of phenotype D seen in this population (Figure 1.6A). However, the joint state of SNPs B and C is able to effectively partition the variance suggesting an epistatic interaction between the two sequence variations (Figure 1.6B). Numerous computational methods for assessing the significance of epistasis between sequence variations exist ranging from simple linear and logistic regression models to more sophisticated Bayesian and machine-learning frameworks.

Currently, mapping genetic interactions using GWAS faces two major challenges: a lack of statistical power for finding genotype-phenotype associations, and a lack of tools for understanding the molecular mechanisms behind the associations found to be significant[25-27]. Because billions of marker pairs must be tested, the power to detect a given epistatic interactions is diluted by the multiple

hypothesis under consideration. One solution to this problem is to initiate searches for pair-wise interactions only for markers with strong individual effects[28]. The subsequent reduction in search space greatly improves power. This technique has been used to great effect in identifying epistatic interactions amongst SNPs controlling gene expression traits in a recent linkage study done in yeast[28,29].

The second problem stems from the so-called "fine-mapping" problem; a significant locus detected in a GWAS many contain dozens of genes due to the large spacing between molecular markers. Consequently it becomes difficult to distinguish between these genes in a given locus[30,31]. A number of recent approaches have attempted to rank the candidate genes within a significant "association" locus based on independent biological information. For instance, Franke *et al.* prioritized genes in loci associated with a particular disease by scoring them based on their network proximity to other genes in these loci[32]. Aerts *et al.* formulated the "Fine Mapping" problem as a classification problem[33]. They used a machine learning approach based on a list of 11 lines of evidence, ranging from protein domains to sequence similarity, to rank genes on their characteristic similarity to known disease genes. Tu *et al.* analyzed eQTL data by modeling the association between locus and phenotype (in this case the expression of a target gene) as a random walk through a protein network[34]. Each random walk originated from the target gene and terminated when it arrived at one of the candidate genes in the locus. The candidate gene that was visited most often was predicted to be the true causal gene for the phenotype.

In Chapters 5 and 6, I describe computational methods for boosting the power to detect epistatic interactions from GWAS data. Our method improves both power and interpretability of such genetic interactions through: (i) exploiting linkage disequilibrium between adjacent genetic markers (i.e., markers which share much of the same information) to identify groups of marker-marker interactions that fall across common genomic intervals, and (ii) projection of such interval-interval interactions within and across protein complexes, thereby elevating the analysis of GWAS from the level of individual markers to global maps of genetic interactions amongst protein complexes.

Together, this thesis reveals how genetic networks are substantially re-wired in response to various stimuli and that genetic networks derived from forward genetic approaches are amenable to similar types of analysis as those networks derived from reverse genetic approaches. It is my sincere hope that both the analytical and experimental approaches established in this thesis will help to guide the analysis of genetic interactions in humans which are believed to underlie numerous diseases that are currently afflicting society[35,36].

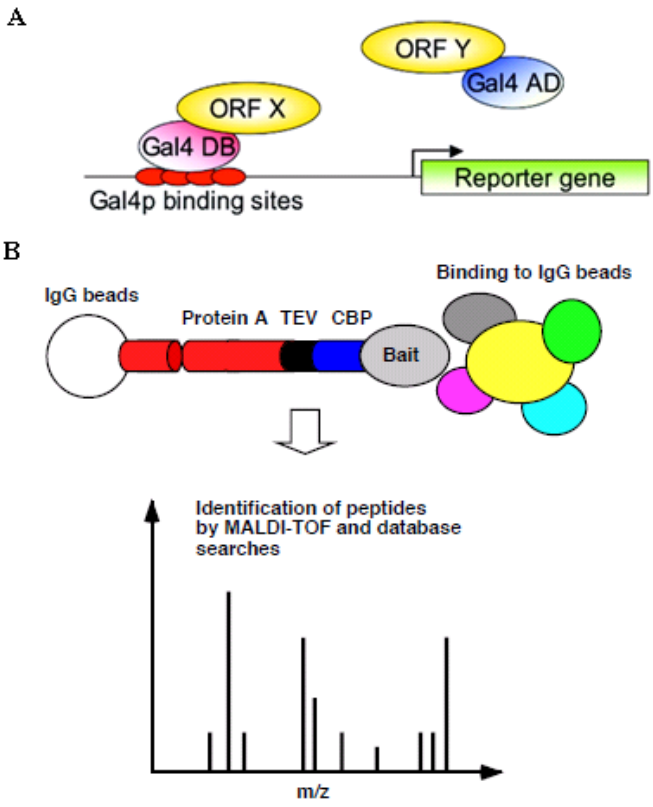


Figure 1.1: Schematic of two high-throughput methods for identifying protein-protein interactions.

(A) The yeast two-hybrid system (Y2H) consists of a separable DNA binding domain from a transcriptional activator fused to one protein and an activation domain from the same transcription factor fused to a second protein. An X-Y protein-protein interaction reconstitutes the transcription factor leading to the transcription of the reporter gene. (B) In the TAG-MS/MS system, a bait protein is fused to particular TAG system (e.g. TAP or FLAG). The bait protein is then purified using an antibody capture system; any interacting proteins are then identified using mass spectrometry. (Adapted from Cusick *et al*[1] and Suter *et al.*[37]).

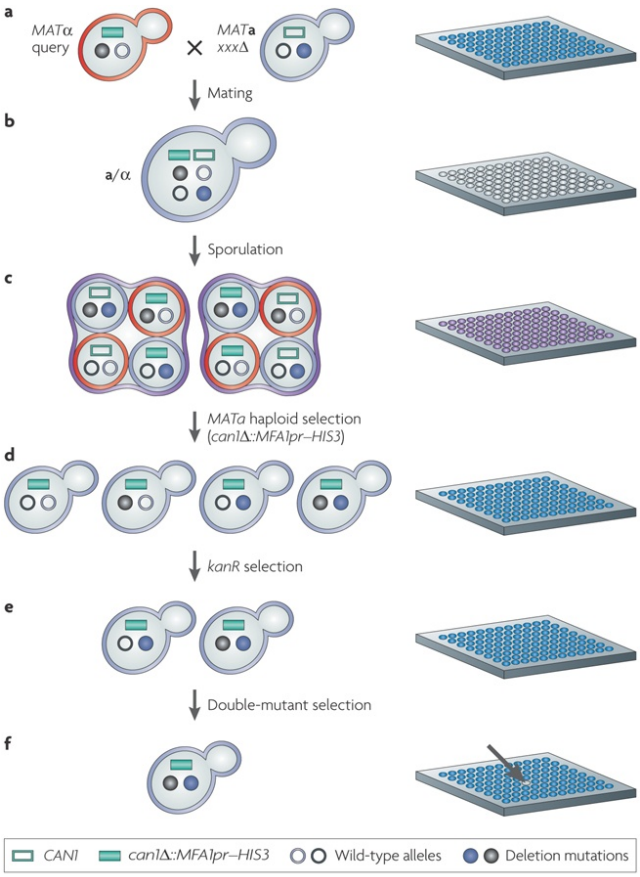


Figure 1.2: SGA Methodology.

(a) A MAT α strain carries a query mutation linked to a dominant selectable markers (e.g. nourseothricin-resistance marker *natMX*), and the SGA haploid selection marker *can1*Δ*::MFA1pr-HIS3*. This query strain is crossed to an ordered array of MATa deletion mutants (harboring a different resistance markers such as *kanMX*). (b) Resultant heterozygous diploids are transferred to a medium with reduced carbon and nitrogen to induce sporulation which leads to haploid progeny. (d-f) Through a series of selection steps, haploid double mutants are finally produced (adapted from Boone *et al.*[8]).

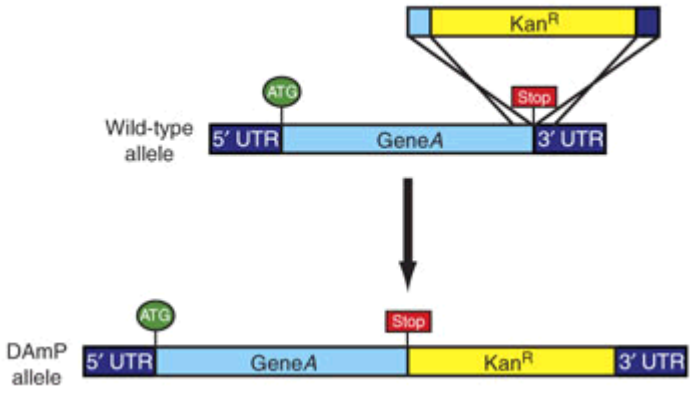


Figure 1.3: DAmP scheme for creating hypomorphic alleles of essential genes.

Using homolgous recombination an antibiotic resistance cassette is integrated downstream of the essential gene leading to the destabilization of the transcript. (adapted from Breslow *et al.*[12]).

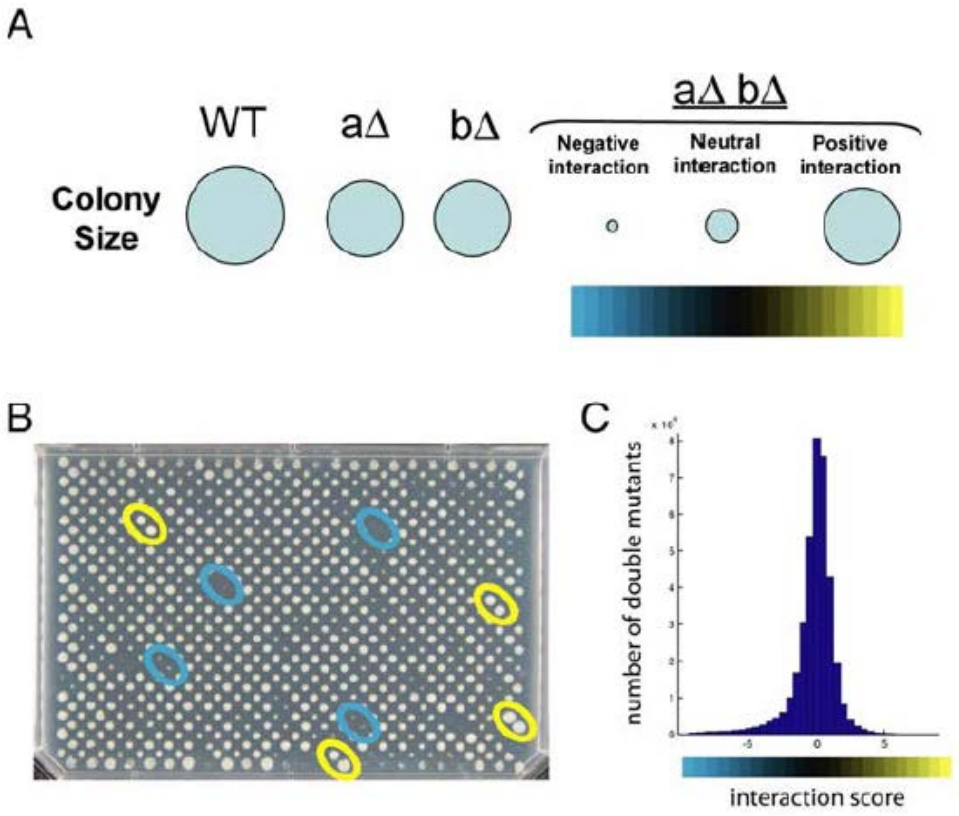


Figure 1.4: E-MAP Enables Measurement of the Continuous Spectrum of Genetic Interactions.

(a) Colony sizes of the final double mutants are measured and compared to an expected colony size (defined as the product of each single mutant colony size). A genetic interaction score is assigned to each double mutant representing both the direction & size of deviation from the expected colony size. (B) An array of double mutants; yellow and blue circles show, respectively, replicates of double mutants which are growing faster than expected (positive interaction) or worse than expected (negative interaction). (C) E-MAP platform allows every single double mutant to be assigned a score allowing for full spectrum of genetic interactions to be analyzed. (Credit to Sourav Bandyophadyay).

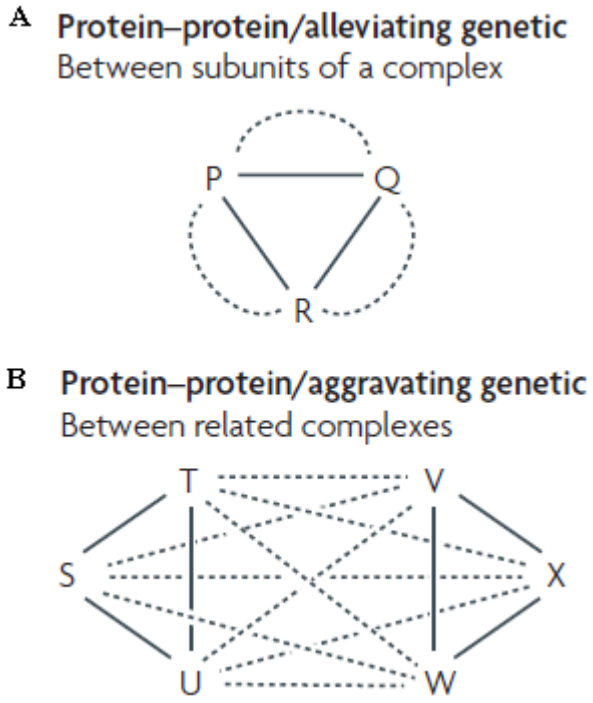


Figure 1.5: Common pathway interpretations of genetic and physical interactions.

(A) An example of a "within-pathway" model. Proteins are connected by higher than expected number of protein and genetic-interactions. An example would be members of a multimeric complex. (B) An example of a "between-pathway" model. Two sets of proteins are connected a higher than expected number of protein interactions. The two sets are spanned by a large number of genetic interactions. An example would be two complexes fulfilling redundant roles. (Adapted from Beyer *et al.*[38])

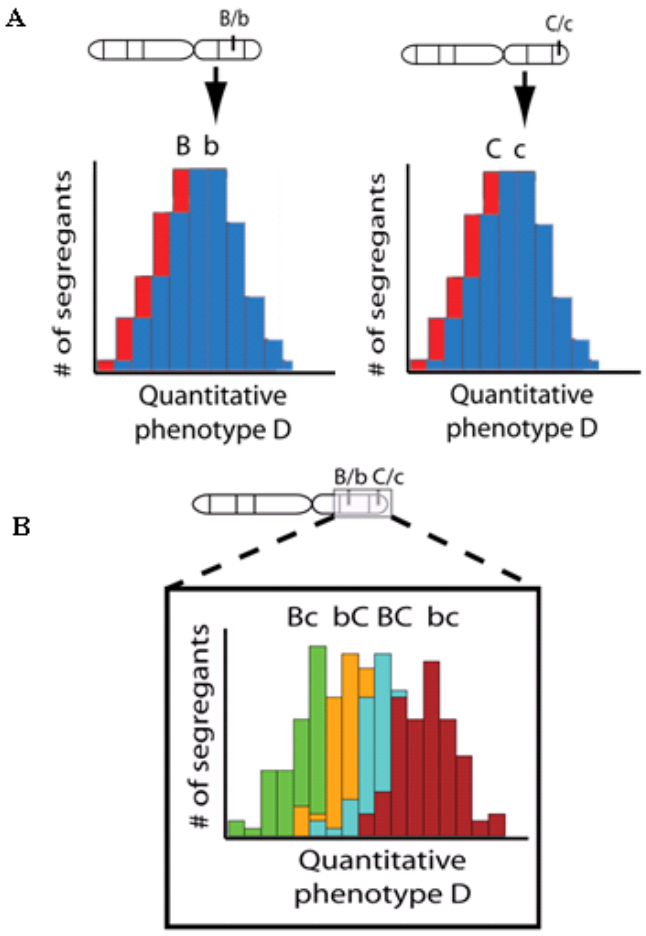


Figure 1.6: Mapping epistatic interactions from forward genetic approaches.

(A) An epistatic interaction between two sequence variants is identified when neither variant alone can segregate the variation of quantitative trait seen in a population , (B) while the joint state of the two markers is predictive.

Chapter 2. Assembling Global Maps of Cellular Function through Integrative Analysis of Physical and Genetic Networks

Chapter 2.1: Abstract

To take full advantage of large-scale genetic and physical interaction mapping projects, the enormous amount of raw data must first be assembled into models of cell structure and function. PanGIA (Physical and Genetic Interaction Alignment) is a plugin for the bioinformatics platform Cytoscape, designed to integrate physical and genetic interactions into hierarchical module maps. PanGIA identifies 'modules' as sets of proteins whose physical and genetic interaction data matches that of known protein complexes. Higher-order functional cooperativity and redundancy is identified by enrichment for genetic interactions across modules. This protocol begins with importing interaction networks into Cytoscape, followed by filtering and basic network visualization. Next, PanGIA is used to infer a set of modules and their functional inter-relationships. This module map is visualized in a number of intuitive ways, and modules are tested for functional enrichment and overlap with known complexes. The full protocol can be completed between 10 minutes and 30 minutes depending on the size of the dataset being analyzed.

Chapter 2.2: Introduction

Genetic interactions are defined as functional relationships between genes that result when the phenotypic effect of one gene is altered by one or several other genes[8,38]. Such interactions have been used to uncover pathway architecture in model

organisms[7,9,14,15]. In humans, genetic interactions are thought to influence numerous phenotypes of interest from expression[39] to complex diseases[23] to drug resistance[40]. Recently a number of technologies such as synthetic genetic arrays (SGA)[7,10,41,42] and heterozygote diploid-based synthetic lethality analysis with microarray (dSLAM)[11] have facilitated the rapid screening of genetic interactions in model organisms. In human cell lines, combinatorial RNAi screening technologies have begun to show promise in uncovering genetic interactions[43,44]. As a result of these high-throughput technologies, the amount of genetic interaction data available in the public domain has increased rapidly. As of December 2010, the BioGRID interaction database houses nearly 175,000 genetic interactions spanning 11 different species[6].

Interpreting the functional significance of each genetic interaction remains a daunting task. One promising solution has been to interpret genetic interactions in the context of their relationships to physical protein-protein interactions (Figure 2.1A)[18-20,45]. At least two distinct models have been put forth to reconcile genetic and physical interactions. The "within-cluster" model seeks to identify clusters of proteins that are enriched for both physical and genetic interactions (Figure 2.1B). We refer to such clusters of proteins and the interactions occurring among them as a module. Modules are often interpreted as functional protein complexes[7,19,20,45] or signaling pathways[40]. In contrast, the "between-cluster" model seeks genetic interactions that are enriched across two clusters of interacting proteins (Figure 2.1B). Such inter-module links have been shown to identify synergistic or compensatory relationships between protein complexes or signaling pathways[9,18,19]. Figure 2.1C shows an example module

map consisting of four modules connected by three inter-module links. The genes in each of these four modules are associated with a strong "within-cluster" signal and furthermore coincide with known *Sacchromyces cerevisiae* physical complexes (Figure 2.1C). Setp3p and Rpd3s are both histone deactylase complexes involved in transcriptional regulation. The Hir complex functions in replication-independent nucleosome assembly, while the UTP-C complex is a component of the 90S pre-ribosome. The inter-module link between Set3p and Rpd3s suggests a functional synergy between the two complexes. Consistent with this hypothesis, several studies have illustrated that the two cooperate in the activation of DNA damage genes through recruitment of RNA polymerase II[46].

Several methods have been previously published[19,20,45,47] for analyzing interactions to identify both within-cluster and between-cluster functional organization. However, these methods have not yet been made available through a publicly-accessible software package. Here, we introduce a novel software tool, PanGIA (Physical and Genetic Interaction Alignment), along with a general bioinformatics protocol for integrative analysis of genetic interactions. PanGIA implements a previously published framework[18] as a plugin for the open-source network analysis platform, Cytoscape[48,49], and allows the user to easily generate maps of modules and module inter-relationships from genetic and physical interaction data (see Figure 2.1 for an overview). A number of options are available to the user for constructing and visualizing the resulting module map. PanGIA is built on the new Cytoscape 2.8 architecture[50] which features the ability to view and manipulate nested

networks, enabling the user to explore both the global map as well as individual modules in an intuitive manner. Finally, individual modules can be interrogated using a number of functional enrichment options.

The computational workflow presented here has been used in the analysis of genetic networks centered on genes involved in chromosomal biology[9,18], RNA processing[13], secretory pathways[15], and DNA-damage response[40]. This analysis has also been employed in comparing genetic networks across two different species[51]. In each case, the module maps generated have helped to identify novel pathways as well as new components and functions for existing complexes[18-20,40,51]. While this workflow has proven useful in the analysis of numerous genetic interaction datasets, the module search process works best when there is a high density of protein and genetic interactions among the set of genes being studied. For species for which there is a scarcity of either genetic interaction or physical interaction data, this protocol may not identify a significant number of modules or inter-module relationships. This limitation will become less relevant as large-scale interaction screens continue to populate the scientific databases.

This protocol is divided into five basic sections (Figure 2.1). The first section, ‘Importing Physical and Genetic Networks into Cytoscape’ describes the available sources of interaction data and means of acquiring these data within Cytoscape. Second, ‘Generating a Module Map Using the PanGIA Plugin’ covers the usage of the PanGIA plugin and is further broken into four sub-sections covering the various aspects of its use (‘Selecting a Physical and Genetic Network’, ‘Setting the Module

Size and Edge Reporting Parameters', 'Training PanGIA', and finally, 'Labeling Modules'). The third section, 'Visualization of the Module Map Using Nested Networks' introduces ways in which the user can navigate and visualize the resulting module map. Fourth, 'Functional Enrichment of the Modules' illustrates methods to identify enriched biological functions and pathways among the identified modules. Finally, 'Exporting the Results' covers the various ways in which the module map can be exported from Cytoscape for further analysis or for inclusion as figures in a publication. We now briefly describe each of the sections comprising this protocol:

The first section ('Importing Physical and Genetic Networks Into Cytoscape') describes the various ways in which a physical or genetic network can be imported for analysis into Cytoscape. A previous protocol has outlined in detail the various file formats Cytoscape can recognize as well as provided detailed instructions on how each file type can be imported[49]. The present protocol will instead focus on importing networks in a tab-delimited format. Table 2.1 provides examples of several different databases from which interaction data (both genetic and physical) can be downloaded in a tab-delimited format for over 50 organisms.

The second section ('Generating a Module Map Using the PanGIA Plugin – Selecting a Physical and Genetic Network') describes the steps necessary to select which physical and genetic networks are to be analyzed. At this point, PanGIA is fully configured and the module search process can be initiated. However, PanGIA is designed with four optional features designed to fine-tune and enhance the search process. We describe these optional features in the subsequent sections.

The first optional feature is the ‘Module Size’ parameter. This parameter helps to control both the size and number of modules by rewarding the formation of larger modules. Thus, higher values of this parameter results in the formation of larger, but fewer modules. Lower values produce the opposite effect (Figure 2.1B). It is recommended that the ‘Module Size’ parameter initially be left at the default value. If the resulting module map contains very large modules, the ‘Module Size’ parameter can be suitably altered and the module search process re-run to produce smaller and more biologically meaningful modules (see Troubleshooting).

The second optional feature is dependent on the presence of quantitative genetic interaction data. Many of the recent experimental technologies for measuring genetic interactions go beyond reporting interactions in a simple binary format (interacting or non-interacting) and provide some measure of confidence in a given interaction. For example, in the SGA technology[42] and a recent variant called E-MAP (epistatic mini-array profiles)[9,10], each double mutant is assigned a quantitative signed score, where positive scores indicate that the double mutant grew better than expected (e.g. suppression) and negative scores indicate pairs for which the double mutant grew worse than expected (e.g. synthetic sick or synthetic lethal)[10,42]. Table 2.1 outlines numerous databases that contain quantitative interaction data..

If quantitative genetic interaction data is provided, each inter-module link can be assessed for significance. A p-value is assigned by comparing the sum of the interaction confidence values for all genetic interactions spanning two modules (i.e., inter-module link) to a distribution of sums of equal interaction confidence values

drawn from random genetic interactions[18] (Figure 2.1C). Thus, inter-module links consisting of very confident genetic interactions will be assigned a significant p-value. The 'Edge Reporting' parameter serves as a threshold used to filter insignificant inter-module links from the final module map. By default, this parameter is set to 0.1, thus filtering inter-module links with a p-value above 0.1 from the final module map.

The next optional feature relies on the presence of a biological annotation set. Examples of an annotation set that can be used include physical complexes, signaling pathways, metabolic pathways, or even broad biological processes. Table 2.2 provides a list of databases where an annotation set can be downloaded for a range of different organisms.

The optional training procedure built into PanGIA is designed to help identify modules which are more likely to be biologically relevant, i.e., modules which contain genes that operate in the same complex or biological process. By default, the module search process is designed to identify sets of genes that are densely connected by physical and genetic interactions. However, some interactions can be given more or less influence based on their quantitative score. PanGIA can determine how likely a certain interaction (either physical or genetic) is to connect two genes within a known complex or biological process using an existing annotation set. Examples of such a set include physical complexes (e.g. INO80 complex), signaling pathways (e.g., the MAPK pathway), metabolic pathways (e.g., glycolysis), or biological processes (e.g. DNA Damage Response Genes). Using this annotation set, PaNGIA assigns each interaction a weight based on the unsigned logistic regression of all interaction

confidence scores of a given type (physical, genetic) against its proteins' co-membership in an annotation. If no quantitative scores are available, PanGIA uses logistic regression to assign a constant confidence score for all interactions of a given type. For specific details regarding the regression procedure please see Bandyophadyay *et al*[18]. The module search process will now seek to identify sets of genes which are connected by highly weighted physical and genetic interactions. Since the weight of an interaction corresponds to how likely it is to connect two genes belonging to the same physical complex or pathways, the modules identified will contain genes which are functionally similar.

The genes comprising a module may function in the same biological process or encode members of the same protein complex. If a biological annotation set is provided, PanGIA will check to see if the set of genes comprising each module overlaps with the set of genes comprising each annotation. Here, overlap is defined using the Jaccard similarity coefficient (intersection/union) which ranges from 0 (no overlap) to 1 (perfect overlap). If the Jaccard coefficient exceeds a user-specified threshold, then the module will labeled with the name of the annotation in the final module map (Figure 2.1C). This sub-section ('Generating a Module Map Using the PanGIA Plugin – Labeling Modules') covers how this labeling feature can be enabled and provides instructions on how to set the overlap threshold.

PanGIA is built on the new Cytoscape 2.8 architecture which features the ability to view nested networks, i.e., each node in a network can represent an entire sub-network. Instructions are provided for laying out the network of modules and

inter-module links and for probing individual modules. This section is broken into three sub-sections, ‘Navigating the Module Map’, ‘Finding Modules of Interest’, and ‘Exploring Modules of Interest’, which cover the various ways in which both the module map and individual modules can be interrogated.

Modules will often contain genes of unknown function. One way to dissect the function of modules uncovered in this workflow is to examine if they are significantly enriched for any functional annotations. This can be used to identify new components of existing complexes or to identify entirely new physical complexes or pathways[9,18,19]. This section (‘Functional Enrichment of the Modules’) outlines the steps for checking for enriched Gene Ontology functional terms[52] using the BiNGO plugin[53].

The final section (‘Exporting your Results’) covers the various options for exporting the resulting module map.

Chapter 2.3: Materials

Equipment

PC with Internet access and an Internet Browser.

Equipment Setup

Hardware Requirements: PanGIA hardware requirements depend on the size of the physical and genetic networks to be imported and analyzed. For networks up to 200,000 edges, we recommend a 2.0 GHz CPU or higher, a medium-end graphics card, 150 MB of available hard disk space, and at least 2

GB of free physical RAM. If analyzing very large networks (>500,000 interactions), at least 8 GB of free physical RAM is recommended. For viewing of the modular map produced by PanGIA we recommend a monitor with a minimum screen resolution of 1024 x 768.

Operating System: PanGIA and Cytoscape are supported on Windows (XP, Vista, and Windows 7), Mac OS X (version 10.6 [i.e., Snow Leopard] or higher) and Linux.

Java Standard Edition: version 1.6 or higher (can be downloaded from http://www.java.com).

A three button-mouse: This is recommended (but not required) as an aid in navigating the module map.

Cytoscape v2.8.0: PanGIA requires Cytoscape version 2.8.0 or higher. The steps for downloading and installing the latest version of Cytoscape can be found in a previously published protocol[49] or online at http://www.cytoscape.org/documentation_users.html.

Plugins: The analysis capabilities of Cytoscape are expandable and extensible through add-on software packages called plugins. This protocol requires the installation of four plugins, PanGIA, BiNGO[53], Enhanced Search[54], and CyThesaurus[55]. Instructions for installing these plugins are outlined in Steps 2–4.

MeV version 4.6 or higher: MeV or MultiExperiment Viewer[56] is an integrated toolkit that that includes sophisticated algorithms for the clustering and

visualization of large-scale genomic data. This protocol uses MeV to view modules as a hierarchically clustered heat map. Instructions for downloading and installing MeV can be found at http://www.tm4.org/mev/.

Data files: PanGIA requires both a physical and genetic network in a tab-delimited format. Sample protein and genetic interaction networks are provided as examples to illustrate the protocol. The physical interaction network (Supplementary Data 1) was taken from a recent computational integration of two large datasets generated using tandem affinity purification followed by mass spectrometry (TAP-MS)[57]. Each physical interaction was assigned a Purification Enrichment score (PE Score), with larger values representing greater confidence in the physical interaction. The genetic interaction network (Supplementary Data 2) was obtained from a large E-MAP screen which measured all possible genetic interactions among 743 genes involved in yeast chromosomal biology[9]. Each gene pair is assigned an S-score representing both the magnitude and confidence of the genetic interaction. The supplementary information can be accessed on the supplementary website (http://prosecco.ucsd.edu/PanGIA/). Table 2.1 lists several public databases where protein and genetic interaction data can be downloaded for many different species.

Additional data files: The file *CYC2008_Complexes.txt* contains a list of 408 protein complexes in the yeast *Saccharomyces cerevisiae* hosted by the CYC2008 database[58,59]. This file illustrates an example of a Cytoscape node attribute file, which

allows nodes in a network to be mapped to a particular. In this case, yeast genes are mapped to the various physical complexes in which they participate. This file is used to demonstrate how a set of known biological modules can be used to train PanGIA to identify more biologically meaningful modules and inter-module relationships (covered in the 'Training PanGIA' sub-section). Additionally, this file is used during the 'Module Labeling' portion of this protocol to check if the identified modules correspond to known protein complexes. Table 2.2 outlines several different public databases from which an annotation set can be downloaded for a variety of species.

Chapter 2.4: Protocol

Chapter 2.4.1: Importing physical and genetic networks into Cytoscape

1. Start Cytoscape. If Cytoscape is not yet installed on your computer, instructions for downloading and installing the latest version can be found at http://www.cytoscape.org/documentation_users.html. Cytoscape can be started by navigating to the directory in which it was installed and executing the file *cytoscape.bat* (Windows users) or *cytoscape.sh* (Linux and Mac OS X users). Critical step. PanGIA requires Cytoscape version 2.8.0 or higher. If your current installation of Cytoscape doesn't meet this requirement, download and install the latest version from http://www.cytoscape.org
2. Next install the required plugins by navigating to the Plugins menu and clicking on 'Manage Plugins'.

3. Double click on the folder titled Analysis located under the 'Available for Install' folder and select the plugin PanGIA version 1.1 or later. Click Install. Accept the Plugin License Agreement and then click Finish.
4. Repeat the above step with BiNGO[53] version 2.42 or later (located in the Functional Enrichment Folder), EnhancedSearch[54] version 1.2 or later (located in the Analysis Folder) and CyThesaurus version 1.2 or later (located in the Network and Attribute I/O Folder).
5. After installing the required plugins, start the PanGIA plugin by navigating to the Plugins menu and selecting Module Finders → PanGIA.
6. After PanGIA has started the PanGIA console will appear (Figure 2.3). The console is divided into three main panels: the Physical Network panel, where details regarding the physical network will be entered, the Genetic Network Panel, where details regarding the genetic network will be entered, and the Advanced Options Panel, which can be expanded by clicking on the triangle located next to the word 'Advanced'. This panel contains multiple advanced options for tuning the module-finding process. Four additional areas of interest are the Cytoscape canvas which displays network visualizations and may be initially blank, the Data panel which is used to display node, edge, and network attribute data, the Toolbar which contains numerous command buttons, and the Network Browser which can be accessed by clicking on the tab titled "Network" (Figure 2.3). The Network Browser provides a list of networks currently available along with the number of nodes and edges in each network.

7. Next we import both a physical and a genetic network to be used in the analysis. Assemble the data in a tab-delimited format. Users wishing to follow this protocol as a tutorial should download the Supplementary Data File 1 (Collins_physical_network_example.txt) and Supplementary Data File 2 (Collins_genetic_network_example.txt) and continue with Step 8. Critical step. PanGIA is designed to work with both quantitative and non-quantitative interaction data. However, any single network (either physical or genetic) must consist of a single type of interactions (i.e., either all quantitative interactions or all non-quantitative interactions).
8. Click on the File menu, then select Import → Network From Table (Text/MS Excel). A window titled 'Import Network and Edge Attributes from Table' will appear.
9. Click on the button titled 'Select File(s)' and specify the file containing the physical interaction network. A preview of the file should appear in the 'Preview' panel located at the bottom. Select the column number representing the gene which is the source node in the selection box titled 'Source Interaction'. Select the column number representing the target node in the selection box named 'Target Interaction'. If the example files (Supplementary Data 1 & 2) are being used, the source and target nodes are, respectively, columns 1 and 2.
10. Specify an interaction type which will enable Cytoscape to differentiate between protein and genetic interactions. Check the box titled 'Show Text File Import Options' and under 'Network Import Options' enter a meaningful string character

in the box titled ‘Default Interaction’ (e.g. ‘pi’ or ‘gi’ depending on whether physical interactions or genetic interactions are being imported).

11. *Optional Step:* Use this step if quantitative interaction strengths are attached to the network. In the file Preview panel launched in Step 8, left click the column which represents the quantitative attribute under the ‘Preview’ panel to enable the import of this attribute into Cytoscape. Right-click the same column and when prompted, type in an appropriate Attribute name (e.g., ‘PScore’ or ‘GScore’ depending on whether the physical or genetic network is being imported) and click ‘OK’. Make sure to note the name used. You will need it later when selecting the attribute to be used in the training process. Critical step. The quantitative attribute provided should be either an integer (e.g., numbers like 1, -2, or 514) or a floating point (e.g., numbers like 2.343, -45.7687, or 74.3).
12. Click the button ‘Import’ located in the lower right hand corner. The physical network should now appear in the Cytoscape canvas area. The title of the network should be the name of the file provided.
13. Repeat Steps 8 – 12 to import the genetic network.
14. *Optional Step:* Steps 14 – 18 should be used if the physical and genetic networks use different gene identifiers (e.g. Uniprot ID versus Ensemble ID). PanGIA requires that the two networks use the same gene identifier. To convert the gene identifiers in a given network, assemble an ID translation file into a tab-delimited format. This file should contain a map between the gene identifier currently being

used and the target gene identifier. Users following this protocol as a tutorial using the sample data provided should skip to Step 19.

15. *Optional Step:* Start the CyThesaurus plugin by clicking on the Plugins menu and then selecting CyThesaurus. A window titled ‘CyThesaurus plugin’ should appear.
16. *Optional Step:* Configure the CyThesaurus plugin to use the ID mapping file generated in Step 14 by clicking on the button ‘ID Mapping Resources Configuration’. A new window titled ‘ID Mapping Source Configuration’ will open up. In the left panel of this window click on the folder titled ‘Local Remote Files’ which will bring up another window titled ‘File-based ID Mapping Resources Configuration’. Under the panel named ‘Data source’, click the button ‘Select file’ and specify the location of the ID mapping file and click ‘Open’. Next, click ‘Okay’ and finally ‘Close’.
17. *Optional Step:* Select both the physical and genetic networks by clicking on them in the ‘Available Networks’ panel, then clicking the right arrow button. The two networks will appear in the ‘Selected Networks’ panel.
18. *Optional Step:* Choose the two different gene identifier names used in the genetic and physical network in the ‘Source ID Type(s)’ selection box. In the ‘Target ID Type’ selection box choose the target gene identifier you with to map to. Finally, in the selection box titled ‘All target ID(s) or first only?’ selection the option ‘Keep the first target ID only’. Now press OK. A message will pop up indicating how many gene identifiers were successfully mapped.

Chapter 2.4.2: Generating a Module Map Using the PanGIA Plugin – Selecting the Physical and Genetic Network

19. In the top-most panel in the PanGIA console ('Physical Network' panel, see Figure 2.3), select the physical network to be used in the 'Network' selection box. The name of the physical network will correspond to the name of the file from which the network was imported.
20. Select the genetic network to be used in the 'Network' selection box located in the 'Genetic Network' panel. Again, the name of the network will correspond to the name of file from which it was imported.
21. *Optional Step:* Use this step if quantitative interaction data are being used. In the 'Attribute' drop-down menu located in the 'Physical Network' panel, select the appropriate attribute name (i.e., the name assigned to the quantitative attribute for physical interactions from Step 11). Similarly, select the appropriate attribute name for genetic interactions in the 'Attribute' drop-down menu located in the 'Genetic Network' panel.
22. *Optional Step:* Use this step if quantitative interaction data are being used and no biological annotation data is present. Even without a set of known complexes or pathways, PanGIA can leverage the confidence values assigned to each interaction (physical or genetic) to identify modules and inter-module links that contain highly confident interactions. However, it is necessary to let PanGIA know how the quantitative information is scaled. In the 'Scale' selection menu located in both the

'Physical Network' and 'Genetic Network' sub-panels (Figure 2.3) choose one of the following options:

a) 'lower – This option indicates that smaller quantitative values (both positive and negative) represent more confident interactions.
b) 'upper' – This option indicates that larger quantitative values (both positive and negative) represent more confident interactions.
c) 'none (prescaled)' - This option should only be chosen if the quantitative attribute attached to either the physical or genetic interactions already represents the likelihood that a given interactions falls within a known biological module. This option enables the user to perform the training procedure outside of PanGIA and use the subsequent results in the module-search process.

If the example files are being used, simply choose 'none'. During the training process, PanGIA will automatically scale the score attached to each interaction to reflect how likely that interaction is to fall either 'within' a module or 'between' two modules.

23. *Optional Step:* Use this step if the gene identifiers in either the physical or genetic network were mapped to a new gene identifier. In the Advanced Optional panel, select the target gene identifier to which genes in both networks were mapped to under the 'Node Identifiers' sub-panel. If no gene identifier mapping was performed or if the user is following this protocol with the sample data, skip to Step 24.

Chapter 2.4.3: Generating a Module Map Using the PanGIA Plugin – Setting the Module Size and Edge Reporting Parameters

24. *Optional Step:* PanGIA features a number of advanced options for tuning the search process. The size and number of modules returned by the search process can be controlled by changing the 'Module Size' parameter (located in the Advanced Options panel). This can be done using the graphical slider in the 'Search Parameters' panel. Dragging the slider to the right will result in fewer modules with larger average size, while dragging the slider to the left will result in more modules with a smaller average size (Figure 2.1B). The value of the 'Module Size' parameter will be displayed in a text box to the right of the slider. It is recommended to leave the slider in its default position for the first run and to adjust it later if the results are unsatisfactory (see Troubleshooting). For the sample data provided, set the 'Module Size' parameter to -1.6 by moving the slider to the left.
25. *Optional Step:* Oftentimes, the physical network being used covers a much larger set of proteins than those examined in the genetic interaction screen. In such a case, it is often useful to trim the physical network to include only proteins which are either present in the genetic network or are neighbors of such proteins within the physical network. This trimming is controlled by setting the 'network filter degree' parameter (located in the Advanced Options panel). A value of 0 will trim the physical network to only include nodes from the genetic network. Higher values represent the acceptable distance (through edges) separating a protein in the

physical network from a node in the genetic network. If no trimming is desired, leave the box blank to prevent PanGIA from filtering any nodes. If the samples data file are being used, leave the ‘network filter degree’ parameter at its default value of two. Critical step. The ‘network filter degree’ parameter provided should be a positive integer (e.g., numbers like 1, 2, or 10).

26. *Optional Step:* Use this step only if quantitative interaction data is present. Every inter-module link found by PanGIA can be assigned a p-value, after which insignificant edges are filtered from the resulting module map. The significance threshold can be set by changing the position of the slider in the ‘Edge Reporting’ sub-panel. Dragging the slider to the left (towards ‘Less’) will result in a higher significance threshold and less inter-module links in the final map (Figure 2.1C). The p-value cutoff will be displayed in a text box immediately to the right of the slider. If the example files are being used, move the slider to the left and set the threshold to 0.05.
27. *Optional Step:* Steps 27–29 should be used only if an annotation set is present. The training and module labeling steps requires a list of annotations to be imported into Cytoscape. Assemble your list of annotations into the node attribute file format. Import this file into Cytoscape by navigating to File → Import → Node Attribute…. Navigate to the appropriate file and click ‘Open’. If using the sample data, the file *cyc2008_complexes.txt* should be used in this step.
28. *Optional Step:* In the ‘Annotation’ sub-panel under Advanced Options, select the annotation attribute that will be used during the training and labeling process. The

name of annotation set is specified in the node attribute file which was uploaded in the previous step. If the sample data have been used, the attribute name will be ‘CYC2008’. Select the annotation set name in the selection box titled ‘Annotation attribute.’

29. PanGIA can be trained to better identify module and inter-module links by examining actual examples of biological modules provided in the annotation set. To train PanGIA, simply check the box titled ‘Train PanGIA’ in the Annotation sub-panel. If the sample data is being used, make sure this box is checked.

Chapter 2.4.4: Generating a Module Map Using the PanGIA Plugin – Labeling Modules

30. *Optional Step:* This step should only be used if an annotation set is present. PanGIA can label individual modules with the name of an annotation, if their member genes overlap with the genes belonging to that annotation (Figure 2.1C). To have PanGIA label modules, check the box titled ‘Label modules’ in the Annotation sub-panel (Figure 2.3). Next, specify the overlap threshold (defined here as the Jaccard index) in the text box named ‘Labeling Threshold’. If the sample data is being used, set the ‘Labeling Threshold’ to 0.2. ?Troubleshooting
31. *Optional Step:* If desired, PanGIA can output a report containing a summary of the module-finding process. This includes: (i) a summary of the networks used by PanGIA, (ii) the results of the training process and, (iii) a summary of the resulting

module map. To have PanGIA output a report, specify an output file in the sub-panel ‘Report’. Following a successful search, an HTML file will be created which can be viewed using any internet browser.

32. At this point, PanGIA is fully configured. The module search process can be initiated by clicking the ‘Search’ button located at the bottom-right corner of the PanGIA console. Depending on the size of the network and the computer hardware, the module-finding process should take anywhere from one to ten minutes. If the sample data is being used, the search process should take less than a minute (see Timing for additional details).

Chapter 2.4.5: Visualization of the module map using nested networks – Navigating the Module Map

33. Once the search process is complete, a window titled ‘Module Overview Network’ will appear in the Cytoscape Canvas panel (Figure 2.4A). This network is the resulting global module map. Each node represents an individual module comprised of a set of genes densely interconnected by genetic and physical interactions. The area of a module scales according to the number of genes that it contains. Links between modules are comprised of genetic interactions; the thickness of the interactions corresponds to the number of genetic interactions spanning the two modules. If the labeling option was chosen, modules that overlap with one of the annotations provided will be labeled as such (Figures 2.4A – B).

34. You can zoom into the module map using the “Zoom In” button on the tool bar. This icon is displayed as a magnifying glass with a ‘+’ symbol in the middle. You can zoom out by clicking on the “Zoom Out’ button (magnifying glass with a ‘-‘ symbol in the middle) Alternatively, one can zoom in and out using the scroll wheel on the mouse. Scrolling up zooms into the area centered on the mouse pointer. Scrolling down zooms out on the area centered on the mouse pointer.
35. To pan around the module map, two options are available:

    (A) Using the mouse. Click the middle button on the mouse anywhere in the active network being viewed in the Cytoscape canvas (or the scroll wheel if present) and drag the mouse in the desired direction.

    (B) Using the network browser. Navigate to the ‘Network Browser’ by clicking on the ‘Network’ tab (Figure 2.3) located to the left of the PanGIA tab. In the bottom half of the ‘Network Browser’ is a bird’s-eye view of the active network being viewed in the Cytoscape canvas; a blue selection box highlights the particular region of the network currently being viewed. To pan around the network, click and hold the blue selection box and move it in the desired direction.

Chapter 2.4.6: Visualization of the module map using nested networks – Identifying Modules of Interest

36. To further investigate modules of interest (i.e., function enrichment or detailed visualization), the module or modules of interest must be selected. This protocol

describes three different options for doing so: direct selection of modules (option A), direct selection of inter-module links (option B), and search-based selection of modules (option C).

(A) Direct selection of modules. Select any single module by clicking on it the with the left mouse button. The selected module will turn yellow. Several modules can be selected by holding down and dragging the left mouse button to define a rectangular selection region. Alternatively, multiple modules may by holding down the ‘Shift button and left clicking on multiple modules.

(B) Direct selection of inter-module links. To select any edge, click on the edge with the left mouse button. The selected edge will turn red. Several edges can be selected by holding down and dragging the left mouse button to define a rectangular selection region.

(C) Search-based selection of modules. To find and highlight modules in the map which contain a gene of interest, enter the name of the gene into the Enhanced Search Plugin search box located in the command toolbar (the box is named ‘Enhanced Search; see Figure 2.3). If your gene of interest falls within a module, that module and its inter-module links will be highlighted yellow.

Chapter 2.4.7: Visualization of the module map using nested networks – Exploring Modules of Interest

37. PanGIA returns numerous useful statistics or attributes regarding the modules identified, including module size, number of physical/genetic interactions among

the genes in this module, etc. A complete list of attributes returned by PanGIA is provided in Table 2.3. The Data Panel (Figure 2.3) can display any/all of the attributes listed in Table 2.3. Select a module(s) of interest from the module map displayed in the Cytoscape Canvas as described in Step 33. When a single module or groups of modules have been selected in the Cytoscape Canvas, the selected modules will be listed in the Data Panel (Figure 2.3). Next, click on the Select Attributes Button located in the upper left corner of the Data Panel. This will cause a list of attributes to appear; select which attributes you wish to view by clicking on their name. Exit this menu by clicking anywhere else.

38. The Data Panel can also display detailed information regarding inter-module links in the map. Select one or more inter-module links of interest in the map as described in Step 36. In the Data Panel, click on the tab labeled 'Edge Attribute Browser'. The panel will display the edges that have been selected. Similar to the modules, inter-module links identified by PanGIA also have several informative attributes as outlined in Table 2.3. These attributes can be viewed by selecting them through the Select Attributes menu (see Step 37).
39. To visually inspect a single module or a group of modules in greater detail, select the module(s) of interest as outlined in Step 36. Next, right-click any of the selected module(s) and choose PanGIA → Create Detailed View. A new window will appear in the Cytoscape Canvas area containing the module (Figure 2.4C) or modules (Figure 2.4D) of interest. In this detailed view, each node represents a single gene. Edges represent either physical interactions (colored black) or genetic

interactions (colored turquoise). If quantitative genetic interaction data is used, positive genetic interactions will be colored yellow, while negative genetic interactions will be colored turquoise (Figure 2.4E).

40. The network displayed in the detailed view can be laid out and manipulated similar to the module map as described in Steps 33 – 35. Individual genes and interactions between genes can be selected similar to the way in which modules are selected in the module map as described in Step 36.
41. *Optional Step:* Steps 41 – 44 should be followed if quantitative interaction data is present. An alternate means of visualizing a single module or a set of connected modules is via a hierarchically clustered heat map (Figure 2.4F). In this view, each row or column represents a single gene. Each cell in the matrix is colored to represent the quantitative value attached to the interaction between those two genes. For example, Figure 2.4F is a hierarchically clustered representation of the 'between-cluster' model shown in Figure 2.4E**.** The colors in the heat map represent the genetic interaction confidence scores between the genes. PanGIA can output a matrix containing either the genetic interaction confidence scores or physical interaction confidence scores between individual genes (option A), between all genes in a module or set of modules (option B):

    (A) Output interaction matrix for a select number of genes. Select the genes of interest from a detailed view as described in Step 36. Right click on any of the selected genes and select PanGIA → Save Selected Nodes to Matrix File. Next, choose the desired quantitative attribute to be outputted (i.e., physical

interaction confidence or genetic interaction confidence). The names of these quantitative attributes will be the ones assigned by the user in Step 11. A dialog box will appear prompting to you enter the output file name. Enter the file name and click ‘Save’.

(B) Output interaction matrix for all genes in a module or set of modules. Select a module(s) of interest as outlined in Step 36. Right click on any of the selected modules and select PanGIA → Save Selected Nodes to Matrix File. Choose the desired attribute to be outputted. Enter the output filename and click ‘Save’. If using the Sample data, select the modules labeled ‘Swr1p complex’ and ‘Set3p complex’ Right click on one of these two modules and select PanGIA → Save Selected Nodes to Matrix File → GScore. Provide an appropriate file name and click Save.

42. *Optional Step:* Start the MeV program. A window titled ‘Multiple Array Viewer’ should pop up. Load the interaction matrix generated in the previous by navigating to File → Load Data. The ‘Expression File Loader’ dialog window will appear. Click the ‘Browse’ button and specify the file containing the interaction matrix. A preview of the interaction matrix should appear in the ‘Expression Table’ panel. Click the upper-leftmost interaction confidence score and then click ‘Load’. A heat map of the interaction matrix will appear in the ‘Multiple Array Viewer’ window.
43. *Optional Step:* To hierarchically cluster the heat map, click on the ‘Clustering’ tab located near the top of the window and then select ‘Hierarchical Clustering’. In the ‘HCL: Hierarchical Clustering’ window which will open, check the boxes

‘Optimize Gene Leaf Order’ and ‘Optimize Sample Leaf Order’. This will ensure that genes with similar interaction profiles will be placed close to one another. Finally, click ‘OK’.

44. *Optional Step:* In the right-most panel of the ‘Multiple Array Viewer’ window navigate to ‘Analysis Results’ → ‘HCL (1)’ → ‘HCL Tree’. A hierarchically clustered version of the heat map will appear. This image can be saved by clicking on File → Save Image. Multiple output formats are available. If using the data, the heat map should look similar to Figure 2.4F.
45. In cases where a module may contain one or more genes with an unknown function, it is useful to be able to query an external web-database like Ensemble or Entrez. Cytoscape features the ability to automatically connect to and query external web-databases. Right-click on a gene of interest within the Detailed View and navigate to the ‘LinkOut’ menu. Numerous databases will be listed including Ensembl, KEGG, Uniprot, and Entrez. Select one of these databases. An internet browser window will open automatically displaying any information the selected database has on the gene of interest. This feature provides an effective way to interrogate the function of unannotated genes.

Chapter 2.4.8: Functional Enrichment of the Modules

46. Start the BiNGO plugin by selecting Plugins → Start BiNGO. The BiNGO Settings window will appear.

47. Select the module or modules of interest which will be examined for an enriched function. Create a Detailed View as outlined in Step 39. Select the genes contained in the module(s) which will be screened for an enriched GO function. To select all genes, simply press ‘Ctrl’ (or ‘Cmd’ if using Mac OS X) and ‘A’ simultaneously.
48. Type in a meaningful name for the set of genes being examined in the box titled ‘Cluster name’. Under the menu titled ‘Select Organism/Annotation’, choose the appropriate organism (for the sample data choose *Saccharomyces cerevisiae*). For the remaining options, the default values will typically suffice. Click ‘Start BiNGO’. Depending on the number of genes selected and the computer hardware, this process will take roughly 5-10 minutes. ?Troubleshooting
49. BiNGO will return an output window containing a list of GO terms that were found to be enriched along with their respective p-values. BiNGO will also return a network of GO terms showing the inter-relationships between the various GO terms that were found to be enriched. The color of each term represents its significance of enrichment.

Chapter 2.4.9: Exporting your results

50. Cytoscape enables multiple ways to export individual modules as well as the global module map. For a thorough explanation of each of these export methods, please refer to the online tutorial (http://www.cytoscape.org/documentation_users.html).

(A) Export Network as a Graphics Object

(i) The module map, as well as individual modules can be exported as a graphics file. Numerous output formats are supported including PDF, JPEG, SVG, PNG, and BMP.

(ii) To export a network as a graphics object, make sure it is the active window and then select File → Export → Network View as Graphics….

(iii) In the 'Export Network View as Graphics' dialog box, select the output file name and choose the desired output format. Click 'OK'.

(iv) If the graphics object will be further manipulated in a graphics software package such as Adobe Illustrator, we recommend exporting the network as a PDF file. Make sure to also check the box titled 'Export text as font', which will enable the manipulation of the text labels in the network image.

(B) Export Modules as a Tab-Delimited File

(i) Each of the individual modules can be exported in a tab-delimited file, where each line consists of two parts separated by a tab character: the name of the module and the genes comprising the module. If multiple genes have been assigned to a module, each gene will be separated by the '|' character.

(ii) To export the modules as a tab-delimited file, right-click on any module in the module map (i.e., the network in the Cytoscape Canvas

titled ‘Module Overview Network’) and select PanGIA → Export → Export Modules to Tab-Delimited file.

(iii) Specify the output file in the dialog box which pops up and click ‘Save’.

(C) Export Module Map as a Tab-Delimited File

(i) The entire module-map can be exported as tab-delimited file, where each single line represents a single interaction between two modules. A single line is split into nine different parts separated by a tab character. The first two parts represent the source and target module. The remaining seven parts represent various attributes describing each interaction as outlined in Table 2.3.

(ii) To export the module-map as a tab-delimited file, right-click on any module in the module map (i.e., the network in the Cytoscape Canvas titled ‘Module Overview Network’) and select PanGIA → Export → Export Module Map to Tab-Delimited file.

(iii) Specify the output file and click ‘Save’.

(D) Export the Entire PanGIA session as a Cytoscape Session File

(i) The entire session PanGIA session can be saved to file. A session file contains all of the results of this entire workflow. This includes all networks which were loaded or generated (physical, genetic, module map, individual modules), any custom visualization styles which were employed, and any enrichment results obtained from BiNGO. Saving to

a session file will enable the user to continue the analysis at a later point.

To save the entire PanGIA session to file, select File → Save As. Type in the name of the output file and click ‘Save’.

Chapter 2.5: Timing

The time required to complete this protocol is almost entirely dependent on the size of the genetic and physical networks being analyzed. Table 2.4 charts the amount of time required for the module search process (under default options) using various sized networks as input. For a physical and genetic network containing less than 100,000 interactions each (~200,000 interactions total) PanGIA takes on average 10 minutes.

Chapter 2.6: Troubleshooting

Troubleshooting advice for specific steps in the protocol can be found in Table 2.5. In addition, we outline two of the biggest problems a user may face and potential solutions to these problems below:

Module Size Issues. In some cases PanGIA may fail to return any modules or it may return modules that are either very large or very small (i.e. that consist of a single gene). The problem may be addressed by moving the ‘Module Size’ slider bar in the ‘Advanced Panel’ (see Step 24). Dragging the slider to the right will generally result in fewer, but larger modules. Dragging it to the left will have the opposite effect. Once

the slider has been set to a new position, make sure the rest of PanGIA is properly configured (Steps 19-31) and hit the ‘Search’ button located at the bottom of the PanGIA console.

Edge Reporting Issues. Another common issue is that the module map may contain either too few or too many inter-module links. PanGIA utilizes a sampling based procedure to assign p-values to every inter-module link and subsequently filters out links which fall below a specified threshold. If the threshold is set too low, this may cause a number of spurious interactions to appear in the module. On the other hand, if the threshold is set too high, this may cause PanGIA to filter out inter-module links of biological interest. This problem may be addressed by adjusting the threshold by moving the ‘Edge Reporting’ slider bar in the ‘Advanced Panel’ (as described in Step 26). Moving the slider to the left will result in a lower threshold and subsequently a larger number of inter-module links in the final map. Moving it to the right will have the opposite effect.

Chapter 2.7: Anticipated Results

Using the sample physical (Supplementary Data 1) and genetic (Supplementary Data 2) interaction networks with PanGIA configured as suggested in this protocol (module size parameter = -1.6, edge filtering parameter = 0.05, network filter = 2, training enabled, labeling threshold = 0.2), will produce a module map containing 82 modules and 164 inter-module links (Figure 2.4A). 34 of these modules overlap with known complexes provided in the file CYC*2008_Complexes.txt*

(Supplementary Data 3) and will be labeled accordingly.

The resulting module map provides a wealth of hypotheses that can form the basis for follow-up experiments. Because PanGIA has been trained on databases of known complexes and pathways, it is likely that many modules will correspond to known protein complexes in the PanGIA results[18-20]. Other modules which do not correspond to prior knowledge are prime candidates for novel complexes or pathways. The module map produced using the sample data contains 21 modules (out of 82) with two or more genes that do not overlap with any known *S. cerevisiae* physical complexes. One could test the members of these 21 modules for co-complex membership. An alternate strategy for revealing novel biological functions is to identify modules that are enriched for a common biological function, yet contain some genes that are not yet annotated to that particular function. For example, Module 24 (Figure 2.4B) is enriched for genes involved in nuclear pore organization ($P<7.05x10^{-11}$). However, two of the genes in Module 24, SEC31 and SEC16, are not annotated to this function. The logical hypothesis in this case would be that these two genes are involved in nuclear pore organization and that a deletion or knockdown of these genes should have an impact on this structure.

Inter-module links, on the other hand, predict functional overlap or synergy between the two connected modules[18,19]. For example, a large number of genetic interactions span the two modules corresponding to the Rpd3S complex and Swr1p complex (Figures 2.4D–E). The Swr1p complex has been well established as a chromatin remodeler which deposits H2A.Z, a histone variant, onto chromatin. The

function of the Set3p complex is much less well understood. The inter-module link between the two complexes suggests that Set3p may play a role in the deposition and remodeling of Htz1-containing nucleosomes. Indeed, a recent publication has provided evidence suggesting that this may be the case[60].

Chapter 2.8: Author Contributions

G.H., R.S., and T.I. conceived and led the project. G.H. coded PanGIA with supporting code from R.S., J.R., K.O., P.W., and M.S. R.S., G.H., and T.I. wrote the paper. All authors have contributed to the design of PanGIA and all have read and approved the paper.

Chapter 2.9: Acknowledgements

The authors gratefully acknowledge Sourav Bandyophadyay and Ryan Kelley for their role in the development of the framework used in PanGIA. Magall Michaut provided useful feedback on the manuscript. Colleen Doherty and Maital Ashkenazi provided helpful beta testing of the PanGIA plugin. This study was supported by grants from the National Institute of General Medical Sciences (GM070743), the National Science Foundation (NSF425926), and Microsoft (Computational Challenges in Genomewide Association Studies).

Chapter 2, in full, is a reprint of the material as it appears in the following publication: “Srivas R*, Hannum G*, Ruscheinski J, Ono K, Wang PL, Smoot M, Ideker T. *Assembling global maps of cellular function through integrative analysis of*

physical and genetic networks. Nat. Protoc. 6(9) (2011)". The dissertation author was the primary investigator and author of this paper. For the sake of brevity, all Supplementary Datasets have not bee included here. These items can be accessed at http://prosecco.ucsd.edu/PanGIA/

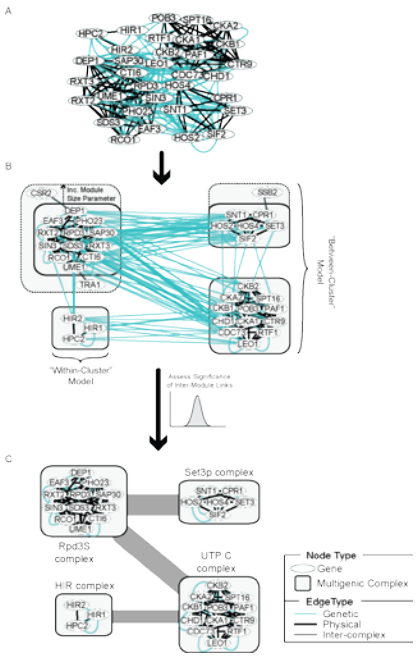


Figure 2.1: Overview of PanGIA's method for identifying a module map of cellular function from physical and genetic networks

(A) PanGIA takes as input a physical and genetic network. Black edges refer to physical interactions, while turquoise edges refer to genetic interactions. (B) Both 'within-cluster' and 'between-cluster' models are identified using the physical and genetic network. A 'within-cluster' model or module consists of a set of genes connected by a large number of physical and genetic interactions. In this example four 'within-cluster' models are identified. A 'between-cluster' model or inter-module link consists of two 'within-cluster' models spanned by a bundle of genetic interactions. Here, five putative 'between-cluster' models have been identified. The size of 'within-cluster' models can be controlled via the 'Module Size' parameter. Higher values of the 'Module Size' parameter lead to larger complexes (denoted by the dashed line). (C) If quantitative interaction data have been made available, the significance of each 'between-cluster' model can be assessed. Only significant inter-module links are displayed in the final module-map (three of the five putative inter-module links are significant in this example). The thickness of the line reflects the score of the inter-module link, which is based on the number of physical and genetic edges spanning the two modules. If a biological annotation set is provided, PanGIA will check the overlap between the set of genes comprising the annotation and the set of genes comprising each module. If the overlap exceeds a user-specified threshold, the module will be labeled with the name of the annotations. Here, all four modules overlap with known complexes and are labeled accordingly.

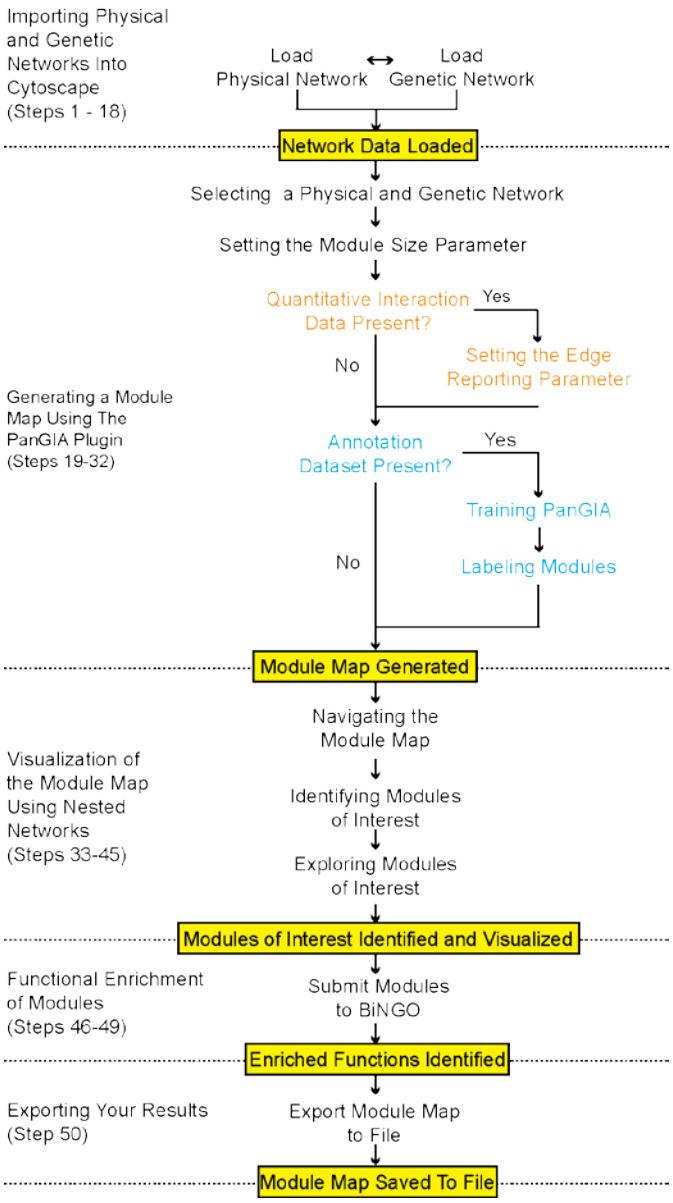


Figure 2.2: Outline of the protocol

Analyses listed in black indicate required steps in the protocol. Analysis listed in orange represent optional steps which may be performed if quantitative interaction data is present; those listed in light blue are optional steps which may be performed if a biological annotation data-set is present. The yellow boxes indicate the desired outcome at the end of each major section in the protocol.

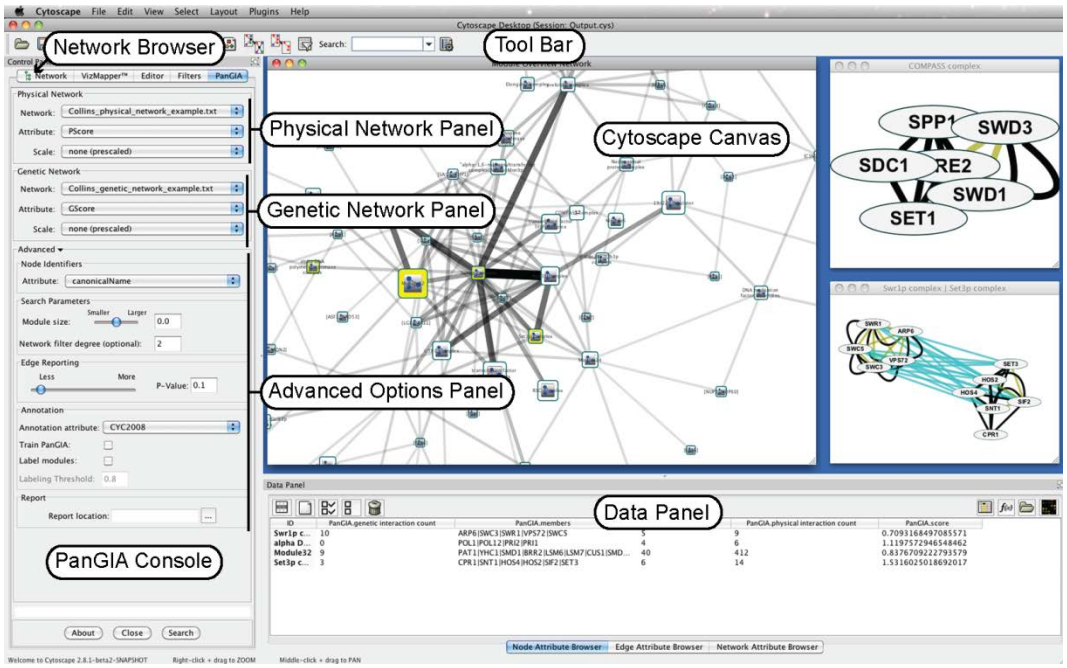


Figure 2.3: The PanGIA console

The Cytoscape canvas displayed the network data and, may initially be blank. The Data Panel (bottom) is used to display node, edge, and network attribute data. The Tool Bar (top) contains numerous command buttons used for navigating the network. The PanGIA console (left) is divided into three main panels including the Physical Network Panel, the Genetic Network Panel, and the Advanced Options Panel. The Network Browser may be accessed by clicking on the Network tab located to the left of the PanGIA console tab.

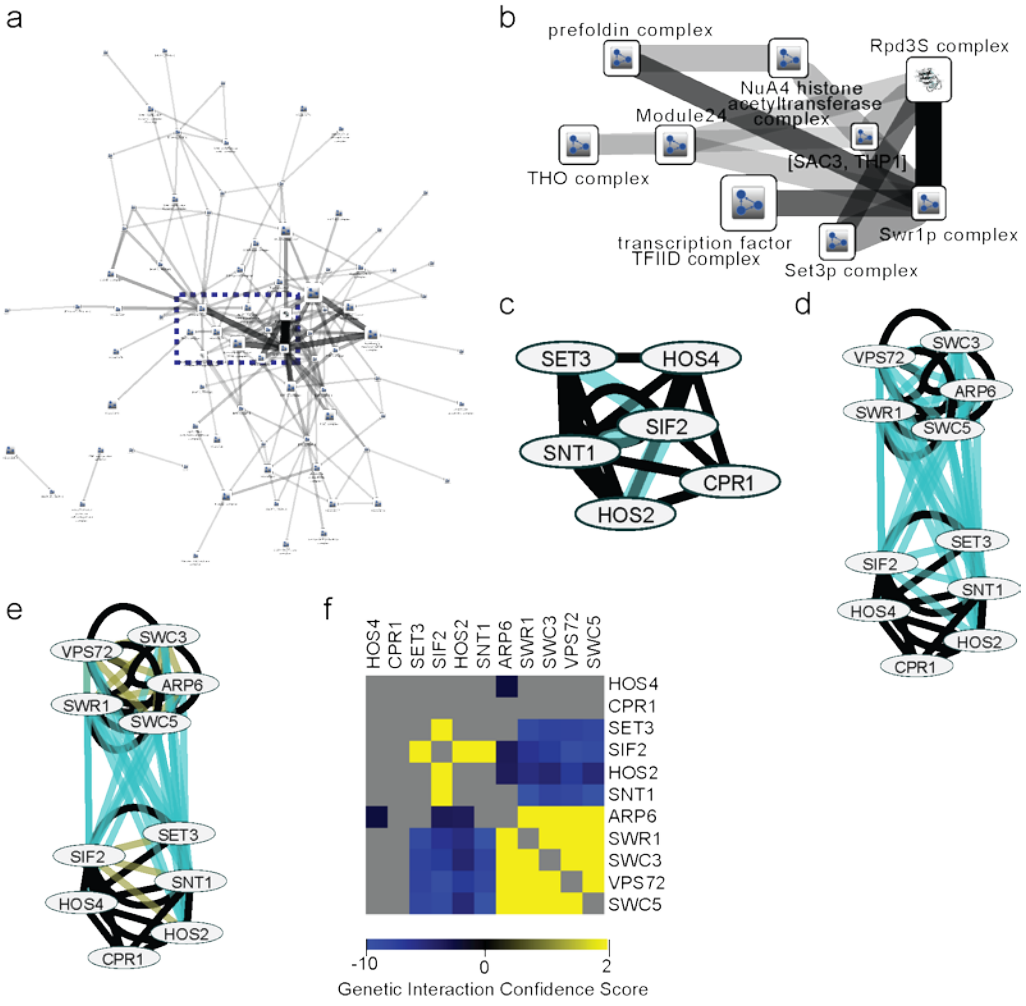


Figure 2.4: PanGIA Output

(A) The module map returned by PanGIA. Each node is a separate module or complex and the area of the node reflects the number of genes contained within the module. (B) A zoomed in portion (blue box) of the module map seen in (A). If an annotation set was provided and the labeling option was chosen, modules which overlap substantially with an annotation are labeled as such (e.g., Rpd3S complex). Modules not overlapping with any of the provided annotations are either given a generic name (e.g. Module 24) or labeled with a gene name (e.g. [SAC3,THP1]) if the module contains only one or two genes. (C) A detailed view for a single module. Each node represents a single gene which was assigned to this module. Physical interactions are colored black, while genetic interactions are colored turquoise. (D) A detailed view for two modules. Edges are colored similar to (C). The layout algorithm seeks to physically separate each module. (E) The same network as shown in (D) but visualized as a hierarchically clustered heat map using MeV[56].

Table 2.1: List of databases of physical and genetic interaction data

Database Name	**URL**	**# of Organisms Covered**	**Physical Interaction Data Available?**	**Genetic Interaction Data Available?**	**Quantitative Interaction Data Available?**
STRING	string-db.org	630	yes	no	yes
DIP	dip.doe-mbi.ucla.edu/dip/Main.cgi	372	yes	no	yes
IntAct	www.ebi.ac.uk/intact/main.xhtml	305	yes	no	yes
ConsensusPathDB	cpdb.molgen.mpg.de	3	yes	no	no
BioGRID	thebiogrid.org	18	yes	yes	no
MINT	mint.bio.uniroma2.it/mint/Welcome.do	30	yes	no	yes

Table 2.2: Examples of databases from which to obtain annotation data

Database Name	URL	# of Organisms Covered	Annotation Type
Gene Ontology (GO)	www.geneontology.org/GO.downloads.annotations.shtml	48	Physical complexes, biological processes, signaling pathways, metabolic pathways
MIPS CORUM	mips.helmholtz-muenchen.de/genre/proj/corum	3	Physical complexes
KEGG	www.genome.jp/kegg/pathway.html	833	Metabolic pathways, signaling pathways
CYC2008	wodaklab.org/cyc2008/	1 (*S. cerevisiae*)	Physical complexes
SGD Pathways	pathway.yeastgenome.org	1 (*S. cerevisiae*)	Metabolic pathways
MetaCyc	metacyc.org	2000	Metabolic pathways
Reactome	www.reactome.org	20	Metabolic pathways

Table 2.3: Description of Module-Level Attributes Returned by PanGIA

Attribute Name	Attribute Type (Node or Edge)	Description
PanGIA Member Count	Node	Number of genes present in module
PanGIA Module Physical Interaction Count	Node	Number of physical interactions present in this module.
PanGIA Module Genetic Interaction Count	Node	Number of genetic interactions present in this module.
PanGIA Source Size	Edge	Member count of the source module
PanGIA Target Size	Edge	Member count of the target module
PanGIA Genetic Interaction Count	Edge	Number of genetic interactions spanning the two modules connected by this edge
PanGIA Physical Interaction Count	Edge	Number of physical interactions spanning the two modules connected by this edge
PanGIA P-value	Edge	Significance of the inter-module link
PanGIA Edge Score	Edge	The total score of genetic interactions spanning two modules minus the score of the physical interactions
PanGIA Genetic Interaction Density	Edge	Represents the Edge Score divided by the Genetic Interaction Count.

Table 2.4: Time Required to Run PanGIA on Networks of Various Sizes

Number of interactions (Genetic + Physical)	Run Time	
	Processor: Dual Core, 32-bit (3.2 GHz) Memory: 2 GB Graphics Card Memory: 256 MB	Processor: 8-core, 64-bit (2.8 GHz) Memory: 8 GB Graphics Card Memory: 256 MB
10,000	<1 minute	<30 seconds
50,000	1 minute	<1 minute
100,000	2 minutes	1.5 minutes
500,000	15 minutes	10 minutes
1,000,000	Insufficient Memory	30 minutes

Table 2.5: Troubleshooting Table

Step	Problem	Possible Reason	Solution
1	Executing cytoscape.bat (Windows) or Cytoscape.sh (Mac OSX, Linux) does not open Cytoscape.	Java is not installed properly.	Make sure Java Version 1.6.014 or higher is installed. Java can be downloaded at http://www.java.com
30	PanGIA fails to label any of the modules in the final module map.	Threshold for labeling may be set too high.	Set the labeling threshold slightly lower to allow more modules to be labeled.
32	The module search process is taking a very long time.	Insufficient memory and/or processing power.	Very large physical or genetic networks (>500,000 interactions) require a larger amount of memory than specified in the Equipment Setup section. See the Timing section for recommendations on the amount of memory and processing power required for larger networks.
45	The queried database fails to return any information on the selected gene(s) of interest.	Mismatched gene identifiers.	When querying an external database, the identifier of the selected gene(s) must be identical to the identifier used by the external database. For example, if querying the Ensemble database, selected genes need to use Ensembl identifiers in order to have any information returned. Use one of the recommended website to map gene identifiers if there is any discrepancy[61,62].
48	BiNGO supplies an error message asking to 'Please select one or more nodes.'	No genes were selected for examining functional enrichment.	Visualize the module(s) of interest as outlined in Step 39. In the detailed view, select one or more genes of interest. All nodes (genes) can be selected in a detailed view by pressing 'Ctrl' (or 'Cmd' if using Mac OS X) + 'A'

Chapter 3. Dissection of DNA Damage Response Pathways using a Multi-Conditional Genetic Interaction Map

Chapter 3.1: Abstract

To protect the genome, cells have evolved a diverse set of pathways designed to sense, signal and repair multiple types of DNA damage. To assess the degree of coordination and crosstalk among these pathways, we systematically mapped changes in the cell's genetic network across a panel of mechanistically distinct DNA-damaging agents, resulting in ~1,800,000 differential measurements. Each agent was associated with a distinct interaction pattern, which, unlike single mutant phenotypes or gene expression data, has high statistical power to pinpoint the specific repair mechanisms at work. The agent-specific networks revealed novel roles for the histone acetyltranferase Rtt109 in the mutagenic bypass of DNA lesions and the neddylation machinery in checkpoint control and genome stability, while the network induced by multiple agents implicates Irc21, a previously uncharacterized protein, in cell cycle regulation and DNA repair. Our multi-conditional genetic interaction map provides a unique resource that identifies both agent-specific and general DNA damage response pathways.

Chapter 3.2: Introduction

Failure of cells to respond to DNA damage is associated with genome instability and the onset of diseases such as premature aging and cancer[63]. To combat

DNA damage, cells have evolved an intricate system, known as the DNA damage response (DDR), which senses DNA lesions and activates downstream pathways such as chromatin remodeling, cell cycle checkpoints and DNA repair[64]. Many studies have sought to use genome-scale technologies to better define and map the DDR, including systematic phenotyping of single mutants[65], RNAi screening[66] and gene expression profiling[67].

While these strategies have met with success in identifying new DDR genes, they have raised a number of questions with regard to how DDR pathways coordinate with one another. For instance, the initial view of DNA damage checkpoints was as a collection of pathways with the sole task of coordinating cell cycle progression with DNA repair[68]. However, recent studies have implicated checkpoints in other processes, including transcription regulation, telomere length maintenance, and apoptosis, suggesting that there is extensive crosstalk between such processes during the DDR[69,70]. Increasing evidence suggests that much of this crosstalk is likely to be dependent on the nature of the DNA lesion. For example, the Bloom syndrome helicase (BLM; Sgs1 in budding yeast) functionally interacts with components of the S-phase replication checkpoint (e.g. Mrc1/Claspin) when replication forks stall, whereas it cooperates with factors of the DNA damage checkpoint (e.g. Rad17/hRAD9 of the 9-1-1 complex) after DNA double-stranded break (DSBs) formation[71]. An important next step is therefore to understand how functional inter-connections between the various components of DDR pathways are formed and altered in response to various genotoxic insults.

To address this issue, we turned to a recently-developed interaction mapping methodology called differential epistasis mapping or dE-MAP[16]. This approach is based on Synthetic Genetic Array (SGA) technology[72] which enables rapid measurement of genetic interactions, i.e., combinations of mutations to two or more genes that produce an unexpected effect on growth. Genetic interactions fall into one of two categories[10]: positive interactions (i.e., epistasis) which typically occur among genes involved in the same complex or pathway, and negative interactions (i.e., synthetic sickness or lethality) which identify genes in compensatory pathways. In the dE-MAP approach, SGA is used to measure genetic interactions under standard conditions as well as under perturbations of interest and, by comparing the resulting networks, interactions that are altered in response to perturbation can be quantitatively assessed. These 'differential' genetic interactions reveal a unique view of cellular processes and their inter-connections under specific stress conditions[17].

Here, we apply our dE-MAP technique to systematically map the genetic modules and networks induced by distinct types of DNA damage, which we anticipate will be an important resource for the study of the DDR and its associated diseases. Based on this map of both agent-specific and general differential interactions, we investigate and validate a number of new pathways and factors involved in DDR. In addition, our work demonstrates that differential interaction mapping across a panel of treatments is a powerful and general approach for disentangling a web of distinct but interrelated signaling processes.

Chapter 3.3: Results

Chapter 3.3.1: Mapping differential genetic networks across distinct types of DNA damage

We constructed a dE-MAP in the budding yeast *S. cerevisiae* centered on the measurement of all possible interactions between a set of 55 query genes and a set of 2,022 array genes (Figure 3.1A). The 55 query genes were chosen to provide coverage of the pathways that define the DDR including representatives of the distinct DNA repair processes (Table S1). The array genes included all of the queries and, to explore crosstalk between DNA repair and other cellular functions, genes involved in cell cycle regulation, chromatin organization, replication, transcription, and protein transport (Table S2). Double mutant strains were constructed six different times for each query–array gene pairing (Methods). Growth rates were measured in standard conditions (Untreated) and after exposure to three chemical agents that induce distinct types of DNA damage: the DNA alkylating agent methylmethane sulfonate (MMS), the topoisomerase I inhibitor camptothecin (CPT), and the DNA intercalating agent zeocin (ZEO). In each condition, double mutants were assigned quantitative S scores[10], which quantify the extent to which the double mutant grew either better (positive S Score) or worse (negative S Score) than expected. In total, the genetic interaction map contained quantitative scores for 97,578 pairs of genes (Table S3). Several routine quality control measures were employed to ensure a high-quality dataset (Supplemental Figure 3.1 and Supplementary Methods).

Using established scoring thresholds to highlight significant positive and negative interactions ($S \geq 2.0$ or $S \leq -2.5$)[10], we uncovered 8222 significant interactions in untreated conditions versus 10584, 9418, and 9969 significant interactions in MMS, CPT, and ZEO, respectively. A first comparison of these sets of interactions reveals numerous differences in genetic interactions between the treated and untreated conditions (Figure 3.1B). On average, 48% of positive interactions and 33% of negative interactions were unique to the treated networks, indicating the presence of DNA damage-induced epistasis and synthetic lethality (Figure 3.1C). To identify which of these differences were statistically significant, we used a previously published scoring methodology[16] to assess the difference in S score for each gene pair before versus after treatment. A p-value of significance was assigned by comparing this quantitative difference to a null distribution of differences derived from replicate genetic interaction screens from the same condition. We refer to this network as the 'differential' genetic network since it is derived by examining the difference between two static networks (Figure 3.1D). At a p-value threshold of 0.002 (FDR = 8.7%; Supplementary Methods), we identified 3150 significant differential interactions when comparing MMS to untreated conditions, versus 1120 and 1474 differential interactions when comparing CPT and ZEO to untreated conditions, respectively (Figure 3.1E).

Across all three differential networks, the number of differential positive interactions (interaction becomes more positive under DNA damage) was roughly equal to the number of differential negative interactions (interaction becomes more

negative under DNA damage, Figure 3.1E). We did note however that many differential interactions become more significant in response to treatment ('Gain of Interaction'), whereas only a small fraction of interactions are lost or reduced in significance ('Loss of Interaction'; Figure 3.1F). This finding is consistent with the notion that the cell activates new pathways in response to DNA damage. Intriguingly, we also observed a third class of interactions which were insignificant in either condition yet had a significant change across conditions, from weakly positive to weakly negative or vice versa ('Sign Switching'; Figure 3.1F). These interactions go unnoticed in either condition yet show changes strong enough to be detected in the differential analysis.

Chapter 3.3.2: Differential interactions effectively discriminate among different DNA damage responses

We next examined all networks, differential and static, for their ability to highlight genes that function in the DDR (Methods). All three differential networks had high enrichment for interactions with known DNA repair genes, while static networks had much less enrichment (MMS) or no enrichment (CPT, ZEO, Untreated) in this regard (Figure 3.2A). Instead, all four static networks showed the strongest enrichment for genes involved in chromatin organization, as had been noted in the original report of the dE-MAP method[16]. Moreover, 15 of the top 20 differential interaction 'hubs' (those with the greatest number of interactions) were annotated as a DNA repair gene, whereas the top interaction hubs in static networks were largely

associated with chromatin organization (Figure 3.2B). Thus, in contrast to static interactions, differential interactions measured across a shift in conditions tend to highlight gene functions related to that condition.

Despite the strong enrichment for DNA repair genes across all differential networks, we found that these networks were strikingly different from one another. Few differential interactions (584 interactions; 11%) were induced by more than one agent and only 45 interactions were induced by all agents (Figure 3.2C). In contrast, a control experiment indicated much better agreement between replicate differential networks generated in response to the same agent (Supplemental Figure 3.2A). To determine whether the distinct interaction patterns induced by each agent were indicative of distinct DNA repair mechanisms, we examined the differential networks for enrichment of interactions with genes involved in six major DDR pathways (Figure 3.2D, Table S4, Methods). The CPT network was highly enriched for DSB repair ($P = 10^{-143}$) and DNA damage checkpoint functions ($P = 10^{-74}$), consistent with the known mechanism of action of CPT which stabilizes DNA topoisomerase 1–DNA complexes. During S-phase the replication machinery collides with these structures resulting in the production of DSB specifically during this phase of the cell cycle[73]. The MMS network displayed only a mild enrichment ($P = 0.009$) for interactions with components of base excision repair (BER), an unexpected result given the mechanism of action of MMS which modifies guanine and adenine bases leading to base mispairing and replication fork blocks[74]. However, replication-blocking lesions can be bypassed by post-replication repair pathways (PRR) such as translesion synthesis

(TLS) and DNA damage avoidance or, in case of fork collapse and subsequent chromosome breakage, are counteracted by DSB repair pathways[75]. All these pathways showed strong enrichment in the MMS network (Figure 3.2D). Finally, the ZEO network was found to enrich for interactions with genes involved in BER and PRR rather than for genes involved in DSB repair (P = 0.002), suggesting that our ZEO treatment leads to the formation of abasic sites rather than DNA strand breakage, consistent with the mode of action of this intercalating agent at lower concentrations[76].

These functional enrichments suggest that the differential networks help decode the particular combination of DDR pathways underlying the response to each agent. To test this hypothesis explicitly, we measured the statistical association between the three agents and the six major DDR pathways as revealed by differential interactions (modified Pearson's Chi-Square Test, see Supplementary Methods). In contrast to functional enrichment, statistical association measures the extent to which interactions induced by each agent implicate a set of genes that discriminates among the six pathways (i.e., genes which associate with some DDR functions but not others). We found that differential interactions were indeed able to elicit a significant association between agents and pathways, especially for the top 5% of interactions (Figure 3.2E). Moreover, differential interactions performed very favorably at this task in comparison to single-mutant fitness[65] or differential mRNA-expression profiles [77,78] gathered previously for the same three agents (Supplementary Methods). Neither of these data types was able to significantly link DNA damaging agents to particular responses (Figure 3.2E). A likely explanation for the better performance of differential

networks lies in the greater sample size afforded by this technology. Whereas single-mutant fitness and gene expression profiling are limited to measurements of individual genes (181 genes across the six DDR pathways), the differential networks cover interactions between DDR genes and over 30% of the yeast genome (39,973 interactions in total). Thus, while single mutant and gene expression profiling are adept at defining high-level biological functions (e.g., DNA repair), differential genetic interactions can begin to tease apart a very specific set of (partially overlapping) mechanisms.

Chapter 3.3.3: Neddylation affects genome integrity and checkpoint control after CPT

Next, we turned to analysis of novel gene functions and pathways implicated by the differential networks. As a starting point for this analysis we examined the genes which were highly connected within the differential network (i.e., hubs). The gene with the greatest overall number of differential interactions was *RAD17* (Figure 3.22B), a component of the 9-1-1 checkpoint complex which is recruited to DSB sites to activate the Mec1-kinase signaling cascade, resulting in cell cycle arrest and repair[68]. Consistent with the role of Rad17 in the DSB response[79], we found that the majority of its interactions were induced specifically in response to CPT (73%, Figure 3.3A). To gain further insight into potential CPT-induced pathways involving the checkpoint, we examined the entire CPT-induced genetic interaction profile of *RAD17* (Figure 3B), which revealed strong differential negative interactions with prominent DSB repair genes (*RAD59*) and checkpoint regulators, such as *TEL1*. This is

consistent with reports showing that Tel1 functions parallel to Rad17 to regulate checkpoint activation following DSBs[80].

Two additional genes, *RUB1* and *UBC12*, which encode key components of the yeast neddylation machinery, displayed strong differential negative interactions with *RAD17* (Figure 3.3B). Neddylation is a process by which the Rub1 protein (NEDD8 in humans) is conjugated to target proteins in a cascade of reactions that involves E1 activating, E2 conjugating (in *S. cerevisiae* only Ubc12) and E3 ligating enzymes in a manner analogous to ubiquitylation and SUMOylation[81]. Whereas ubiquitylation and SUMOylation have been shown to regulate a myriad of cellular processes, including DDR[82], those that involve neddylation remain largely unknown due to the limited number of neddylation substrates that have been identified[83]. In further support of a potential link between neddylation and checkpoint pathways, the CPT network revealed a number of additional differential negative interactions between *RUB1*/*UBC12* and other checkpoint genes, including *DDC1, RAD9* and *RAD24* (Figure 3.3C). These interactions were also observed via spot dilution assays, confirming that cells defective for neddylation and DNA damage checkpoints are hypersensitive to CPT (Figure 3.3D).

To investigate a role for the neddylation machinery in DNA damage checkpoint control, we assessed *rub1*Δ and *ubc12*Δ mutants for their progression through the cell cycle in the presence of CPT. After arrest in G1 and release into medium containing CPT, *rub1*Δ and *ubc12*Δ mutants had significant accumulation of cells in G2 at 90 and 105 minutes whereas wild-type cells efficiently progressed

through G2 and M-phase into the next cell cycle (Figure 3.3E and Supplemental Figure 3.3A). As this delay was not observed in the absence of CPT (Supplemental Figure 3.3B), we demonstrate for the first time that neddylation mutants display perturbations in cell cycle progression. Since defects in cell cycle checkpoints have been shown to contribute to genome instability[63], we decided to measure the rate of gross chromosomal rearrangements (GCR) in the neddylation mutants (Supplementary Methods). The rate of GCR events in the *ubc12*Δ mutant was nearly 2.7-fold greater than in wildtype, whereas the *rad17*Δ*ubc12*Δ double mutant showed, respectively, a 7- and 2-fold increase in GCR rates when compared to the *ubc12*Δ and *rad17*Δ mutants (Figure 3.3F), suggesting that neddylation and checkpoint pathways are likely to cooperate in promoting genome stability.

The best-studied NEDD8/Rub1 targets are cullin proteins, which are scaffolds for the assembly of multi-subunit cullin-RING ubiquitin ligases (CRLs)[81,84]. CRLs are responsible for the turnover of a vast majority of proteins and consequently play a major role in maintaining cellular homeostasis[85]. Strikingly, another interaction hub in the differential network was the cullin Rtt101 (Figure 3.22B), which has been shown to play a critical role in regulating the G2/M checkpoint by promoting proteasomal degradation of Mms22[86]. Given the role of neddylation in CRL modification, we examined whether this process would affect the steady state levels of Mms22. We observed a faster degradation of Mms22 in a *rub1*Δ strain when compared to wildtype, suggesting that neddylation, in contrast to Rtt101-dependent ubiquitylation[86], promotes Mms22 stability (Figures 3.3G–H). Taken together, these data implicate the

neddylation machinery as a novel factor that regulates cell cycle progression in response to DNA damage and contributes to genome stability, most likely by regulating the steady state levels of DDR factors such as Mms22.

While CRLs are the most well-studied Rub1 substrates to-date, emerging evidence suggest that many other proteins may be modified by neddylation[83]. For example, ribosomal proteins and E3 ubiquitin ligases, such as the p53 regulator Mdm2, have been shown to be substrates for neddylation[83,87]. We infer from this that the stability of DDR factors such as Mms22 may be regulated by direct neddylation, or indirectly by the neddylation of E3 ubiquitin ligases or CRLs (Figure 3.3I). Although further work will be required to resolve the precise mechanisms, these data confirm the power of differential genetic data to identify novel agent-induced functional connections.

Chapter 3.3.4: Irc21 is a general response factor in checkpoint control, repair and genome stability

While the interactions induced by the three agents were largely divergent, the differential analysis did implicate a 'conserved' network of 584 interactions that were altered in response to at least two agents (Figure 3.4A). Many known DNA repair factors were highly connected within this network including DSB repair factors (Rad52, Sae2, Mre11, Rad59), PRR genes (Rad18), and chromatin remodelers (Swr1) which have well-documented roles in the DDR[88,89]. In particular, our analysis once again highlighted the damage checkpoint gene *RAD17* as a hub not only of the CPT

network (see above), but also of conserved interactions across agents (Figure 3.4A, top inset). These included a differential positive interaction with *IRC21*, an as yet uncharacterized gene, in response to both CPT (differential $P = 4.7x10^{-7}$) and MMS ($P = 8.3x10^{-7}$), but not ZEO ($P = 0.53$).

We confirmed that Irc21 is expressed *in vivo* in yeast (Supplemental Figure 3.4A), and that deletion of *IRC21* in a *rad17*Δ mutant suppresses its sensitivity to CPT and MMS (Figure 3.4B). Analysis of the Irc21 protein sequence revealed the presence of a cytochrome b5-like domain (Supplemental Figure 3.4B), which is usually found in proteins that are involved in cytochrome P450-dependent metabolic processes[90]. To rule out that the suppression was due to Irc21 affecting drug metabolism via its cytochrome b5 domain, we exposed cells to ultraviolet light (UV) and ionizing radiation (IR) and were able to re-produce the suppressive phenotype in both cases (Figure 3.4B). Ectopic expression of Irc21 in the *irc21*Δ*rad17*Δ mutant restored the sensitivity to DNA damaging agents to that observed for the *rad17*Δ mutant (Figure 3.4B and Supplemental Figure 4.4E). These results suggest that Irc21 affects cell survival in response to genotoxic insult by modulating the DNA damage checkpoint rather than affecting drug metabolism.

To further explore this possibility, we profiled *rad17*Δ, *irc21*Δ and *rad17*Δ*irc21*Δ mutants for their cell cycle progression in the presence of MMS. While the wildtype and *irc21*Δ strains displayed slow S-phase progression and accumulated in G2 two hours after release from G1, the checkpoint-deficient *rad17*Δ strain rapidly progressed through S-phase and accumulated in G2 within an hour (Figure 3.4C and

Supplemental Figure 3.4C). Remarkably, deletion of *IRC21* in the *rad17*Δ strain partially suppressed the checkpoint deficiency as we noted an increased fraction of cells remaining in S-phase (20.7% versus 10.7% at two hours after release; Figure 3.4C). In support of this observation, we noted that while the *rad17*Δ mutant failed to activate the central checkpoint kinase Rad53, denoted by the absence of phosphorylated forms of Rad53 (Figure 3.4D), the *irc21*Δ*rad17*Δ double mutant displayed almost a complete restoration of this phenotype with Rad53 becoming hyperphosphorylated at near wild-type levels (Figure 3.4D).

Checkpoint proteins detect DNA lesions, arrest the cell cycle and trigger DNA repair[63,88]. Given that Irc21 modulates the DNA damage checkpoint, we examined whether it also functions in the timing of repair. The Rad52 repair protein has been shown to accumulate into subnuclear foci representing active repair center [91]. We used this phenotype to investigate the capacity of wildtype, *rad17*Δ, *irc21*Δ and *rad17*Δ*irc21*Δ strains to repair MMS-induced DNA damage (Figure 3.4E). In all strains the maximum number of Rad52 foci was reached one hour after exposure to MMS. While Rad52 foci gradually disappeared by 2–4 hours, persistent foci were observed in the *rad17* mutant, indicating abrogation of repair. However, deletion of *IRC21* alleviated the repair defect seen in the *rad17*Δ strain, as indicated by the enhanced dissolution of Rad52 foci in the *irc21*Δ*rad17*Δ strain compared to that in the *rad17*Δ strain (4 hour time point, Figure 3.4E).

Finally we found that, whereas *irc21*Δ cells showed no alterations in genomic stability, *rad17*Δ cells displayed an 8.2-fold increase in GCR events compared to

wildtype (Figure 3.4F). However, *irc21*Δ*rad17*Δ cells only showed only a 4.5-fold increase, suggesting that deletion of *IRC21* partially rescues the deleterious impact of Rad17 loss on GCR (Figure 3.4F). Together, these results suggest that Irc21 not only modulates DNA damage checkpoints, but also promotes efficient repair of DNA damage and contributes to genome stability.

In contrast to previous high throughput localization studies, which reported Irc21 localization in the cytoplasm[92], we found that Irc21-GFP localizes in both the cytoplasm and nucleus (Supplemental Figures 3.4D–E). Irc21-GFP, however, did not accumulate into MMS-induced sub-nuclear foci as observed for Rad52-YFP (Supplemental Figure 3.4D), suggesting that it may not operate directly at DNA lesions. Interestingly, we observed that *irc21*Δ strains are hypersensitive to MMS when combined with the TOR inhibitor rapamycin, a compound that leads to increased autophagy (Supplemental Figure 3.4F), suggesting that Irc21 may affect the DDR through TOR signaling and autophagy-mediated protein degradation, a process which has been recently linked to DDR[93]. Although further work will be required to work out this intriguing connection, our analysis of the set of commonly perturbed genetic interactions identified Irc21 as a novel DDR factor that regulates cell cycle progression, repair and genome stability.

Chapter 3.3.5: An integrated module map reveals a novel role for Rtt109 in translesion synthesis

An especially powerful approach for interpreting genetic interactions is in conjunction with knowledge of physical protein-protein interactions and protein complexes[45,94]. We have previously demonstrated that, while static genetic interactions are enriched among components of the same physical complex, differential genetic interactions tend to occur between distinct but functionally-related complexes[16,95,96]. Based on this idea we used a recently-described integrative clustering algorithm[97] to transform our differential genetic interaction data for all agents into a map of 179 modules and 452 module-module interactions (Figure 3.5A and Tables S5– 6). Modules group genes with similar patterns of both genetic and physical interactions, many of which were found to coincide with known DNA repair complexes. Module-module interactions represent bundles of differential genetic interactions which span across the genes in the two modules and point to DNA damage-induced cooperativity.

The low overlap we had observed among the genetic networks of the three agents (Figure 3.2C) was reproduced in the module map, as the vast majority (~90%) of module-module interactions were found to occur in response to a single agent. Indeed, each of the agents highlighted a different module as a central hub of interactions (Figure 3.5A); these were the 9-1-1 DNA damage checkpoint complex (CPT), the Mms2/Ubc13 E2 ubiquitin conjugase complex (MMS) and the MRX double strand break repair complex (ZEO). Many of the interactions involving these hub modules recapitulate known drug-specific DDR mechanisms. For example, the 9-

1-1 complex was found to genetically interact with the S-phase checkpoint complex Csm3/Tof1, which is consistent with recent work showing that both complexes are required for the response to CPT[98]. We also observed an MMS-dependent link between the INO80 chromatin remodeling complex and the Mms2/Ubc13 and Rad6/Rad18 ubiquitin E2 conjugase/E3 ligase complexes, which are involved in DNA damage tolerance. These observations are in line with recent work implicating a role for INO80 in this pathway that operates to overcome MMS-induced replication fork blocks[61].

The module map also highlighted several MMS-dependent interactions with the histone acetyltransferase Rtt109, including an interaction with the replicative polymerase Polδ, which is consistent with its known role in maintaining the integrity of replisomes (Figure 3.5B)[99]. Unexpectedly, we also observed a differential positive relationship between Rtt109 and the TLS polymerases Rev1 and Polζ, a complex composed of Rev3 and Rev7 (Figure 3.5B), which we validated using a spot dilution assay (Supplemental Figure 3.5A). Polζ-dependent TLS enables cells to replicate through DNA lesions, ensuring that such lesions do not result in the collapse of replication forks[100]. Moreover, Polζ, in conjunction with Polδ, is responsible for as much as 85% of the bypass events at abasic sites[101], with much of this occurring in an error-prone fashion[100].

To validate the link between Rtt109 and TLS, we utilized a *CAN1* forward mutation assay (Supplementary Methods) which reports any mutation that disrupts Can1 function, resulting in a canavanine-resistance (*can1*^{r}) phenotype. Cells with

proficient TLS activity will accrue mutations at this locus at a much higher rate enabling them to survive selection on media containing canvanine. As expected, deletion of *REV3,* which impairs Polζ function, produced almost no *can1*^{r} colonies mutants, indicating an almost complete loss of TLS activity (Figure 3.5C and Supplemental Figure 3.5B). Surprisingly the *rtt109*Δ strain showed a 2-fold decrease in the rate of *can*^{r} colonies when compared to wild type (Figure 3.5C and Supplemental Figure 3.5B). A similar phenotype was observed for cells expressing H3K56R, a mutant form of histone H3 that cannot be acetylated by Rtt109 (Figure 3.5D and Supplemental Figure 3.5C). This suggests that Rtt109 affects TLS through acetylation of H3K56. Finally, the *rev3*Δ*rtt109*Δ and *rev3*ΔH3K56R mutants displayed a reduction in the rate of *can*^{r} colonies which was comparable to that of the *rev3*Δ mutant (Figures 3.55C– D), suggesting that the TLS defect in the absence of Rtt109 activity may depend on the Polζ complex.

Recent work has shown that Rtt109 mediates acetylation of newly synthesized histones that are deposited onto synthesized DNA during DNA replication[102]. In support of this, *rtt109*Δ or H3K56R mutants have been found to genetically interact with several genes involved in DNA replication, including DNA polymerase α and PCNA[103]. Moreover, these mutants fail to stabilize Pol α and PCNA at stalled replication forks[104]. Since PCNA serves as a clamp for the loading of TLS polymerases[105], we suggest a model in which Rtt109-dependent H3K56 acetylation regulates PCNA-dependent loading of TLS polymerases, including Pol ζ, at sites of MMS-induced fork stalling.

Chapter 3.4: Perspective

Here, we have measured ~100,000 differential interactions of DDR genes in response to three genotoxic agents with distinct modes of action. While the networks induced by each agent were largely divergent (Figure 3.2C), each was highly effective in pinpointing a set of pathways involved in specific types of DNA repair. To further aid in the discovery and mapping of these pathways, we integrated our multi-conditional genetic network with protein interaction data to uncover a global map of gene modules and their inter-functional relationships (Figure 3.5A). Here again, the module interactions induced by each agent were largely divergent, pointing to many agent-specific repair mechanisms. For example, the module map identified a link between the histone acetyltransferase Rtt109 and the Polζ complex suggesting a role for Rtt109 in the bypass of alkylated bases (Figures 3.5B–C). Together, the multi-conditional dE-MAP and module map provide a major resource for further discovery, with hundreds of agent-specific functional links involving novel combinations of DDR pathways as well as connections to uncharacterized proteins.

Because DNA repair pathways are remarkably well conserved[64], this resource is not limited to yeast biology but also informs the response to DNA damage in humans [106] and related diseases such as cancer. For example, the dependency we uncovered between neddylation and the DNA damage checkpoint (Figure 3.3C) is echoed in a recent study in humans, in which an inhibitor targeting the NEDD8 (ortholog of Rub1) activating enzyme (NAE1) lead to extreme sensitivity to ionizing radiation in pancreatic cancer cells expressing mutated p53. Here, we have found that

cells lacking both neddylation and checkpoint machinery exhibit not only increased cell killing following exposure to CPT (Figure 3D), but also display a marked G2/M arrest (Figures 3.3E and Supplemental Figure 3.3A) and higher levels of genomic instability (Figure 3.3F). Moreover, our data suggests neddylation may regulate the turnover of DDR factors, such as Mms22 (Figure 3.3G) indicating a potential mechanism for this functional link. In addition to shedding light on human DDR pathways, differential synthetic lethal interactions may serve as a key resource in the emerging "synthetic-lethal" approach to cancer therapy[63,108]. In this respect, our multi-conditional dE-MAP provides an extensive catalog of genes that display differential synthetic-lethal interactions with orthologs of genes implicated in tumorigenesis. Such genes could be targeted (e.g., NAE1) to enhance the killing power of chemotherapeutics in specific cancer types (e.g., p53 deficient cancers).

Finally, this study illustrates that differential network analysis is a powerful approach for annotating gene function that is complementary to existing functional genomics technologies. Widespread availability of genome-wide knockout/RNAi libraries coupled with advances in sequencing technology has driven down both the cost and effort required to phenotype a large collection of gene knockouts or mutations or conduct differential mRNA expression profiling across dozens of conditions. However, the resolution of these technologies is ultimately limited to the total number of genes in a genome. In contrast, differential interaction mapping taps into a much larger (quadratic) space of gene–gene interactions, which we have shown enables the dissection of gene function in greater detail (Figure 3.2D). This power comes at a cost,

as screening all gene pairs is presently arduous and expensive even in model organisms such as *S. cerevisiae*, requiring us to restrict coverage of the network map to a focused set of query genes. This tradeoff in precision versus coverage is analogous to the two complementary strategies that have been employed in mapping disease-causing mutations: analysis of genotyped pedigrees, involving no more than two to three generations, provides a 'coarse' mapping to identify a large candidate region of the genome[109], after which 'fine mapping' techniques such as gene association studies, which leverage large unrelated populations, are used to pinpoint the location of the causal mutation more precisely[110].

Here, we have pursued a similar strategy by seeding our differential genetic interaction screen with genes which have been previously annotated to high-level DDR processes. The resulting network highlights dynamic functional connections between numerous pathways and complexes at high resolution (Figure 3.5A), suggesting a new paradigm for dissecting the mechanism of action of a family of related drugs or cellular responses.

Chapter 3.5: Experimental Procedures

Chapter 3.5.1: Differential genetic interaction screens

Genetic interaction screens were performed as described[41], except that the last selection step was performed by replica-plating cells on medium containing 1% DMSO (Untreated), 0.01% MMS, 5µg/ml CPT or 75µM ZEO. Pictures were taken 48 hours after the final replication step. Colony sizes were then quantified using the HT

Colony Grid Analyzer program and genetic interaction scores (S Scores) were calculated using the E-MAP toolbox[10]. Differential p-values were calculated as previously described[16].

Chapter 3.5.2: Functional enrichment analysis

Both static and differential genetic networks were examined for enrichment of interactions with various sets of functionally related genes (Table S4). Significance was assessed using the hypergeometric distribution where the four parameters were defined as follows:

k. Total number of significant interactions containing a gene involved in a function of interest (e.g. DNA repair).

m. Total number of tested interactions containing a gene involved in a function of interest.

n. Total number of significant genetic interactions.

N. Total number of tested genetic interactions.

A significant differential genetic interaction was defined as having a differential p-value below 0.002. A significant static genetic interaction was defined as $S \geq 2.0$ or $S \leq -2.5$. The total number of tested interactions was the same across all networks (97,578). The enrichment results presented in Figure 2A were robust to the choice of significance threshold (Supplemental Figures 3.2B–C) as well as alternate definitions of DNA repair genes or chromatin organization genes (Supplemental Figure 3.2D).

Chapter 3.6: Acknowledgements

The authors thank G. Hannum, M. Chen, R. DeConde and N. de Wind for helpful discussions; J. Shen, M. Tijsterman and H. Vrieling for critical reading of the manuscript; H. Braberg and A. Roguev for assistance with the genetic interaction screens; and M. Lisby, KJ Myung, S. Ben-Aroya and P. Hieter for providing reagents. R.S. and T.I. were generously supported by grants from the U.S. National Institutes of Health (ES014811, GM084279). H.v.A. was supported by grants from the Netherlands Organization for Scientific Research (NWO-VIDI) and a CDA grant from Human Frontiers Science Program (HFSP).

Chapter 3, in full, is a reprint of a manuscript currently in revision: "Guénolé A*, Srivas R*, Vreeken K, Licon K, Wang Z.Z, Wang S, Krogan NJ., Ideker T., van Attikum H. *Dissection of DNA Damage Response Pathways using a Multi-Conditional Genetic Interaction Map*. Mol. Cell. In revision". The dissertation author was the primary investigator and author of this paper. For the sake of brevity, all Supplemental Data and Supplemental Tables can be found at the publication website.

Chapter 3.7: Supplemental Experimental Procedures

Chapter 3.7.1: Assessing the Quality of the Genetic Interaction Data

To ensure a high-quality dataset we examined several different quality control metrics:

(i) Correlation of replicate colony size measurements: Each query mutant in this screen was crossed against all array mutants six different times for each

condition (Untreated, MMS, ZEO, and CPT). Supplemental Figure 3.1A displays the histogram of the average correlation seen amongst the colony size measurements made across the six replicates for each query in each condition. The average Pearson's correlation seen amongst replicates was 0.78.

(ii) Correlation of NAT x KAN swaps: A subset of interactions in this dataset were screened twice with the only difference being the orientation of the drug resistance markers, i.e., *xxx*Δ::KAN *yyy*Δ::NAT versus *xxx*Δ::NAT *yyy*Δ::KAN. Each of these 'swaps' were scored independently. Across all four conditions, we observed a high correlation between the genetic interactions scores (S Score) for these 'swap' replicates (Supplemental Figures 3.2B-E). These correlation values are in line with previously published datasets[10,51].

(iii) Examination of linkage plots: Although each query mutant was checked via colony PCR for insertion of the drug resistance marker at the proper genomic location, it is not uncommon for ~10-15% of strains screened to be incorrect[10,111]. One useful tool for identifying these incorrect strains is to examine the interaction scores for pairs of genes that are located relatively close to one another on the genome. Due to linkage, such strains will fail to inherit both drug resistance markers following sporulation and as a result will appear as a negative interaction when plated on double selection media. As Supplemental Figure 3.1F shows gene-pairs located within 100 kbp of each other tended to exhibit a much more negative S score, indicating that the

majority of strains are indeed correct. Individual strains which deviated from this trend were identified and removed from the final dataset.

Chapter 3.7.2: Assessing the False Discovery Rate (FDR) of Differential Genetic Interactions

We assessed the false discovery rate (FDR) of the differential genetic interactions using two different methods. In the first method, we corrected the differential p-value assigned to each gene-pair across all three conditions using the Benjamini-Hochberg procedure. At $P \leq 0.002$, the threshold used in this study, we observed an FDR of 6.2%, 17.4%, and 13.2% respectively for the MMS, CPT, and ZEO differential networks (Supplemental Figure 3.1G). As an alternate method, we obtained five previously published E-MAP datasets, which had been generated in untreated conditions[13,14,16,112,113], and scored each pair of networks (which had tested at least 100 interactions in common) for differential interactions. Any interaction which appears significant in this analysis is essentially a false discovery since the comparison is being made between genetic interactions measured under the same condition. Supplemental Figure 3.1H shows the average false discovery rate across all pair-wise comparisons as a function of the differential significance threshold. At a threshold of $P \leq 0.002$, we observed an average FDR of 8.2%.

Chapter 3.7.3: Description of Single Mutant and Gene Expression Datasets Used In This Study

We obtained single mutant data from a previously published chemogenetic screen which had examined the fitness of 4,722 homozygous diploid mutants in response to hundreds of different compounds[65]. To ensure the closest comparison with the perturbations in this study, we used fitness data generated under the following concentrations: MMS (0.002%) and CPT (30 µg/mL). For ZEO, we used the fitness data generated under bleomycin (1.7 µg/mL), a chemical compound which induces similar effects to zeocin. Differential gene expression data for MMS (0.12%) and ZEO (again using data generated under bleoymycin [0.15 U/ml]) was obtained from[78]. Expression data for CPT was obtained from [77].

Chapter 3.7.4: Testing for Association Between Agent and DNA Repair Pathway

For each experimental technology, we generated a 3x6 contingency table (C), where each cell ($C_{i,j}$) was defined as follows:

Differential networks: $C_{i,j}$ is the number of significant differential interactions observed in condition *i* containing a gene in pathway *j*.

Single mutant data: $C_{i,j}$ is the number of genes displaying a significant sensitivity in condition *i* which also fall in pathway *j*.

Gene expression data: $C_{i,j}$ is the number of differentially expressed genes observed in condition *i* which are also in pathway *j*.

where i ∈ {MMS, CPT, ZEO} and j ∈ {Double-strand break repair, DNA damage checkpoint, nucleotide excision repair, mismatch repair, post-replication repair, base excision repair}. All pathway definitions have been provided in Table S4.

A standard chi-square statistic (X^2) was then computed on this contingency table. To assess significance, the chi-square statistic was re-calculated over 1000 permutations in which the set of sensitive genes or differentially expressed genes were randomly re-assigned to a different compound, while ensuring (i) the total number sensitive/differentially expressed genes seen in the actual data was maintained in each permutation and, (ii) any dependencies seen amongst compounds was maintained in the permutation, i.e., if there were 40 differentially expressed genes seen in both MMS and CPT, the same number was maintained in each permutation.

For the differential networks, we generated 1000 randomized networks by scrambling the node labels, after which the chi-square statistic was re-computed. This null distribution of chi-square statistics was subsequently used to assign each experimental technology a p-value for the association between agent and pathway.

Chapter 3.7.5: Spot Dilution assays

Cells were grown to mid-log phase in YPAD and 10-fold serial dilutions were spotted on YPAD plates containing the drug of interest. Images of the colonies were taken after 2 – 3 days at 30°C.

Chapter 3.7.6: Cell cycle checkpoint analysis

Exponentially growing cells were synchronized in G1 with α-factor (7.5μM). Cells were either exposed to MMS in G1 for 30 minutes and then released in fresh medium, or released in fresh medium containing CPT. FACS analysis was performed using a BD™ LSRII instrument and FACS data were processed using WinMDI software. Rad53 phosphorylation analysis was performed as previously described using anti-Rad53 antibody (Santa Cruz Biotechnologies, SC-6749)[114]. Membranes were imaged using the Biorad Universal Hood II instrument and signal intensities were quantified using the Quantity One software package.

Chapter 3.7.7: GCR and Mutagenesis assays

Gross chromosomal rearrangement and mutagenesis assays were performed as previously described previously[115,116].

Chapter 3.7.8: Analysis of Mms22 turnover

Gal1-HA-MMS22 containing cells were grown in rich medium 3% glycerol, 2% lactic acid and 0.05% glucose for 12 hours. Next, 2% galactose was added to induce HA-MMS22 expression for 3 hours. Finally, cells were washed and resuspended in glycerol and lactic acid-based rich medium containing 2% glucose. HA-Mms22 expression was analyzed by western blot analysis using an anti-HA antibody (Santa Cruz Biotechnology, SC-7392).

Chapter 3.7.9: Analysis of Rad52 foci

Cells containing a Rad52-YFP expression vector were grown to mid-log phase, exposed to MMS for 1 hour, washed and concentrated in 1% low melting agar (Cambrex). Images were captured using a Leica AF6000 LX microscpe at 100-fold magnification using a HCX PL FLUOTAR 100x 1.3 oil objective lens.

Chapter 3.7.10: Integrative analysis of differential genetic interactions and protein interactions

Protein interactions were obtained from a previous integration of several primary protein interaction screens[57], from which we selected interactions with PE >= 2.0. Each static network was analyzed using a previously published workflow to identify multi-genic modules, i.e. sets of genes spanned by many physical and genetic interactions[97]. This list was further augmented with a set of literature-curated protein complexes[59] resulting in a final set of 332 modules after removing overlapping modules (Table S5). To identify functional relationships between modules, we searched for enrichment of differential genetic interactions between each module pair. Significance was assessed using the hypergeometric distribution as previously described[117]. The full list of module-module interactions ($P \leq 0.05$) is provided in Table S6.

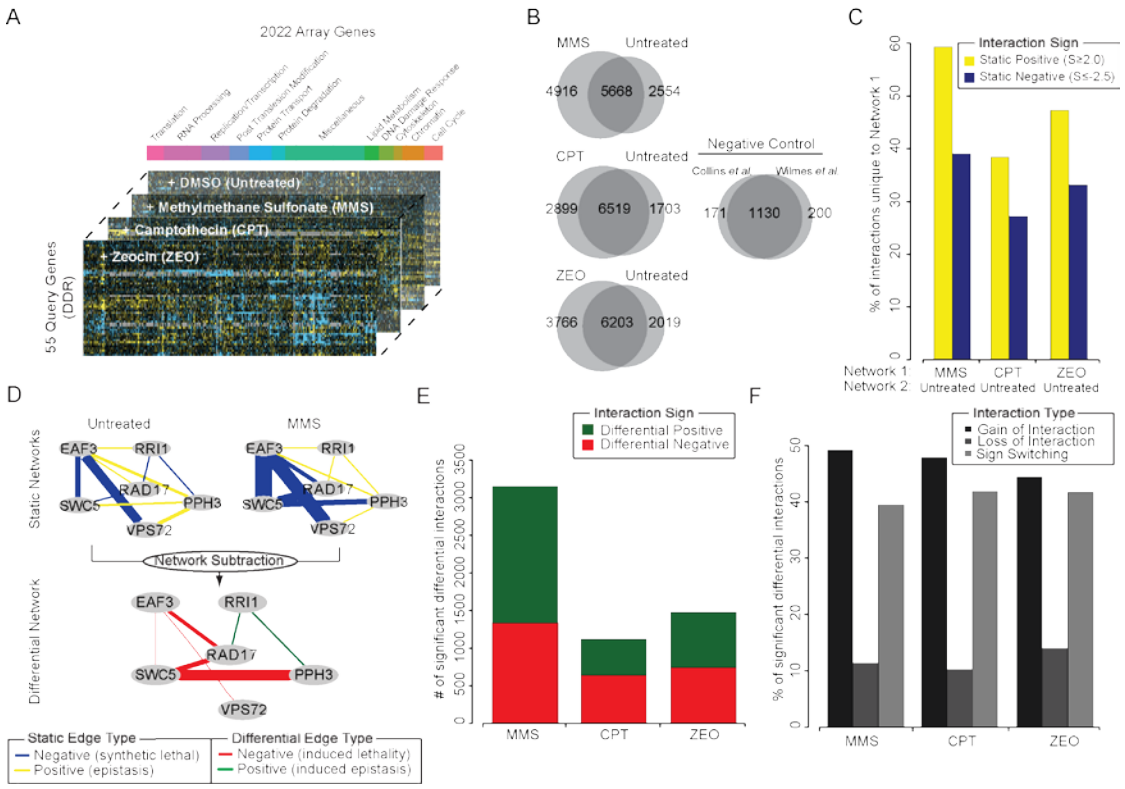


Figure 3.1: Overview of the Multi-Conditional Differential Network

(A) Experimental design of the differential genetic interaction screen. The stacked barplot illustrates the functional breakdown of array genes. A full list of the query and array genes is provided in Tables S1 and S2. (B) Overlap in significant static interactions ($S \geq 2.0$, $S \leq -2.5$) between treated and untreated conditions. The negative control represents the overlap seen amongst previously-published networks measured in untreated conditions [13,103]. (C) The percentage of positive and negative interactions that are unique to the treated network (Network 1) when compared to the untreated network (Network 2) for all three conditions. (D) Schematic overview of how differential genetic networks are derived by examining the difference between static treated and untreated genetic networks. The thickness of the edge scales with the magnitude of the genetic interaction. (E) Overview of the number of significant positive and negative differential interactions uncovered in each condition. (F) Breakdown of significant differential interactions into three categories: 'Gain of Interaction', 'Loss of Interaction', and 'Sign Switching' (see text).

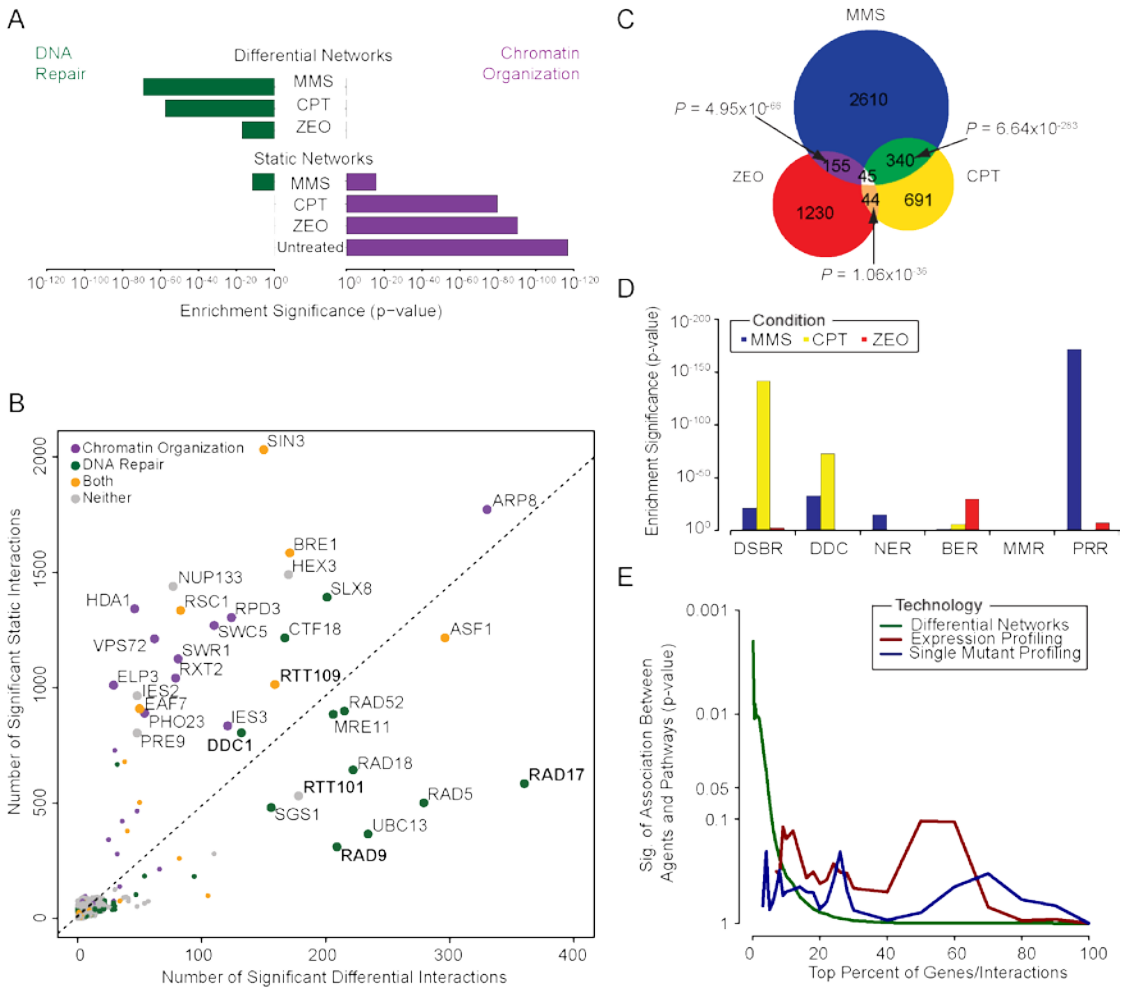
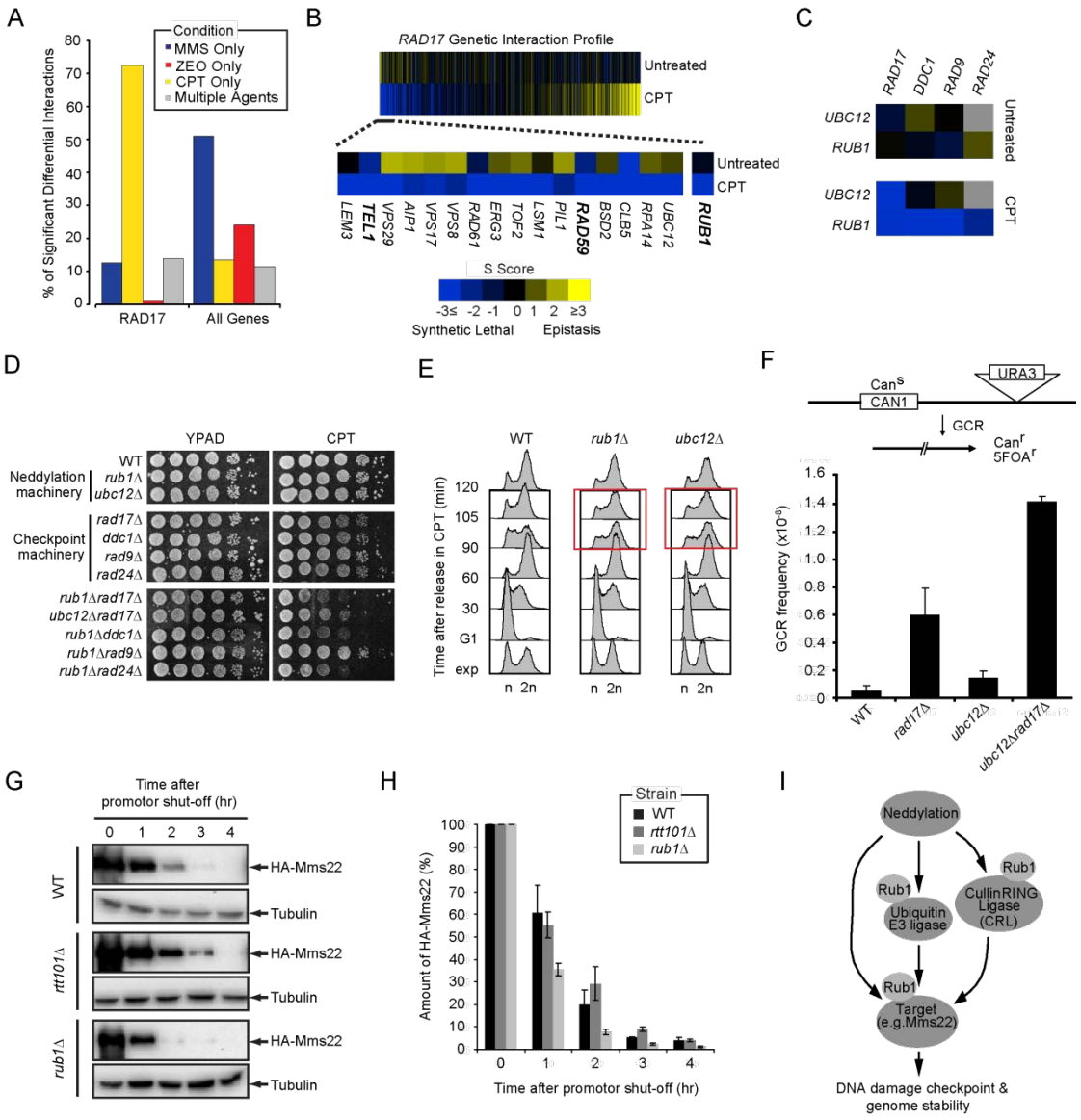


Figure 3.2: Differential networks reveal specific pathways induced by different types of DNA damage

(A) The significance of enrichment for interactions with genes that function in either DNA repair (green bars) or chromatin organization (purple bars) is plotted for all static and differential genetic networks. (B) The total number of significant static (y-axis) and differential (x-axis) interactions for each gene considered in this study. (C) The differential networks have very different patterns of interactions. The overlap in significant ($P \leq 0.002$) differential interactions induced by each agent (MMS, CPT, and ZEO) is shown. (D) Enrichment of differential interactions containing genes involved in six different specific DNA repair pathways: double-stranded break repair (DSBR), DNA damage checkpoint (DDC), nucleotide excision repair (NER), base excision repair (BER), mismatch repair (MMR) and post-replication repair (PRR). (E) The significance of association between agents and DNA repair pathways is computed using differential networks, single mutants, and differential gene expression across a range of thresholds.

Figure 3.3: Neddylation regulates cell cycle progression after DNA damage and preserves genome integrity

(A) Percentage of RAD17's significant differential genetic interactions arising in response to MMS, CPT, ZEO, or multiple agents. As a control, the average percentage of significant differential interactions in each of these categories across all genes is shown. (B) Entire CPT-induced genetic interaction profile for *RAD17* sorted (left to right) in order of most differential negative to most differential positive. A subset of the top differential negative interactions is also shown. (C) Genetic interactions seen between components of the neddylation machinery and the DNA damage checkpoint. (D) Viability of cells deficient for both neddylation and checkpoint processes is impaired in the presence of CPT. 10-fold serial dilutions of log-phase cells of the indicated genotypes were spotted onto YPAD and YPAD containing CPT (15 μM) and incubated for 3 days at 30°C. (E) Neddylation mutants display perturbed G2/M progression in the presence of CPT. Exponentially (exp) growing WT, *rub1*Δ and *ubc12*Δ cells were arrested in G1 with α-factor and released in fresh medium containing 50 μM CPT. Cells were analyzed by FACS at the indicated timepoints. (F) Cells deficient for both neddylation and checkpoint processes have increased Gross Chromosomal Rearrangements (GCR). GCR frequencies were determined as described in the Supplemental Experimental Procedures. The mean ± standard deviation of three independent experiments is presented. (G) Neddylation-deficient cells display a more rapid turnover of the G2/M checkpoint protein Mms22. The expression of *GAL1*-HA-Mms22 expression was induced in WT, *rtt101*Δ and *rub1*Δ cells by growing the cells in 2% galactose for 3 hours. Cells were released in 2% glucose to shut-off expression of HA-Mms22, after which levels of HA-Mms22 were monitored by Western blot analysis. (H) Bar-plot showing the rate of HA-Mms22 protein degradation in WT, *rtt101*Δ and *rub1*Δ. The levels of HA-Mms22 protein were quantified and normalized to tubulin. The ratio at the start of shut-off was set to 100%. The mean ± standard deviation of four independent experiments is presented. (I) Schematic illustrating proposed mechanisms by which the Neddylation machinery may regulate DNA damage checkpoints and genome stability. See text for details.


A
Condition
MMS Only
ZEO Only
CPT Only
Multiple Agents
% of Significant Differential Interactions
RAD17
All Genes
B
RAD17 Genetic Interaction Profile
Untreated
CPT
LEM3
TEL1
VPS29
AIP1
VPS17
VPS8
RAD61
ERG3
TOF2
LSM1
PIL1
RAD59
BSD2
CLB5
RPA14
UBC12
RUB1
S Score
Synthetic Lethal
Epistasis
C
RAD17
DDC1
RAD9
RAD24
UBC12
RUB1
Untreated
CPT
D
YPAD
CPT
Neddylation machinery
WT
rub1Δ
ubc12Δ
Checkpoint machinery
rad17Δ
ddc1Δ
rad9Δ
rad24Δ
rub1Δrad17Δ
ubc12Δrad17Δ
rub1Δddc1Δ
rub1Δrad9Δ
rub1Δrad24Δ
E
WT
rub1Δ
ubc12Δ
Time after release in CPT (min)
G1
exp
n 2n
F
Can^S
CAN1
URA3
GCR
Can^r
5FOA^r
GCR frequency (x10^-8)
WT
rad17Δ
ubc12Δ
ubc12Δrad17Δ
G
Time after promotor shut-off (hr)
HA-Mms22
Tubulin
WT
rtt101Δ
rub1Δ
H
Strain
WT
rtt101Δ
rub1Δ
Amount of HA-Mms22 (%)
Time after promotor shut-off (hr)
I
Neddylation
Rub1
Ubiquitin E3 ligase
CullinRING Ligase (CRL)
Target (e.g.Mms22)
DNA damage checkpoint & genome stability

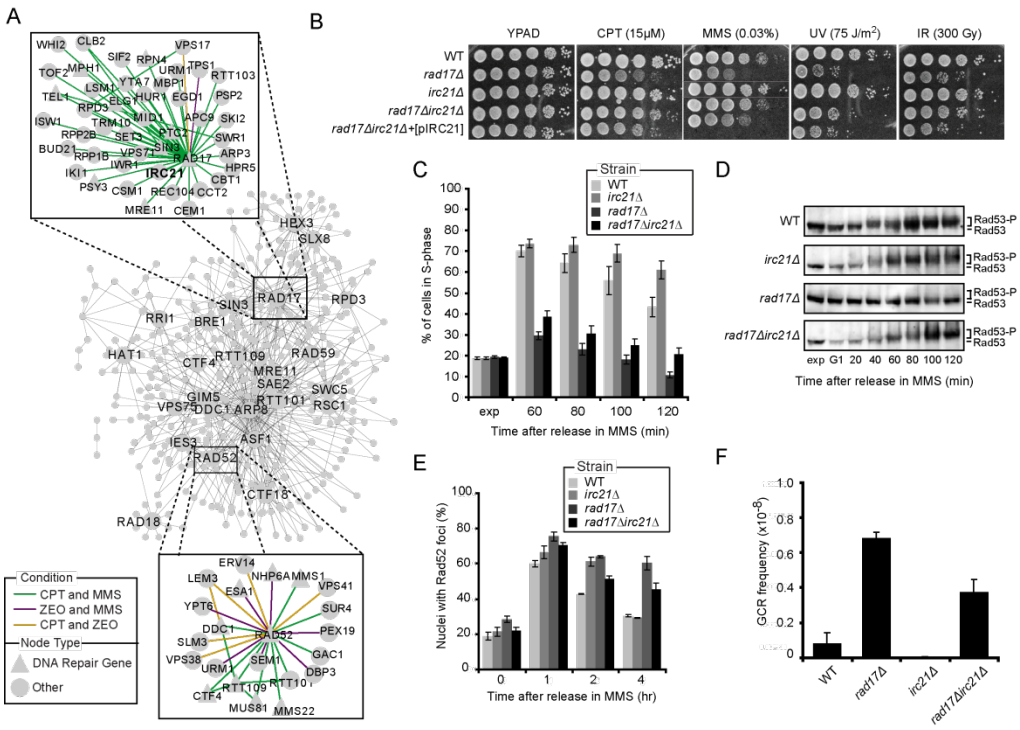


Figure 3.4: Irc21 affects checkpoint control, DNA repair and genome stability

(A) Network of all 584 differential genetic interactions induced by at least two agents. The top 25 hubs in this network have been labeled. The sub-network of interactions involving *RAD17* and *RAD52* is also shown. (B) *IRC21* deletion rescues the viability of *rad17*Δ cells in the presence of DNA damage. 10-fold serial dilutions of log-phase cells of the indicated genotypes were either spotted onto YPAD plates containing 0.03% MMS, 15 μM CPT, or spotted on YPAD and exposed to UV (75J/m^2) or IR (300 Gy), followed by incubation for 3 days at 30°C. (C) Exponentially (exp) growing WT, *irc21*Δ, *rad17*Δ, *rad17*Δ*irc21*Δ cells were arrested in G1 with α-factor and released in fresh medium containing 0.02% MMS and 15μg nocodazole. Cells were analyzed by FACS at the indicated timepoints (see Figure S4C for FACS plots). The bar-plot shows the percentage of S-phase cells. The mean ± standard deviation of three independent experiments is presented. (D) Western blot analysis of Rad53 phoshorylation in cells from (C). (E) Irc21 affects the dissolution of MMS-induced Rad52 foci. Exponentially growing WT, *irc21*Δ, *rad17*Δ, *rad17*Δ*irc21*Δ cells expressing Rad52-YFP were exposed to 0.02% MMS for 1h and then released in fresh medium. Images were taken at the indicated timepoints and scored for the presence of Rad52-YFP foci. At least 100 nuclei were analyzed per strain and per time point. Data represent the mean ± standard deviation from three independent experiments. (F) Irc21 affects genomic instability. GCR frequency was determined in WT, *irc21*Δ, *rad17*Δ, *rad17*Δ *irc21*Δ cells. Data represent the mean ± standard deviation from three independent experiments.

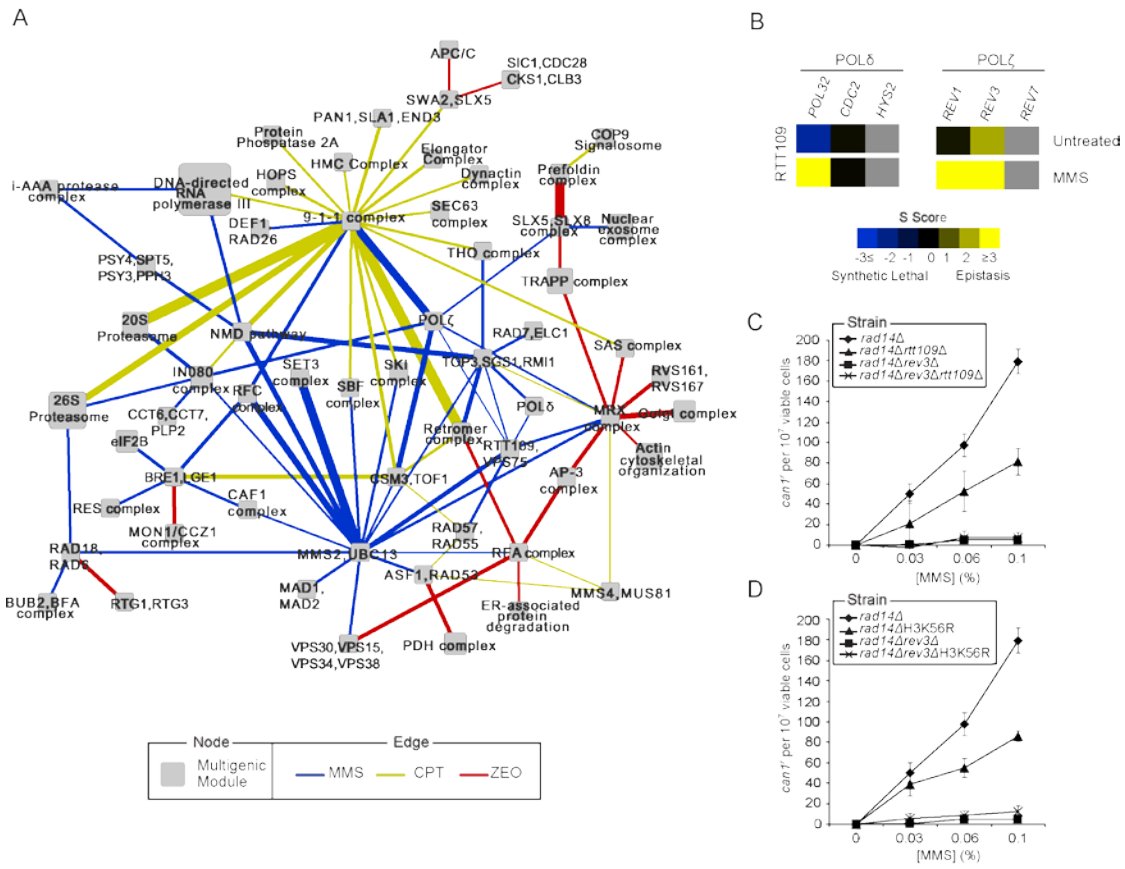
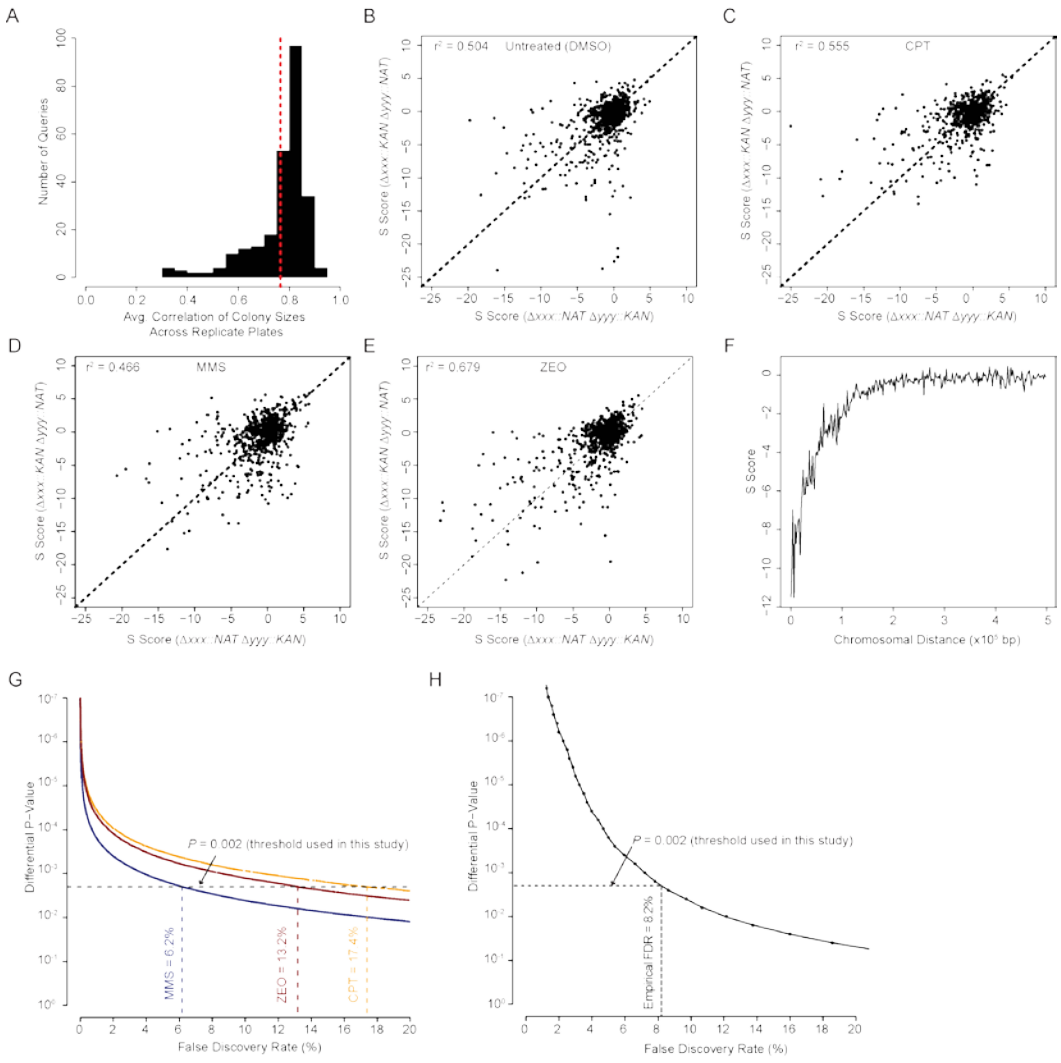


Figure 3.5: A global map of DDR modules reveals a novel role for RTT109 in translesion synthesis

(A) A map of multi-protein modules connected by bundles of differential genetic interactions. Node size scales with the number of proteins present in the module. Edge size scales with the significance of the enrichment for differential interactions spanning the two modules. For the sake of clarity only a portion of the entire map has been shown. The full list of module-module interactions is provided in Table S6. (B) Genetic interactions observed between Rtt109 and members of the Polδ and Polζ complexes. (C) Rtt109 and H3K56 acetylation affect MMS-induced mutagenesis in NER-defective *rad14*Δ cells. MMS-induced *can1*^{r} mutation frequencies were examined in *rad14*Δ, *rad14*Δ*rtt109*Δ, *rad14*Δ*rev3*Δ and *rad14*Δ*rev3*Δ*rtt109*Δ cells. (D) As in C, except that *rad14*ΔH3K56R and *rad14*Δ*rev3*ΔH3K56R cells were used. The data represent the mean ± standard deviation of three independent experiments.

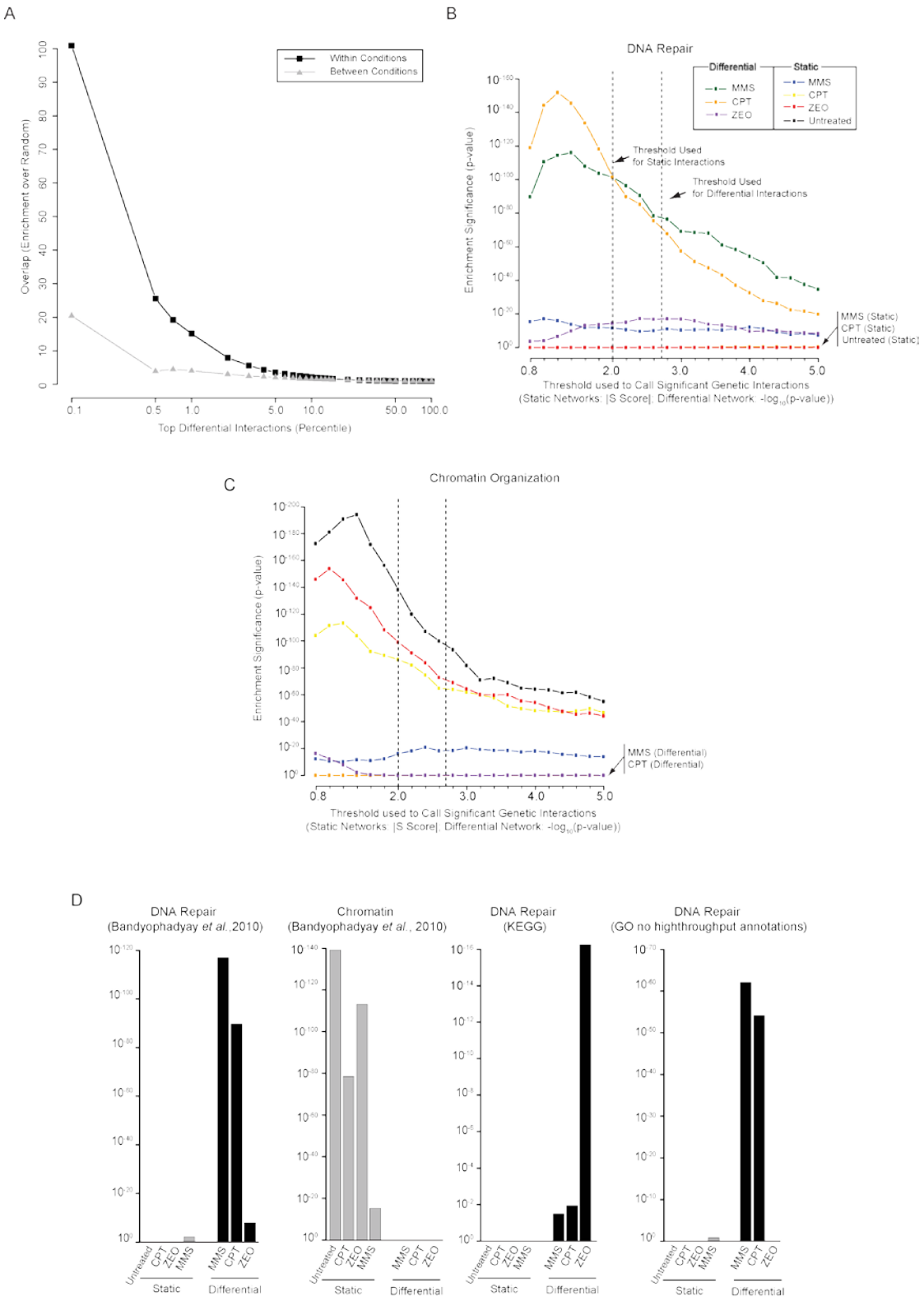


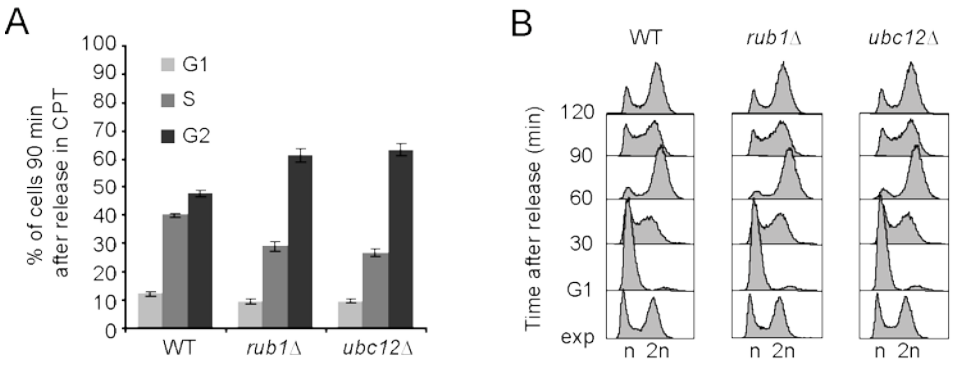
Supplemental Figure 3.1: Quality of Genetic Interaction Data.

(A) Each query mutant was crossed against the set of array mutants six different times. This histogram displays the average correlation seen in colony size measurements between the six replicates for each query across all four conditions (Untreated, MMS, CPT, and ZEO). The dotted red-line indicates the average correlation seen across all queries and all conditions (r^2 = 0.78). (B–E) Correlation of genetic interaction scores derived from 'marker swap' experiments for (B) Untreated, (C) CPT, (D) MMS, and (E) ZEO. (F) Genetic interaction scores for pairs of genes in linkage. (G) For each condition, the differential p-value (x-axis) is plotted versus the corresponding multiple hypothesis corrected false discovery rate (FDR) (corrected using the Benjamini-Hochberg procedure). (H) Same as (G), but FDR was determined empirically by comparing five previously published genetic interaction datasets (see Supplementary Methods).

Supplemental Figure 3.2: Comparison of Replicate Differential Networks and Robustness of Functional Enrichment Results

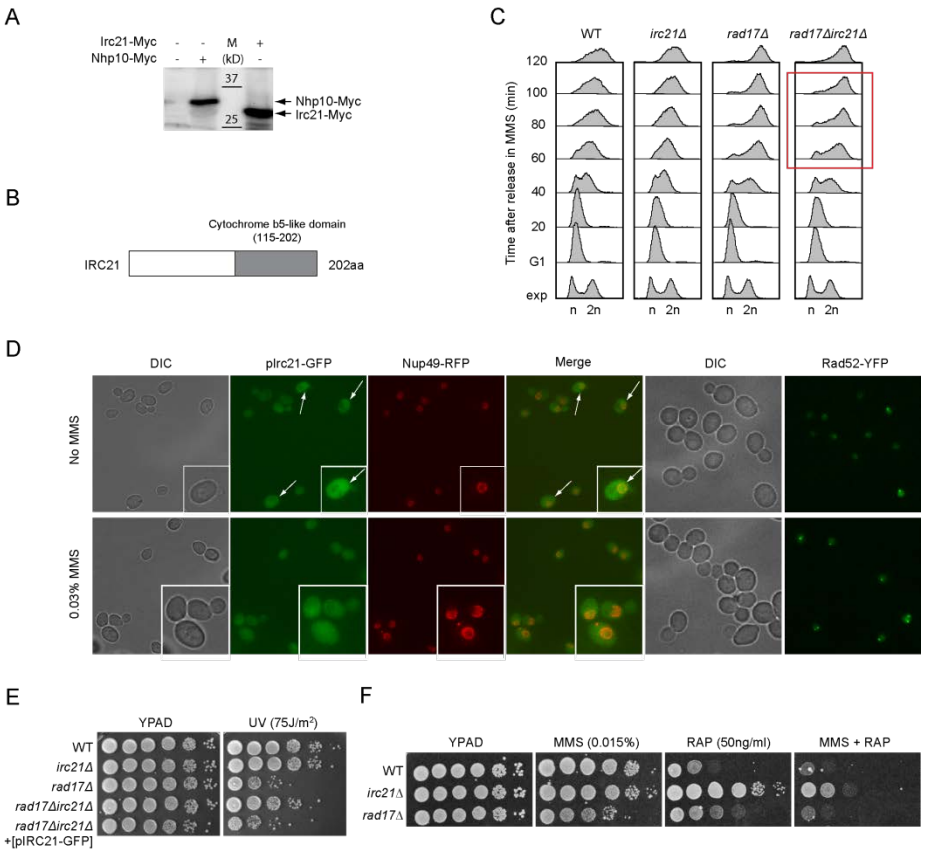
(A) The overlap in replicate differential networks seen amongst the same condition (black line) or between two different conditions (grey line). Replicate networks are derived by splitting the six replicates obtained for each double mutant into two sets and scoring each set independently. Enrichment over random (y-axis) is defined as the ratio of overlapping interactions seen amongst the top percent of differential interactions (x-axis) to the number of overlapping interactions expected at random. (B – C) The significance of enrichment with either (B) DNA Repair or (C) chromatin organization genes is plotted for all static and differential genetic networks across a range of thresholds. For static networks the absolute-value of the S Score is used as a threshold. For differential networks the $-\log_{10}$(differential p-value) is used as a threshold. (D) Enrichment results using different databases to define DNA repair and chromatin organization gold-standard genes.


A
Within Conditions
Between Conditions
Overlap (Enrichment over Random)
Top Differential Interactions (Percentile)
B
DNA Repair
Differential
Static
MMS
CPT
ZEO
Untreated
Threshold Used for Static Interactions
Threshold Used for Differential Interactions
Enrichment Significance (p-value)
MMS (Static)
CPT (Static)
Untreated (Static)
Threshold used to Call Significant Genetic Interactions
(Static Networks: |S Score|; Differential Network: -log10(p-value))
C
Chromatin Organization
Enrichment Significance (p-value)
MMS (Differential)
CPT (Differential)
Threshold used to Call Significant Genetic Interactions
(Static Networks: |S Score|; Differential Network: -log10(p-value))
D
DNA Repair (Bandyophadyay et al.,2010)
Chromatin (Bandyophadyay et al., 2010)
DNA Repair (KEGG)
DNA Repair (GO no highthroughput annotations)
Untreated
CPT
ZEO
MMS
Static
Differential



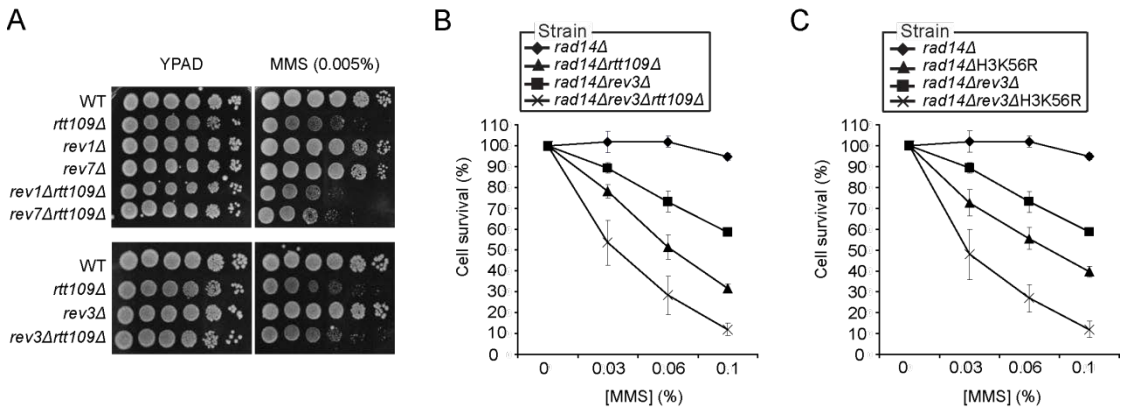
Supplemental Figure 3.3: Neddylation regulates mitotic progression after DNA damage

(A) Quantification of FACS data from Figure 3E. The bar-plot represents the percentage of WT, *rub1*Δ and *ubc12*Δ cells in G1, S or G2 phase 90 minutes after their release from G1 in fresh medium containing CPT. Data represent the mean ± standard deviation from three independent experiments. (B) Exponentially (exp) growing WT, *rub1*Δ and *ubc12*Δ cells were arrested in G1 with α-factor and released in fresh medium. Cells were analyzed by FACS at the indicated time points.



Supplemental Figure 3.4: Irc21 localizes in both the cytoplasm and nucleus and may be linked to autophagy

(A) Western blot analysis of cells expressing Myc-tagged Irc21. Cells from non-tagged and Nhp10-Myc expressing strains were used as negative and positive controls, respectively. (B) Schematic of the Irc21 protein showing a cytochrome b5-like domain in its C-terminus. **(**C**)** Exponentially (exp) growing WT, *irc21*Δ, *rad17*Δ, *rad17*Δ*irc21*Δ cells were arrested in G1 with α-factor and released in fresh medium containing 0.02% MMS and 15μg nocodazole. Cells were analyzed by FACS at the indicated time points**.** (D) Exponentially growing *irc21*Δ cells expressing Irc21-GFP and Nup49-RFP were treated with 0.03% MMS for 1 hour and then examined for Irc21 localization. Wild-type cells expressing Rad52-YPF were treated similarly and examined for Rad52 focus formation. (E) Ectopic expression of Irc21-GFP in *rad17*Δ*irc21*Δ renders cells as sensitive to UV as *rad17*Δ cells, demonstrating the functionality of GFP-tagged Irc21. 10-fold serial dilutions of log-phase cells of the indicated genotypes were spotted onto YPAD plates and exposed to UV (75J/m^2) followed by incubation for 3 days at 30°C. (F) *irc21*Δ cells are hypersensitive to MMS when combined with the TOR inhibitor rapamycin (RAP). 10-fold serial dilutions of log-phase cells of the indicated genotypes were either spotted onto YPAD plates containing 0.015% MMS, 50 ng/ml RAP or both and incubated for 3 days at 30°C.



Supplemental Figure 3.5: Integrative Analysis of Differential Genetic Interactions Reveals a Role for RTT109 in Translesion Synthesis.

(A) *RTT109* displays epistatic interactions with components of the polymerase ζ complex (*REV1*, *REV3*, *REV7*) in the presence MMS. 10-fold serial dilutions of log-phase cells of the indicated genotypes were spotted onto YPAD plates containing 0.005% MMS and incubated for 3 days at 30°C. (B) MMS survival of NER-deficient *rad14*Δ, *rad14*Δ*rev3*Δ, *rad14*Δ*rtt109*Δ and *rad14*Δ*rev3*Δ*rtt109*Δ, and of (C) NER-deficient *rad14*ΔH3K56R and *rad14*Δ*rev3*ΔH3K56R cells were examined. Log-phase cells were exposed for 20 minutes to the indicated MMS concentrations. Appropriate dilutions of were plated on SC(-Arg) and colony formation was scored. The data represent the mean standard deviation of three independent experiments.

Chapter 4. Genome-wide association data reveal a global map of genetic interactions among protein complexes

Chapter 4.1: Abstract

This work demonstrates how gene association studies can be analyzed to map a global landscape of genetic interactions among protein complexes and pathways. Despite the immense potential of gene association studies, they have been challenging to analyze because most traits are complex, involving the combined effect of mutations at many different genes. Due to lack of statistical power, only the strongest single markers are typically identified. Here, we present an integrative approach that greatly increases power through marker clustering and projection of marker interactions within and across protein complexes. Applied to a recent gene association study in yeast, this approach identifies 2,023 genetic interactions which map to 208 functional interactions among protein complexes. We show that such interactions are analogous to interactions derived through reverse genetic screens, and that they provide coverage in areas not yet tested by reverse genetic analysis. This work has the potential to transform gene association studies, by elevating the analysis from the level of individual markers to global maps of genetic interactions. As proof of principle, we use synthetic genetic screens to confirm numerous novel genetic interactions for the INO80 chromatin remodeling complex.

Chapter 4.2: Author Summary

One of the most important problems in biology and medicine is to identify the catalog of genes that underlie disease. Currently, genome-wide association studies are employed to detect genes that are associated with a particular disease or phenotype. However, this approach to date has yielded very few genes leading to growing interest in mapping gene-gene interactions that may be more prevalent due to the complex nature of most diseases. Mapping epistatic interactions, however, presents numerous challenges. The search space of all possible gene-gene interactions that must be examined is enormous leading to a loss in statistical power to detect such interactions. Furthermore, simple lists of gene-gene interactions provide no insight into the molecular mechanisms behind these interactions. Here, we present a fundamentally new approach that leverages genome-wide association data to reveal how gene-gene interactions function together within a unified map of complexes and pathways. This map elevates the analysis of genome-wide association studies from the level of individual gene-gene interactions to functional interactions between physical complexes. We demonstrate that such pathway-based interpretations provide novel hypothesis regarding the mechanism through which combinations of polymorphisms may affect a phenotype.

Chapter 4.3: Introduction

A central challenge in genetics is to understand how interactions among different genetic loci contribute to complex traits[7,9,11,13,31,39,118]. In model organisms

such as yeast, genetic interactions have been rapidly elucidated using reverse genetic approaches, in which double gene knockouts are introduced into various strains and subsequently scored for their effects on fitness. Genetic interaction is indicated when the growth rate of the double mutant is slower than expected (e.g., synthetic sickness or lethality) or faster than expected (e.g., suppression)[7,8,10]. The systematic and high-throughput screening of such interactions has been made possible through a variety of methods including Synthetic Genetic Array (SGA) analysis[7], diploid Synthetic Lethality Analysis by Microarray (dSLAM)[11], and epistatic miniarray profiles (E-MAP)[9,13,14,51].

While the above techniques have been instrumental in model organisms, performing genetic interaction analysis in higher eukaryotes has been less straightforward. First, genetic screens have relied on easy-to-measure cell-based phenotypes, such as fitness in rich growth conditions. However, genetic interactions governing complex traits in humans (such as body weight, blood pressure, or incidence of disease) are difficult to study using cell-based assays and are highly condition-dependent. Second, systematically engineering a series of double gene disruptions in mammals remains technically difficult, although combinatorial RNAi knockdowns show promise in this regard[44].

As an alternative to engineered genetic perturbations, high-throughput genotyping and sequencing platforms have made it possible to characterize the millions of polymorphic genetic markers present in the genome. Genome-wide linkage or genome-wide association studies (GWAS) attempt to identify polymorphic markers

that have associations, ideally causal associations, with a phenotype of interest[23]. Numerous technologies are currently available for measuring upwards of 10^5 Single Nucleotide Polymorphisms (SNPs) in the human genome[24].

Similar to reverse genetic interaction screens, GWAS have been screened for pairs of loci underlying complex trait variation[25,29,119]. Mapping pair-wise locus associations has proven remarkably difficult, however. The most basic approach is to perform an exhaustive two-dimensional (2D) scan, in which all pairs of genetic markers are tested for joint association with the phenotype. Because billions of marker pairs must be tested, 2D scans are computationally demanding and suffer from low statistical power due to multiple hypothesis testing. One method to address this problem is to initiate searches for pair-wise interactions only for markers with strong individual effects[26,27]. Two recent studies by Storey *et al.* and Litvin *et al.* used this approach while accounting for information shared across multiple traits to further enhance statistical power[28,120]. These results indicate a major role for genetic interactions in the heritability of complex traits. However, it is likely that the interactions uncovered to date represent only a fraction of the true genetic network.

Here, we show that both the power and interpretation of genetic interactions derived from association studies can be significantly improved through integration with information about the physical architecture of the cell. We apply this integrative approach to an association study conducted in yeast, yielding a genetic network that complements, extends, and validates networks assembled through reverse genetic methods.

Chapter 4.4: Results

Chapter 4.4.1: Bi-clustering of marker pairs defines a network among genomic intervals

We analyzed a recent GWAS in yeast which analyzed a population of 112 segregants resulting from a cross of a laboratory *S. cerevisiae* strain with a wild isolate[31]. For each segregant, the states of 1,211 unique markers (genotypes) were mapped along with the expression profile of 5,727 genes (traits) (Table S1). To identify pairs of markers that genetically interact— i.e. for which the joint state of the marker pair was associated with one or more gene expression traits— we considered the method of Storey *et al*.[28] which provides the best marker pair for each expression trait, resulting in a set of 4,687 distinct marker-marker interactions (removing redundancies due to marker pairs that associate with multiple traits).

A preliminary examination of the genotype data showed few recombinations between neighboring markers, indicating that markers in close proximity were in linkage disequilibrium (LD). As a result, neighboring markers were often found to display similar patterns of interactions (Figure 4.1A). In much the same way that LD has allowed neighboring markers to be grouped into haplotype blocks[121], we reasoned that LD between neighboring markers could also be exploited to enhance marker-marker interactions. To this end, we developed a bi-clustering algorithm to identify groups of marker-marker interactions that fall across common genomic intervals (Figure 4.11B**;** see Methods). We reasoned that bi-clustering the marker pairs might provide two distinct advantages: First, it allows many statistically insignificant

marker-marker interactions to reinforce a single interval-interval interaction. Second, it leverages the structure between neighboring marker pairs to identify with greater precision the interval of DNA underlying the variance in a given trait.

Applied to the marker pairs from Storey *et al.*, the bi-clustering procedure yielded a network of 2,023 interactions between 1,977 genomic intervals (Figure 4.1C). Of these, 695 interval pairs garnered support from multiple marker pairs (five on average). The remaining 1,328 interval pairs consisted of singleton marker-marker interactions, which were not found to cluster with any others. The complete network of interval-interval interactions can be found in Table S2. We refer to this network as a natural genetic network since it is derived from natural rather than engineered mutations.

Chapter 4.4.2: Natural interactions define a map of functional links between protein complexes

A common interpretation of genetic interactions measured in reverse genetic screens has been the “between-complex” or “between-pathway” model, in which interactions are found to span pairs of protein complexes or functional annotations. Such complex-complex interactions have been instrumental in identifying synergistic or compensatory relationships[7,8,19]. Similarly, pairs of functional terms have served to identify functions that are cooperative or buffer one another[7].

To evaluate natural networks in this fashion, we examined all pairs of documented protein complexes (out of 302 in Gavin *et al.*[5] or the Munich Information

Center for Protein Sequences [MIPS][122]) and all pairs of functional terms (out of 1,954 terms in the Gene Ontology [GO][52]) for enrichment for natural genetic interactions. As further described in Methods, we inspected all complex pairs and found 208 significant interactions in the natural network (False Discovery Rate < 5%; Table 1). Similarly, we identified 17,714 significant interactions between functional terms. In contrast, far fewer results were found for complex or term interactions derived from the raw marker pairs of Storey *et al*. prior to bi-clustering these data into intervals (Table 4.1). The full set of complex-complex and term-term interactions are available as a resource in Table S3 and on the Supplemental Website (http://www.cellcircuits.org/qtlnet/).

Figure 4.2A shows a map of the 50 most significant complex-complex interactions. Because gene expression is the phenotypic trait, each complex-complex interaction is linked to a cluster of gene expression levels that it regulates (with each cluster containing an average of 287 genes). As the map integrates many traits simultaneously, it is distinct from previously-published genetic networks which have relied on cell viability as the single readout of interest. We found that two-thirds of the complex-complex interactions were linked to gene expression clusters that were highly functionally coherent (Figure 4.2A). In contrast, less than one one-hundredth of interval-pairs were found to influence a set of genes belonging to a single pathway or function. Thus, we conclude that integration of epistatic interactions with protein complex maps helps to filter spurious interactions while simultaneously providing a putative mechanism for the pair-wise associations.

As an illustrative example, Figure 4.2B shows the natural genetic interactions supporting a functional link between the synaptonemal complex and RNA Polymerase II. Mutations in the *TOP2* gene of the synaptonemal complex have been shown to lead to higher levels of mitotic recombination in rDNA which can result in amplification and deletion of the rDNA array[123]. RNA polymerase II is responsible for the transcription of small nucleolar RNAs (snoRNAs) that physically and functionally interact with many other proteins required for ribosomal biogenesis[124]. Indeed, we found that the gene expression traits linked to this interaction were enriched for ribonucleoprotein complex biogenesis and ribosome biogenesis (both $P' = 10^{-8}$ by hypergeometric test; P' is a Bonferroni corrected p-value).

Figure 4.2C centers on the Tim9-Tim10 complex, an essential component of the TIM machinery responsible for the transport of carrier proteins from the cytoplasm to the inner mitochondrial membrane[125]. Tim9-Tim10 is genetically connected with two other complexes, Mannan Polymerase II and the TRAPP complex. Mannan Polymerase II is a component of the secretory pathway and is involved in lengthening the mannan backbone of cell wall and periplasmic proteins[126]; the TRAPP complex plays an important role in trafficking of proteins from the golgi to the cell periphery[127]. The abundant genetic interactions between Tim9-Tim10 and these two complexes suggest they may jointly influence the make-up of cell surface proteins, possibly through control of trafficking. Consistent with this hypothesis, disruption of mitochondrial function has been shown to influence cell wall composition, including levels of phosphopeptidomannans[128].

For comparison to the between-complex model, we also examined the natural genetic network for support for a "within-complex" model, in which single functional terms or complexes are enriched for genetic interactions among their member genes[7,8,19]. Searching across the 1,954 GO terms and 302 complexes, the natural network identified only 12 enriched GO terms and no significant complexes (Table 4.1 and Table S3). Thus, genetic interactions in naturally-derived networks are far less likely to occur within a single pathway than to span between pathways. This result mirrors what has been observed in analysis of reverse genetic interaction networks, particularly amongst interactions characterized as synthetic lethal or synthetic sick, which have been shown to interconnect different pathways that are functionally synergistic or redundant[19,20].

Chapter 4.4.3: Complementarity between natural and synthetic genetic networks

Next, we asked whether the natural genetic network had any direct overlap with "synthetic" networks derived using reverse genetic approaches such as SGA, dSLAM, or E-MAP platforms. To address this question, we considered four synthetic interaction networks: a work by Tong et al.[7] reporting comprehensive interaction screens for 132 genes using SGA, a genetic network governing DNA integrity identified using dSLAM[11], and E-MAPs centered on chromosomal biology[9] and RNA processing[13]. The combined network from these four sources consisted of 2,117 genes linked by 29,275 genetic interactions. As with the natural network, we confirmed that

interactions in the combined synthetic network were more likely to fall between functional terms and protein complexes than within them (Table 4.1 and Table S4).

To evaluate overlap, an interaction in the synthetic network was considered "supported" if the two genes mapped into two different intervals that were found to interact in the natural network. As shown in Figure 4.3A, the natural network supported on average 8.7% of interactions across the four synthetic networks as opposed to 5.7 ± 0.5% expected by chance (Supplementary Methods). Thus, some regions are shared in common between natural and synthetic networks, but these regions appear to represent a minority of all genetic interactions.

We found that these common genetic interactions took place among genes encoding basal transcriptional activators ("regulation of nucleotide metabolism", Figure 4.3B) including components of RNA polymerase II, Kornberg's mediator complex, the holo TFIIH complex, INO80, SET3, and COMPASS (Figure 4.4A). Moreover, the expression traits linked to these common interactions were for genes encoding the cytosolic ribosome ($P' < 10^{-47}$), cell cycle checkpoints ($P' < 10^{-15}$, including *RAD9* and *DDC1*), and mitochondrial electron transport ($P' < 10^{-12}$). Thus, interactions that overlap between natural and synthetic genetic networks seem to take place among core transcriptional activators and influence expression of core metabolic processes.

Chapter 4.4.4: Novel interactions of the INO80 complex as suggested by natural networks

One prominent complex highlighted by both natural and synthetic interactions was INO80, a multi-subunit ATP-dependent chromatin remodeling complex (Figure 4.4A). At its core is the Ino80 protein, an ATPase of the SNF2 family which functions as the catalytic subunit. Recent studies have demonstrated that INO80 chromatin remodeling activity contributes to a wide variety of pivotal processes, including transcription, DNA replication, and DNA repair[114,129-131]. Consistent with these processes, both the natural and synthetic networks supported interactions of INO80 with TFIIH and alpha(I)-primase. However, INO80 had far more interactions in the natural network than the synthetic one. This result is reflected in Figure 4.4A (large height versus width of the INO80 node) and more explicitly in Figure 4.4B, which plots the p-values in the natural versus synthetic network for all complex pairs involving INO80. This plot shows that the reason for few synthetic interactions is lack of coverage: most complex pairs (82%) have simply not yet been tested for interaction using reverse genetic screens, placing them at a significance score of P = 1 (i.e., on the y-axis of Figure 4.4B).

To fill this gap, we genetically analyzed three genes encoding members of the INO80 complex (Arp8, Ies3, Nhp10) using the quantitative E-MAP approach. Complete genomic deletions of each gene were screened against a standard array of 1,536 mutants to select double mutant combinations whose growth rates were slower or faster than expected (Methods). This screen uncovered 496 novel genetic

interactions (Table S5) supporting 20 complex-complex relationships ($P < 0.05$; Table S6). Nine of the complex-complex interactions were also supported by the natural network, including interactions with four complexes (tRNA splicing, RNA polymerase II, Actin-associated proteins, and the Vps35/Vps29/Vps26 complex) that were already present in the common complex interaction map (see Figure 4.4B and Figure 4.4C).

The relationships identified here implicate a number of novel links between INO80-mediated chromatin remodeling and a wide range of important cellular processes. For example, numerous genetic interactions were identified between INO80 and RNA Polymerase II. There is substantial evidence demonstrating that the rate of transcriptional elongation by RNA Polymerase II is reduced in the presence of nucleosomes and requires chromatin-modifying activities[132]. Since INO80 has been shown to mobilize/remove nucleosomes[130,133], this functional link may indicate that the two complexes co-operate: INO80 may exchange histones at a particular location to facilitate transcriptional elongation by RNA polymerase II. Indeed, while this manuscript was in review, a new report has implicated a role for INO80 in histone redeposition during RNA polymerase II-mediated transcription of stress-induced genes[134].

Four of the nine novel INO80 interactions are involved in various aspects of vacuolar protein degradation including transport of hydrolases to the vacuole (Vps35/Vps29/Vps26 complex and Vps27/Hse1 complex), vacuole biogenesis (Vacuolar assembly complex), and targeting of proteins for degradation (Rubiquitin-activating complex). Given INO80's role in transcription[130], the new interactions

suggest that these complexes work in tandem to regulate the expression level of certain proteins, with INO80 controlling the level of transcription and these four complexes controlling the rate of protein degradation. This work serves as an example of how the broad coverage in the natural network can be used to focus future genetic screens and provide the basis for many mechanistic follow-up studies.

Chapter 4.5: Discussion

Currently, mapping genetic interactions using GWAS faces two major challenges: a lack of statistical power for finding genotype-phenotype associations, and a lack of tools for understanding the molecular mechanisms behind the associations found to be significant[25-27]. In this study, we have demonstrated that such challenges can be partly overcome by (1) accounting for bi-cluster structure in the data and (2) by integrating genetic interactions derived from GWAS with protein complexes and functional annotations. The result is a map of protein complexes and pathways interconnected by dense bundles of genetic interactions, which raises statistical power and provides biological context to the genetic interactions uncovered in natural populations.

Despite exhibiting some overlap (8.7%), there was also much divergence between the natural and synthetic networks. Such divergence might be explained by a number of factors. First, the two types of genetic networks have major differences with respect to coverage and power. Natural networks are based on genome-wide variations and thus nearly all gene pairs are tested for pairwise interaction— i.e., the

coverage of gene pairs is practically complete. This large coverage comes at the price of low statistical power: gene association studies are limited by the number of individuals that can be surveyed which, in turn, limits the power of natural genetics to detect any given genetic interaction. On the other hand, a reverse genetic interaction screen explicitly tests the growth rate of gene pairs, with high power to detect interaction. However, the set of gene pairs that can be tested in a single study is limited by the throughput of the screening technology. The synthetic genetic network used here was a combination of four such studies which collectively cover approximately 5% of yeast gene pairs. Future efforts may seek to complement the coverage of reverse genetic screens by using natural genetics, or to improve the power of gene association studies through focused reverse genetic analysis. Here, we have demonstrated this concept by expanding the coverage of the synthetic network around the INO80 complex, based on the conserved interactions we found for this complex in both types of networks.

Even with equivalent coverage and power, the two types of network would still likely diverge due to their different means of perturbation. The natural network is driven by variations in genome sequence including SNPs, repeat expansions, copy number variations, and chromosomal rearrangements which lead to a variety of effects on gene function such as hypo- and hypermorphic alleles, null alleles, and so on. In contrast, synthetic networks predominantly consist of complete gene deletion events, which are rarely experienced in nature and lead exclusively to null alleles.

A final difference is phenotype— the natural and synthetic networks in this study differ markedly in the underlying phenotypic traits they have measured, relating to gene expression versus cell growth, respectively. It is important to note, however, that the differences in traits are specific to the currently available data sets. They are not inherent to either mapping approach, and in general one can imagine synthetic genetic interactions related to gene expression (see Jonikas *et al*. for a recent example[135]) or natural interactions related to a single phenotypic trait such as cell viability or disease (which in fact describes the majority of GWAS data generated to-date for humans)[23].

Despite all of these differences, we did observe a significant number of natural and synthetic genetic interactions in common. It is tempting to speculate that these common interactions might share certain characteristics with regard to cellular function. In particular, we found that natural interactions also present in the synthetic network were linked to expression levels of ribosomal genes as well as to core components of respiration and cell cycle. Several studies have noted a correlation between the expression levels of ribosomal or mitochondrial genes and growth rate[136,137]. Thus, the overlap between natural and synthetic interactions seems to occur among genes that strongly influence expression traits related to growth.

A common issue in association studies, known as the "fine mapping problem"[30,138], is that a strongly associated marker will fall near many candidate genes, leaving it ambiguous as to which of these candidates is the causal factor. Numerous methods have been developed to refine or prioritize these candidates, often

through incorporation of orthogonal information[139]. An extension of this problem applies to marker-marker interactions, which typically implicate one of many possible pairs of genes. Here, we have mitigated this problem by summarizing markers into protein complexes and functional terms. However, ambiguities can still arise in cases where several complex-complex interactions are supported by the same underlying set of marker pairs. Since it is likely that only one of these interactions is causally linked to phenotype, further work may be necessary to prioritize these candidates. It is important to note, however, that fine-mapping issues will be less of a concern in humans than in yeast, given the higher density of available markers which will improve the resolution in identifying causal genes.

In summary, we have demonstrated that the logical framework developed for analysis of synthetic genetic networks can also be readily applied to natural genetic networks. Biologically and clinically, the clear and immediate application is towards the analysis of genome-wide association studies in humans. Many diseases, both common and rare, have so far been opaque to genome-wide association analysis[140]. The key question will be whether, using integrative maps such as those developed here, they can become less so.

Chapter 4.6: Methods

Chapter 4.6.1: Marker pair bi-clustering

An interval is defined as a set of one or more contiguous markers along the chromosome. A pair of intervals induces a set of m tested marker pairs of which k

pairs are found to interact, drawn from a total genome-wide pool of N tested marker pairs of which n are found to interact. An exhaustive genome-wide scan is performed to identify interacting interval pairs, i.e. those that are enriched for marker-marker interactions, as follows. The counts (*m*, *k*) are tallied for all possible pairs of intervals (up to a maximum of 60 markers per interval) using a recursive algorithm in which the entire space of marker pairs is represented as an upper-triangular matrix A with each row and column denoting a marker. An interval pair is represented by a submatrix Ai,j,a,b, where i,j are the starting row and column indices and a,b are the dimensions of the submatrix. The number ki,j,a,b of interacting marker pairs in a submatrix is determined using the formula:

$$k_{i,j,a,b} = k_{i,j,a-1,b-1} + k_{i+a,j,1,b-1} + k_{i,j+b,a-1,1} + k_{i+a,j+b,1,1}$$

An identical formula is used to count the number of tested marker pairs in each interval pair (substitute *m* for *k*). Following computation of the (*m*, *k*) counts, every interval pair is assigned a p-value of enrichment for marker-marker interactions based on the four parameters *m*, *k*, *N*, *n* using the hypergeometric distribution. The natural network is then assembled in an iterative fashion, where the most significant interval pair is selected from among all possible interval pairs, after which all interval pairs which contain any overlapping marker pairs (interacting or non-interacting) are removed from consideration. The process is repeated until there are no interval pairs remaining, which ensures that the final set of interval-interval interactions comprising the natural network is disjoint.

Chapter 4.6.2: Comparison of bi-clustering to a naïve algorithm

We considered that the improved performance of bi-clustering might be non-specific, i.e., that simpler methods for expanding marker-marker pairs to form genomic intervals might perform equally well. As one possibility, we compared the bi-clustering approach to a naïve algorithm for generating interval-interval interactions, in which raw marker pairs were expanded to encompass the nearest *x* neighboring markers on either side. However, as shown in Supplemental Figure 4.1 this naïve expansion method performed substantially worse than bi-clustering at identifying term-term or complex-complex interactions, for any choice of *x*, suggesting that bi-clustering identifies more appropriate interval boundaries for each natural genetic interaction.

Chapter 4.6.3: Mapping genes to intervals

The chromosomal coordinates of open reading frames (ORFs) for all yeast genes were obtained from the *Saccharomyces* Genome Database[141]. Each gene was assigned to all markers found within its ORF and to the nearest marker within a window of x = 100 kb on either side (Supplemental Figure 4.2). This mapping procedure resulted in a discrete number of genes mapped to a given marker. Intervals were mapped to all genes assigned to their constituent markers, again resulting in a discrete number of genes mapped to an interval.

The complex-complex interactions identified in the natural network were robust to the particular choice of window size *x*. We varied *x* over a range of distance

thresholds from 0 to 100 kb. As shown in Supplemental Figure 4.3, the resulting complex-complex interactions implicated by the natural network had a high degree of overlap with the results obtained using the original mapping procedure.

Chapter 4.6.4: Enrichments of interactions within and between complexes and terms

A within-complex (within-term) model is defined as the set of all gene pairs falling within a given physical complex (functional GO term). A between-complex (between-term) model is defined as the set of all gene pairs that span two complexes (terms), such that one gene belongs to the first complex, the other gene belongs to the second complex, and neither gene belongs to both. For each model we compute *k*, the number of gene pairs "supported" (see main text) by the network. The significance of this support is assessed using the hypergeometric distribution, governed by *k* and three additional parameters:

n. The total number of gene pairs induced by the model.

m. The total number of gene pairs having support in the entire network.

N. The total number of gene pairs in the tested space of the entire network.

Counts for all four parameters are based only on pairs of genes found in the corresponding space of interactions tested by the network and covered by the given annotation set (complexes or terms). Further details are given in Text S1. All models are visualized using Cytoscape[48].

Chapter 4.6.5: Removing the effects of non-random gene order on annotation enrichment

The above enrichment tests assume independence of genetic interactions from protein complexes and functional terms. However, intervals in the natural network typically cover several consecutive genes, which are more likely to be of similar function than genes chosen at random[142]. To correct for this effect, each complex/term annotation is assigned a score $P_{min} \in [0, 1]$ measuring the degree to which its member genes are clustered [$P_{min} \to 0$] versus dispersed [$P_{min} \to 1$] along the genome (see Text S1 for more details). Annotations with $P_{min} < p_T$ are removed from further consideration. We use a stringent threshold of p_T=0.1 for physical complexes and p_T=0.3 for functional terms resulting in less than one erroneous complex-complex or term-term interaction identified in randomized networks (Supplemental Figures 4.4 and 4.5). Further details regarding the randomization procedure is provided in Text S1. A list of the complexes used in this study is provided in Table S8.

Chapter 4.6.6: INO80 Epistatic Mini-Array Profile (E-MAP)

The *arp8*Δ, *nhp10*Δ, and *ies3*Δ knockout strains were constructed and E-MAP experiments were performed as described previously[41]. The array used to generate the double-knockout strains contained 1,536 strains involved in chromatin metabolism (including chromatin remodeling, repair, replication, and transcription) as well as global cellular processes like protein trafficking and mitochondrial metabolism (see Table S5). Genetic interaction scores were computed as described previously[10].

Chapter 4.7: Acknowledgements

We thank Sourav Bandyopadhyay for numerous comments and suggestions. Tune H. Pers, Karen Kapur, and Ryan Kelley provided helpful reviews of the manuscript.

Chapter 4, in full, is a reprint of the following published work: "Hannum, G*, Srivas, R*, Guenole, A., van Attikum, H., Krogan, N.J., Karp, R.M., Ideker, T. *Genome-wide association data reveal a global map of genetic interactions among protein complexes*. PLoS Genetics 5(12):e1000782 (2009)". The dissertation author was the primary investigator and author of this paper. For the sake of brevity, all Supplementary Data and Supplementary Tables have not provided here. These items can be found at the publication's website.

Chapter 4.8: Supplementary Methods

Chapter 4.8.1: Mapping interacting marker pairs using an exhaustive 2D scan

For comparison to the Storey *et al.*[28] approach, we performed a complete 2D scan of the Brem et al.[29] linkage data to identify interacting marker pairs. First, redundant markers were merged in a manner identical to Storey et al. (see Table S1 for a list of all markers and genomic positions). Next, for each gene-expression trait marker-pairs were assigned a baseline F-score for interaction using a two-way analysis of variance with a fixed-effects model[143]. To assess significance, the complete scan was repeated over 100 permutations in which each segregant strain was randomly re-assigned a gene-expression value. The best F-score for each trait in each permutation

was used to construct an empirical null distribution. This distribution was subsequently used to assign a p-value to each marker pair. For comparison with the Storey *et al.*[28] marker-pair list, marker pairs across all traits were pooled together and thresholded at $P < 0.18$ which produced an identical number of marker pairs (4,687; Table S7A). This network was bi-clustered (Table S7B), and examined for annotation enrichment (Table 4.1 in the main text). It performed substantially worse than the network derived from Storey et al.

Chapter 4.8.2: Annotation datasets

The following annotation datasets were used in this study:

1. GO terms: We obtained gene functional annotations from the Gene Ontology (GO) Database revision 5.814 (July, 2008) [52].

2. Physical complexes: A set of yeast protein complexes was obtained from MIPS [122] and from Gavin *et al.* [5] ("Core" Set). The union of these sets was filtered to ensure no two complexes shared a Jaccard score (intersection / union) greater than 0.1. When two complexes exceeded this threshold, priority was given to MIPS literature-curated complexes followed by complexes with greater numbers of proteins. A list of all complexes used can be found in Table S8.

Chapter 4.8.3: Defining the marker pair test spaces

Two sets of interacting marker pairs were considered in this study: (1) marker pairs from the Storey *et al.* dataset and (2) marker pairs identified through an exhaustive 2D scan. For the Storey *et al.* dataset, a marker pair was defined as tested if it had been examined for joint linkage with at least one trait. This produced 691,039 tested marker pairs. For the exhaustive 2D scan, all marker pairs were tested except those with markers that were highly correlated in the segregant population. Correlated marker pairs violate the assumption of balance in two-way ANOVA, the method used in the exhaustive 2D scan as described above. Two markers were considered highly correlated if the number of segregants with the most common pair-wise genotype was more than twice that of the least common pair-wise genotype. Using this criterion, the exhaustive scan tested a total of 623,073 marker pairs (representing approximately 85% of all pairs).

Chapter 4.8.4: Defining the gene pair test spaces

To determine the significance of enrichment of genetic interactions within or between annotations, we determined four parameters for the hypergeometric distribution k, m, n, and N as described in the main Methods. To compute these values, it was first necessary to determine the set of gene pairs tested by each genetic network. This space was computed differently depending on the network type, consisting of either (1) raw marker-marker interactions, (2) interval-interval

interactions, or (3) synthetic gene-gene interactions. A test space (4) was also constructed for the physical complexes and functional terms.

A gene pair (g1, g2) was considered "tested" iff:

Case (1): There exists a *tested* marker pair (m1, m2) such that m1 → g1 and m2 → g2 (the arrows → denote mapping of markers to genes as described in the main Methods; the definition of tested marker pairs is given in the section above).

Case (2): There exists *any* marker pair (tested or untested) such that m1 → g1 and m2 → g2. The rationale is that an interval covers a contiguous range of markers, regardless of whether any individual marker was explicitly tested for interaction.

Case (3): Its corresponding double mutant had been created and examined as part of a synthetic screen, regardless of the growth rate of the mutant. This information is not reported in every genetic interaction study but was available for the four included in our paper.

Case (4): The physical complex and functional term sets were each assigned a test space consisting of all pair-wise interactions between genes annotated in each set.

All four parameters of the hypergeometric distribution were considered only within the subset of *N*, the intersection of the test space of the analyzed network and the test space of the complexes/terms.

Chapter 4.8.5: Defining a colocalization score and determining a suitable threshold

For each annotation A, a co-localization score P_{min} was computed as follows. Define:

G = $\{g_1, g_2, \ldots, g_{|G|}\}$, the set of genes in annotation A. Define a partition $G_c \subseteq$ G which contains all g in G that fall on chromosome c.

x = $(x_1, x_2, \ldots, x_{|G|})$, the genomic position of each g in G on its chromosome, measured in bp from the left chromosome end to the middle of the ORF encoding g (as given in the SGD database [141]).

m = $(m_1, m_2, \ldots, m_{16})$, the co-localization score to be determined for each chromosome.

For each chromosome c = (1,2, …16):

For all (g_i, g_j) in G_c, compute the intergene distance $d_{ij} = |x_i - x_j|$.

Define $\mathbf{d_c} = (d_1^*, d_2^*, \ldots)$ as the sorted list of these intergenic distances.

Compare each d_i^* to a corresponding null distribution of distances $\mathbf{d_i^0}$. $\mathbf{d_i^0}$ is the distribution of gene-gene distances at rank i in the sorted list produced by sampling without replacement the same number of genes $|G_c|$ at random from the chromosome 10^6 times. Define **p** = $(p_1, p_2, \ldots)$ where p_i is the p-value of d_i^* indexed against $\mathbf{d_i^0}$.

m_c = min(**p**).

P_{min} = min(**m**)

Given this metric, P_{min}, co-clustered annotations were filtered by removing those annotations with $P_{min} < p_T$ from further consideration. We chose a suitable threshold p_T so as to ensure that no complex-complex or term-term associations would be reported if the network were permuted. Permuted interval networks were generated by randomly assigning a new starting marker-index to each interval in the natural network, while ensuring that interval pairs remain disjoint and that no interval crosses the edge of a chromosome. We computed the number of significant associations found by 100 permuted networks over a range of p_T values (Figure S4). Based on this analysis, we chose a stringent colocalization threshold of $P_{min}>0.1$ for physical complexes and $P_{min}>0.3$ for functional terms (blue arrows in Supplemental Figure 4.4) resulting in less than one erroneous complex-complex or term-term interaction identified per permuted network.

To further validate the annotation models, we performed two additional permutation methods and examined how many significant complex-complex interactions could be identified in either case. First, we re-assigned interactions between intervals in the natural network. Second, we performed 100 randomized scans for marker-marker interactions using the method of Storey *et al.*[28], in which the associations between segregant strains and gene expression traits were randomly permuted while leaving the associations between segregrant strains and genotypes intact. Each of these random marker-marker networks was subsequently bi-clustered to produce interval-interval networks. Across, 100 permuted interval networks derived

through these two methods very few complex-complex interactions were identified (<2 on average; Supplemental Figure 4.5).

Physical complexes and terms with a score above these thresholds were removed prior to enrichment analysis. The filtering process removed approximately 16% of the physical complexes and 40% of the functional terms, resulting in a reduced set of 302 physical complexes (Table S8) and 1,954 functional terms after further processing as described above.

Chapter 4.8.5: Determining significance of overlap between natural and synthetic networks

The overlap between the natural and synthetic networks was based on the number of synthetic genetic interactions that were supported by the natural network. A genetic interaction was considered supported if the two genes mapped into two different genomic intervals that were found to interact. Significance was determined using 1,000 natural network permutations using a procedure based on the re-positioning of interval-interval interactions. In this scheme, each interval of an interval-interval interaction in the natural network was randomly assigned a new starting marker-index, while ensuring that interval pairs remain disjoint and that no interval crosses the edge of a chromosome. This effectively disrupts any biological signal, while preserving the distribution of interval sizes.

Chapter 4.8.6: Mapping broad GO terms

As shown in Figure 4.3B of the main text, we characterized the functional relationships for the natural and synthetic networks by mapping all identified functional term and term-term interactions to a set of “broad” terms defined at the fifth and sixth levels of the GO hierarchy (1,285 possible terms). For each of these broad terms, the number of term and term-term interactions among the mapped children was tabulated. Similarly, for each pair of broad terms, the number of term-term interactions between the respective children was tabulated. The 10 broad terms and 30 term-term interactions with the most counts were considered a good representation of the functional relationships evident in the natural and synthetic networks.

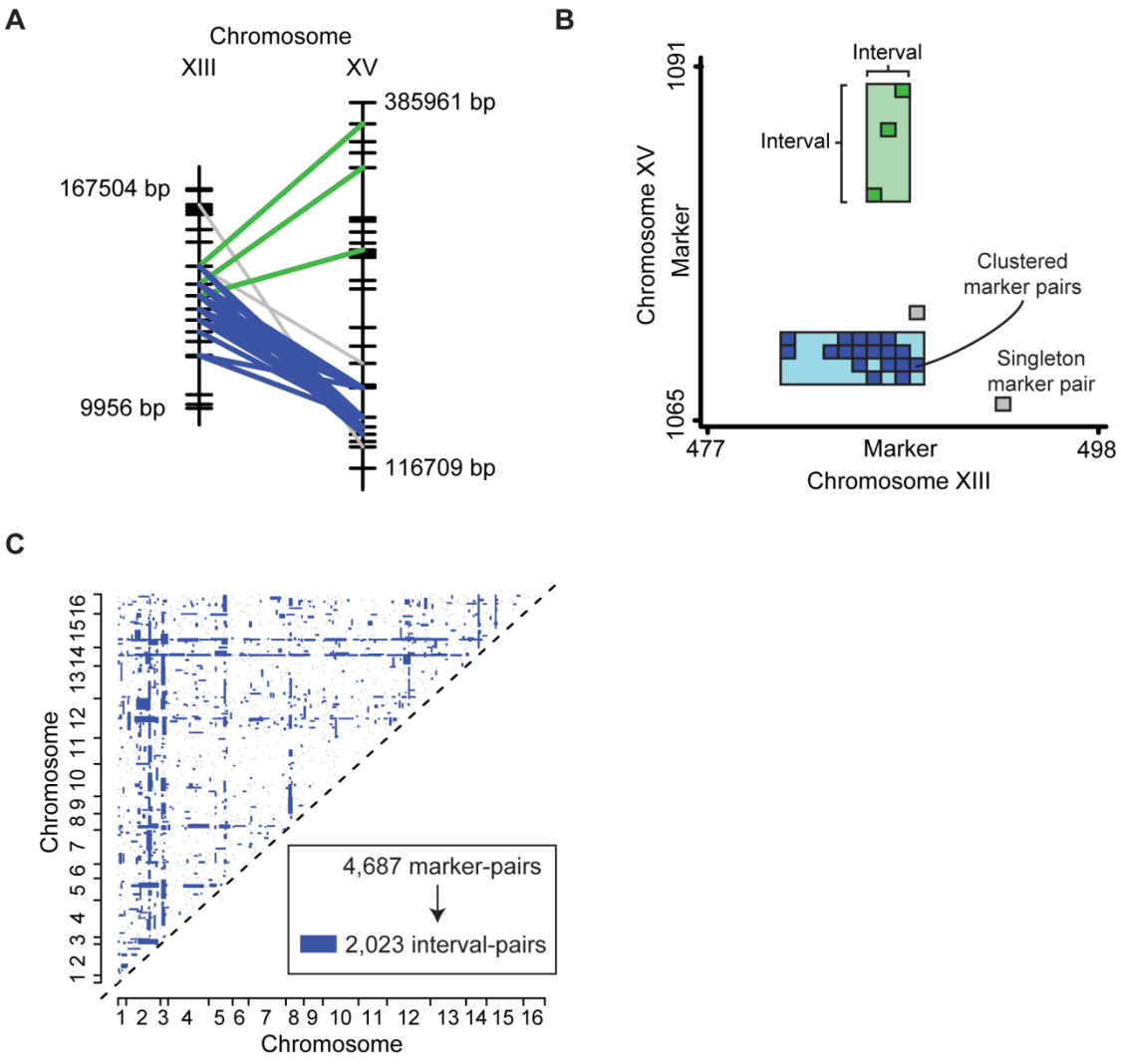


Figure 4.1: Using genome-wide linkage data to identify natural genetic interaction.

(A) Two interacting interval pairs (green and blue) which represent significantly dense groups of marker-marker interactions are shown. (B) A matrix view of the same genomic regions. The blue and green interval pairs appear as two rectangles. (C) The entire set of marker pairs was bi-clustered to form a set of high-confidence interval pairs (blue rectangles).

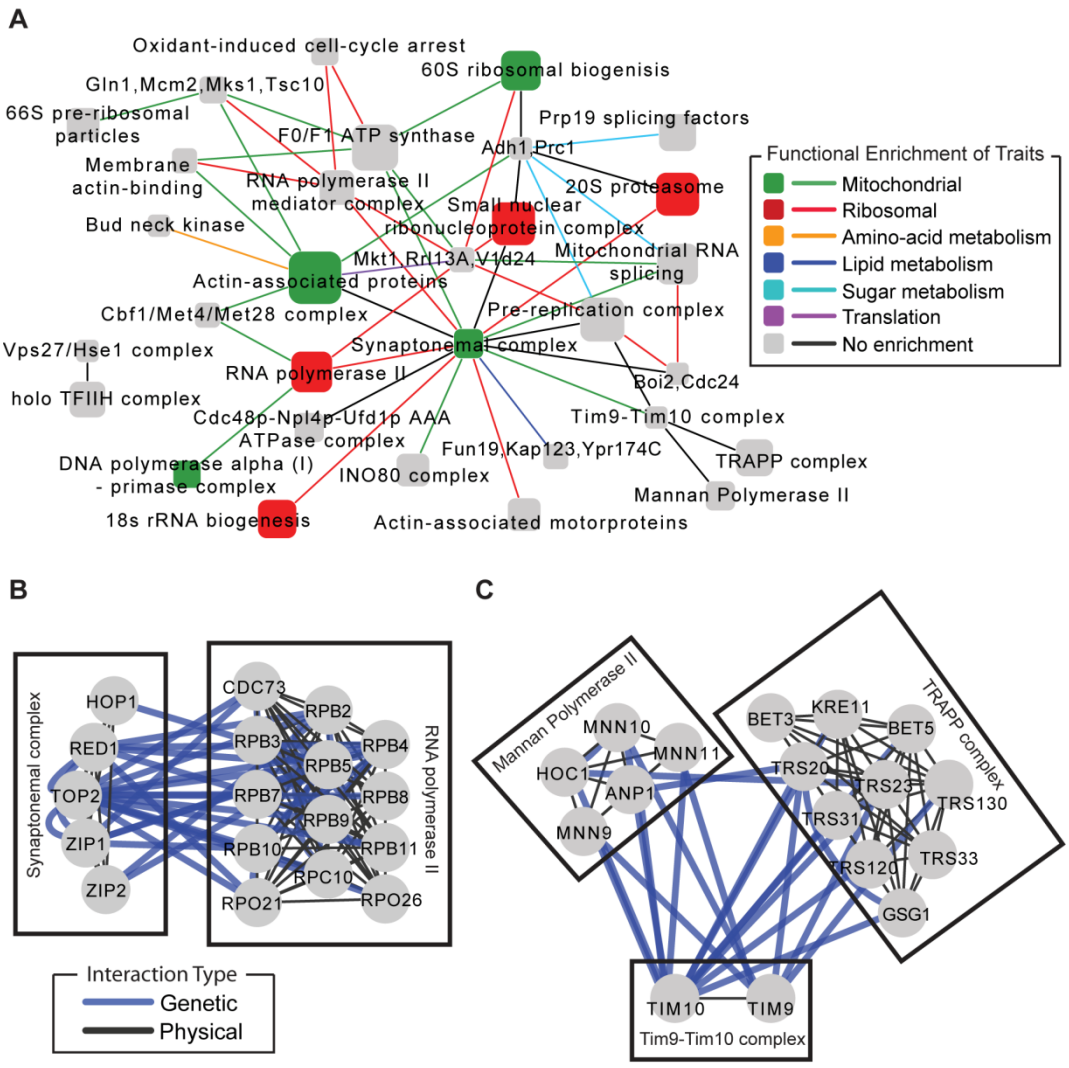


Figure 4.2: Natural genetic networks elucidate pathway architecture

(A) A global map of the complex-complex interactions found using the natural network. Each node represents a protein complex and each interaction represents a significant number of genetic interactions (False Discovery Rate < 5%)[144]. We analyzed the set of gene expression traits associated with each complex-complex interaction for functional enrichment using the hypergeometric test. Nodes and edges are colored according to the functional enrichment of gene expression traits underlying the natural interactions (Bonferroni $P' < 0.05$). Node sizes are proportional to the number of proteins in the complex. When known, nodes have been labeled with the common name of the complex. (B – C) Two specific examples of complexes spanned by dense bundles of natural genetic interactions.

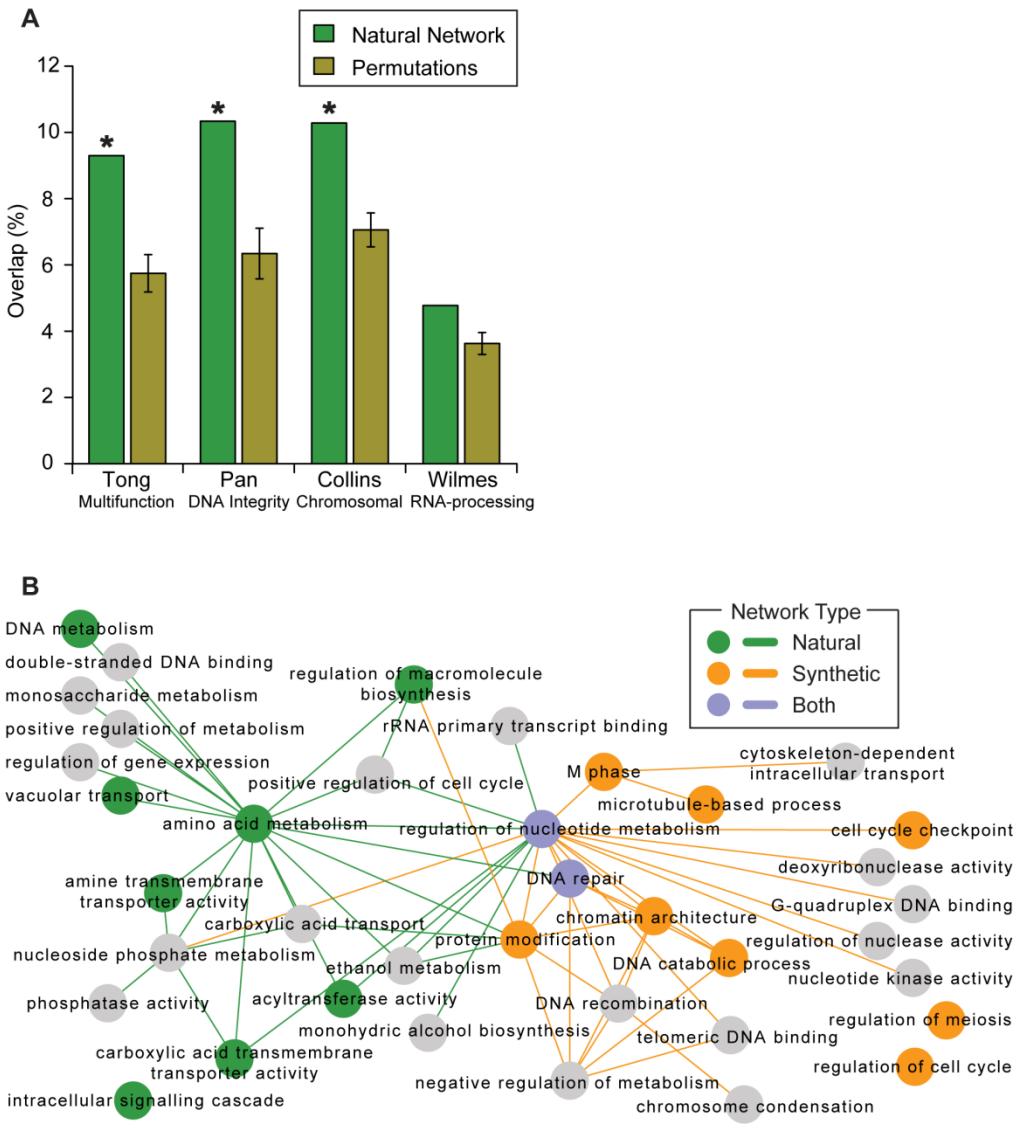


Figure 4.3: Comparison of the natural and synthetic networks

(A) The overlap between the natural network and four previously-published synthetic genetic networks (Tong[7], Pan[11], Collins[9], Wilmes[13]) is shown as a percentage of the synthetic network size. An asterisk indicates significance at $P < 0.05$. (B) A map of the functions and functional relationships supported by either the natural or synthetic networks. Each node represents a broad GO term, with colors (green, orange, blue) indicating terms that contain many within-term interactions. Edges show the top 30 between-term interactions for each of the natural and synthetic networks. Two broad GO terms (regulation of nucleotide metabolism and DNA repair) contained many within-term interactions in both the natural and synthetic networks.

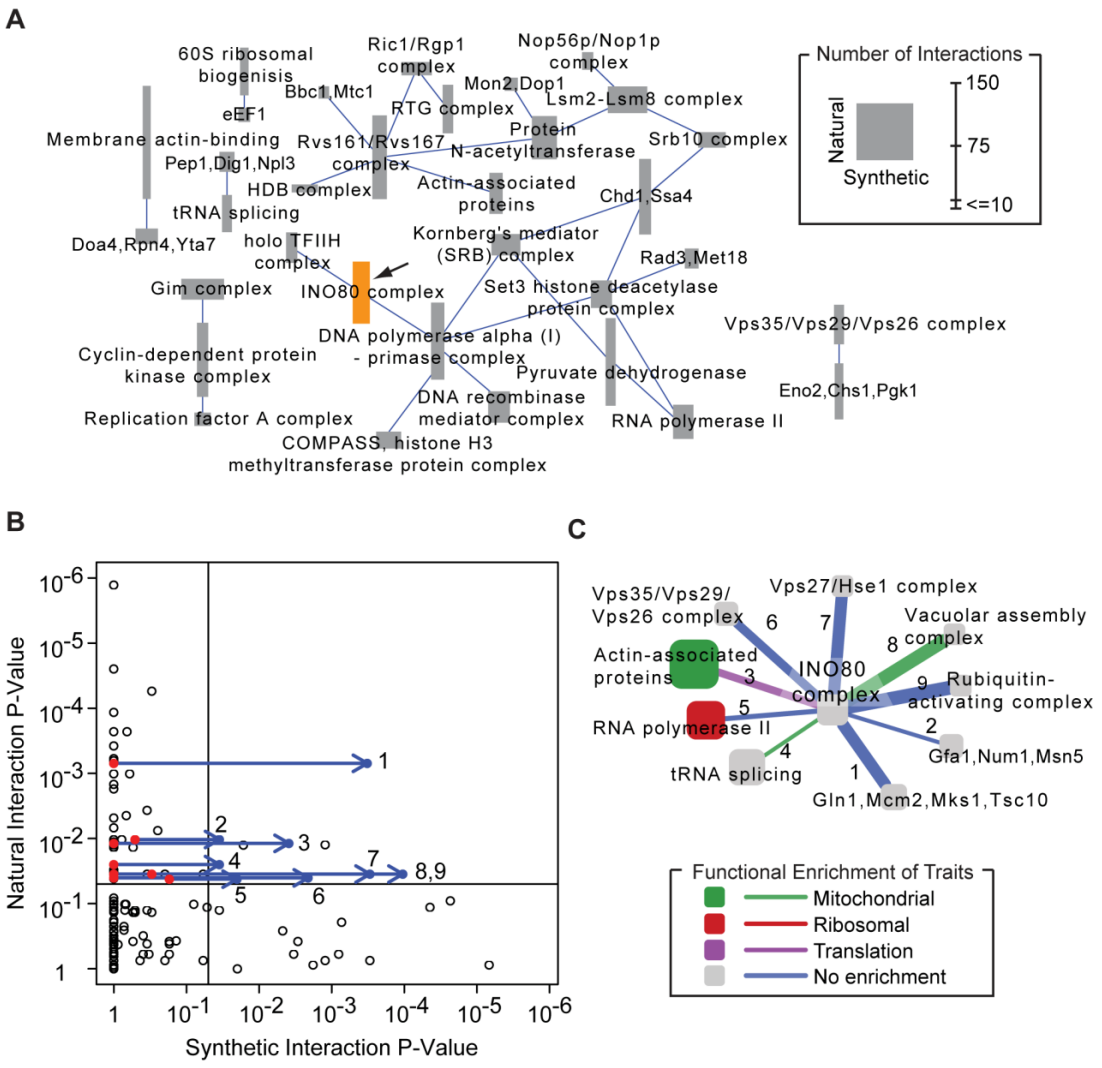


Figure 4.4: Guiding synthetic genetic screens using natural genetic networks.

(A) Complex-complex interactions common to both the natural and synthetic networks at a relaxed threshold of $P < 0.05$. Many of these complexes, including INO80 (orange), have more coverage in the natural network (node height) than in the synthetic network (node width). (B) Each point in the scatter plot represents the significance of support for a possible complex-complex interaction with INO80 from the natural (x-axis) versus synthetic (y-axis) networks. Due to low coverage, comparatively few complex pairs have support in the synthetic network. New E-MAP data for INO80 support nine new complex-complex interactions predicted by the natural network (blue arrows). (C) A network of natural genetic interactions for INO80 validated by the new E-MAP. Functional enrichment for traits is shown as in Figure 4.2. The thickness of each link is proportional to its support in the new genetic interaction screen.

Table 4.1: Correspondence of interval and marker pairs with complexes and functions

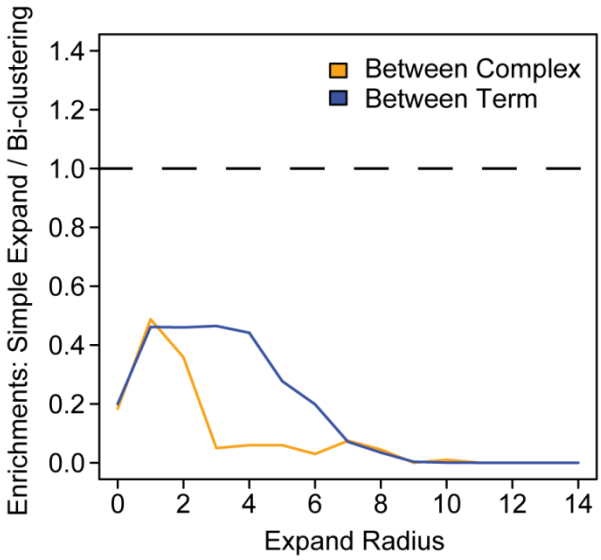
			Between		Within	
	Nodes†	Edges‡	Complexes	Terms	Complexes	Terms
Storey *et al.*						
Bi-clustering*	1,977	2,023	208	17,714	0	12
Raw Marker Pairs	1,157	4,687	38	3,546	0	3
Full 2D ANOVA scan**						
Bi-clustering	1,387	964	0	19	0	0
Raw Marker Pairs	1,141	4,687	0	0	0	0
Synthetic Genetic Analysis	2,117	29,275	140	1,833	13	33

† Node definition: For Storey *et al.* and Full 2D ANOVA, nodes represent genomic intervals. For the synthetic network, nodes represent genes.

‡ All cases report the number of distinct interactions in the network, removing redundancies due to marker pairs that associate with multiple traits (Storey *et al.*, Full 2D ANOVA) or gene pairs scoring positive in multiple data sets (Synthetic Genetic Analysis).

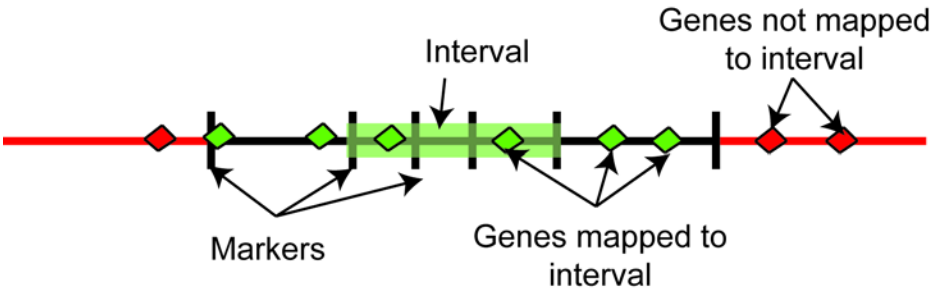
* These bi-clustered interval pairs were used to define the "Natural Network" explored in this work.

** We also considered an exhaustive scan of all marker pairs using two-way analysis of variance (ANOVA). The most significant 4,687 marker-marker interactions (Table S7) were taken to match the number of interactions from Storey *et al.* (Text S1). Both the raw marker-pairs and the bi-clustered interval network identified substantially fewer enrichments than the Storey *et al.* method.



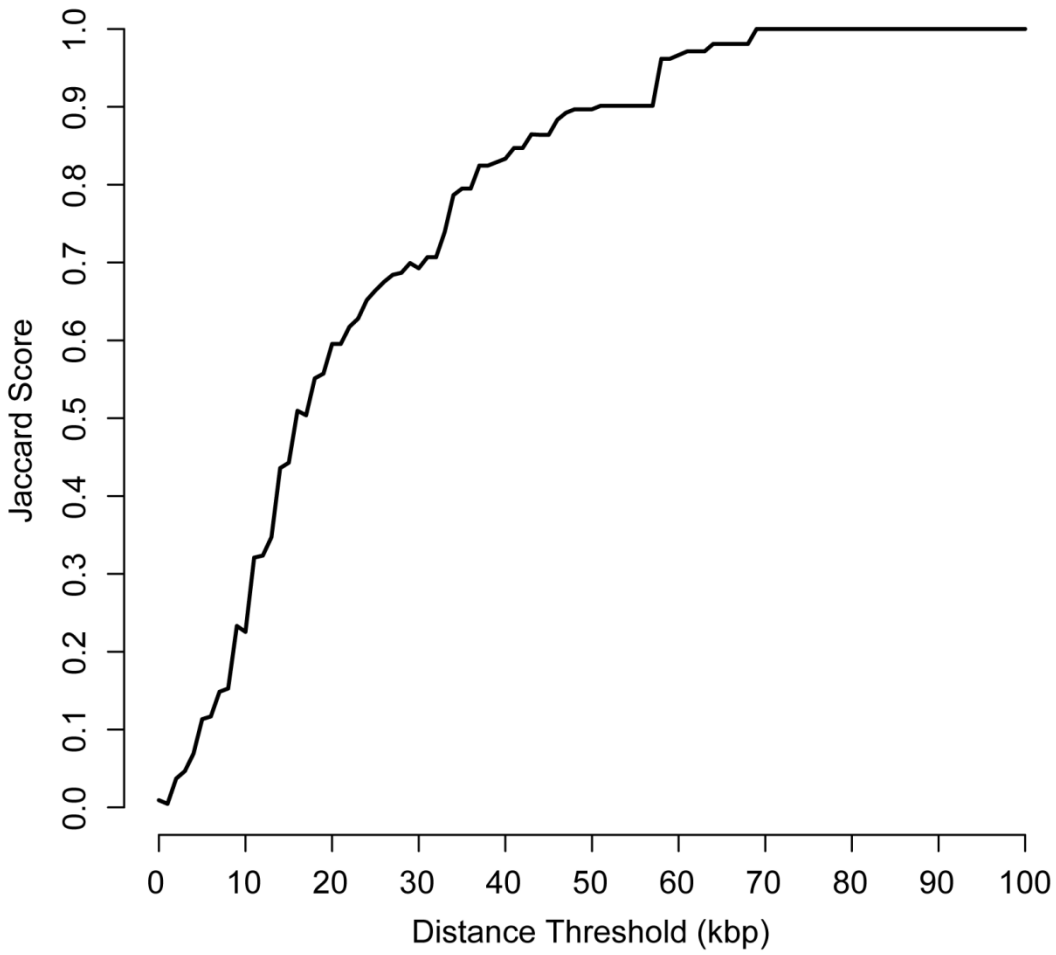
Supplemental Figure 4.1: Comparison of the bi-clustering method to a naïve approach

A naïve approach for identifying interval-interval interactions was compared to the bi-clustering approach. In the naïve approach, markers involved in a marker-marker interaction were expanded to encompass the nearest k neighboring markers on either side. The naïve approach identified substantially fewer between-pathway enrichments.



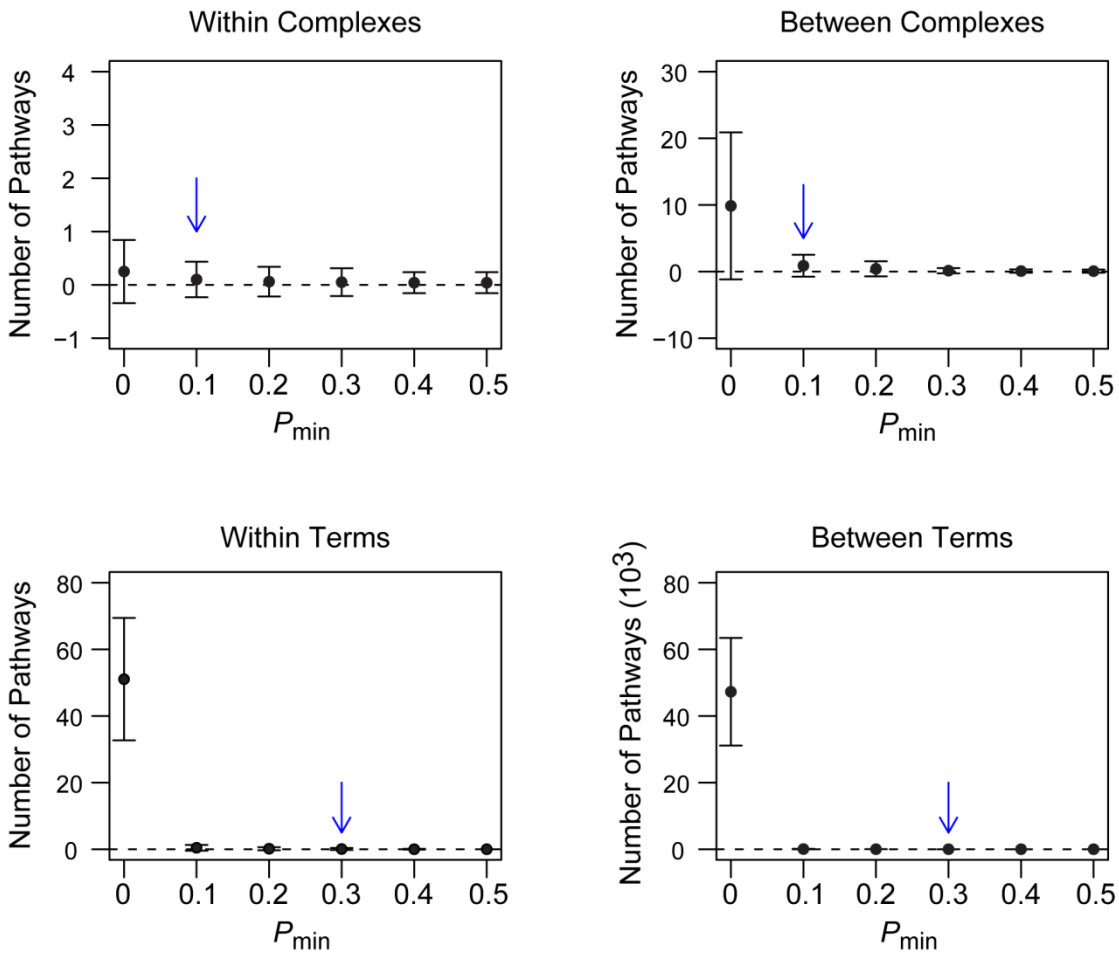
Supplemental Figure 4.2: Interval to gene mapping

Each gene (diamond) was assigned to all markers (vertical bars) found within its ORF and to the nearest marker within a window of x = 100 kb on either side. Each interval (green bar) inherited the mapping of all constituent markers.



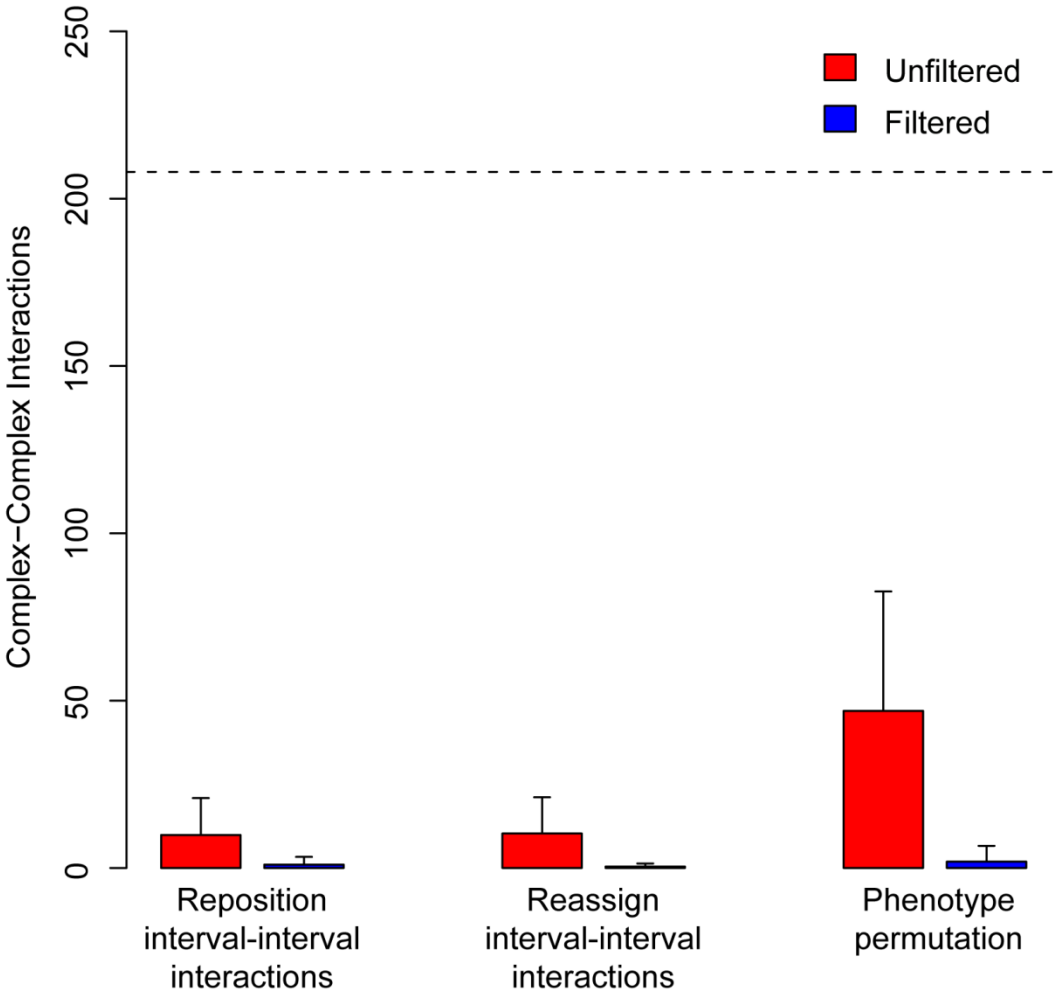
Supplemental Figure 4.3: Sensitivity of pathway identification to marker-gene mapping threshold

Genes were mapped to their nearest marker within 100 kbp. We varied this threshold from 0 kbp to 100 kbp to determine what effect it would have on the resulting complex-complex interactions. Overlap of the resulting complex-complex interactions with the results in the manuscript are shown as a Jaccard score.



Supplemental Figure 4.4: Choosing a colocalization threshold

The number of interactions identified from permuted natural networks were examined at several colocalization thresholds. Thresholds were chosen which resulted in fewer than one interaction in a typical permuted network (blue arrows).



Supplemental Figure 4.5: Additional permutation methods for pathway validation

The number of complex-complex interactions identified by the natural network (dotted line) is compared to the average number of complex-complex interactions identified across 100 permuted interval networks generated using three different procedures. Complex-complex interactions were mapped using either all complexes (unfiltered) or only those with a co-localization p-value above 0.1 (filtered). Error bars indicate one standard deviation.

Chapter 5. Allele Specific Compatibility of Locus Interactions Underyling Yeast DNA Repair Phenotypes

Chapter 5.1: Abstract

Recent studies revealed that many phenotypes, including common diseases, are complex and are likely governed by multiple loci and interactions between these loci. The need to preserve the functionality of these interactions may lead to co-evolution of interacting components. In genetic crosses such interactions can be perturbed, leading to a phenotype change. Such perturbation can, for example, explain the emergence a drug sensitive progenies in genetic cross of two drug resistant strains. Here, we develop a novel computational method, called LoCAp (Locus Compatibility Approach) which is based on proposed interaction models explaining such phenotype change. LoCAp searches for interacting loci consistent with the model using a graph based method. Application of LoCAp to a recent genome-wide linkage study in yeast, identified 12 significant allele specific locus interactions underlying the sensitivity to a panel of genotoxic agents. Moreover, we identified the DNA helicase Rad5 as an interaction hub, which we verified experimentally. These findings serve as a proof of principle for exploring allele specific compatibility in identifying interactions between genetic loci.

Chapter 5.2: Background

It is generally accepted that complex traits are subject to polygenic inheritance, and are controlled by many, potentially interacting, loci with small individual effects[35,145]. Therefore, inferring interactions between various loci remains one of the most important challenges in mapping the genetic basis of complex traits.

Genetic interactions can be detected by observing phenotype changes in response to genetic perturbations. Large-scale knockouts experiments have been used to uncover genetic interactions on a genome-wide scale[146,147] or in the context of particular biological pathways, such as DNA repair[16] or RNA processing[13]. In contrast, natural genetic variability can often be used for uncovering more subtle dependencies[28,29].

The basic approach to identify the relation between variability in a genomic locus and a phenotype, Quantitative Trait Loci (QTL) analysis, looks for correlations between changes in genotype and quantitative changes in a phenotype. Such basic QTL analysis can be extended to allow identification of interacting loci, that is, loci which jointly control a complex trait in a nonadditive (non-independent) way. Empowered by large scale genotyping techniques, QTL analysis has led to uncovering numerous cases of interacting loci in several model organisms including *S. cerevisiae*[148] and *C. elegans*[149]. In humans, genetic interactions have been linked to diseases such as diabetes[150], autism[151], and bipolar disorder[152].

There are several obstacles that render computational detection of multiple locus interactions to be very challenging. First, the search space increases

exponentially as the number of considered loci increases leading to a computationally intractable problem. Second, the huge number of possible locus interactions exasperates the multiple testing problem, making it difficult to uncover statistically significant associations[27].

To mitigate challenges of exhaustive search methods[153,154], a number of methods have been developed to detect significant pairwise interactions by reducing the number of interactions which must be tested[26,28,120,155-163]. For example, step-wise methods[26,28,120] first detect a set of primary loci that have relatively significant effect on the phenotype individually and then search for a secondary interacting loci amongst this primary set. As another example, the SEE (Symmetric Epistasis Estimation) method, developed to detect eQTL epistasis, considers pairs of loci simultaneously, but reduces the number of tests by focusing exclusively on specific patterns of genotypes and gene expression that are expected to be enriched in such interactions[164]. For interested readers, comprehensive reviews on various methods designed to detect multiple locus interactions can be found in[165-168].

Genetic crosses provide an informative context that can be used for capturing interacting loci. In this work, we considered the scenario where both parental strains share a specific phenotype, for example they are both drug resistant, but this phenotype is lost in a fraction of their progeny. In such a case, the emergence of the alternative phenotype in the progeny strains cannot be explained by a perturbation in one locus only, but rather is likely to involve perturbation of interaction between at

least two loci. Building on the between and within pathway models for genetic interactions[8,94,97], one can propose two basic models for loci interaction.

The between pathway model explains a genetic interaction via the existence of two parallel pathways converging on the phenotype of interest, where each of the two interacting genes belongs to one of these two pathways. As long as one of the two pathways is "active" the phenotype is preserved. In contrast, the within pathway model describes a situation where both genes are in the same pathway, but perturbing of only one of them is not disruptive enough to change the phenotype. These two concepts can be translated to genetic interactions in the context of genetic crosses. Namely, assume that two loci, *l* and *l'* function in two parallel biological pathways, each controlling the phenotype and as long as one of these pathways is "active" the cell is resistant (Figure 5.1A). In such a context loss of the original phenotype, in our case drug resistance, can be explained as follows. Suppose that the drug resistance phenotype is mediated by the pathway involving locus *l'* in the first parental strain and by the pathways involving locus *l* in the second parental strain. Then the progenies which inherit locus *l* form the first parent and locus *l'* from the second parent lose drug resistance (Figure 5.1A). Note that we refer to this as a "asymmetric" loss of complementarity since allele 1 in locus *l* and allele 0 in locus *l'* jointly cause different phenotype compared to the case of allele 0 in locus *l* and allele 1 in locus *l'*.

The alternative model parallels the within pathway model and builds on the assumption that loci from the same pathway are likely to co-evolve to maintain the interaction properties as interacting proteins do[169-173]. Such co-evolution of interacting

loci might lead to allele specific locus interaction compatibility as illustrated in Figure 1b. Namely, assume that interacting loci, *l* and *l'*, have been a subject of compensatory mutations in one strain, say strain 0. However, in a genetic cross where loci *l* and *l'* are inherited form different parent the interaction might be perturbed or lost. We have two possibilities: (i) the symmetric case: the interaction is perturbed causing phenotype change whenever loci *l* and *l'* are inherited form different parents or (ii) asymmetric case – only one combination leads to phenotype loss (Figure 1b). In fact a recent study by Heck et al. identified such an interaction between two DNA repair genes, MLH1 and PMS1 from a genetic cross of two strains of yeast[174]. The resulting progeny which had inherited MLH1 and PMS1 from different parents were observed to display a severe DNA repair defect not seen in either parent.

Such results indicate that data generated from genetic crosses can serve as fruitful source in understanding how interacting amongst different genetic loci can contribute to complex traits. Here we develop a computational framework, termed LoCAp (Locus Compatibility Approach), which models loss of the parental phenotype as a loss of compatibility between interacting loci, where the interactions are modeled by the within pathway or between pathway loci interaction models described above. Finding interacting loci where potential loss of compatibility is correlated with loss of phenotype is reduced to identification of specific genotype-phenotype patterns. For computational efficiency and to reduce statistical challenges, LoCAl encodes progeny and phenotype data in a graph and identifies corresponding patterns by an efficient graph mining strategy.

We applied LoCAp to a QTL study conducted in *S. cerevisiae*[175], which mapped the genetic basis of sensitivity to four unique DNA damaging agents, and found 12 significant locus interactions. These results not only demonstrate that the proposed allele specific compatibility approach provides considerable power in detecting interacting loci, but also highlights the importance of allele specific interaction in controlling complex phenotypes.

Chapter 5.3: Results and Discussion

Chapter 5.3.1: LoCAp: A novel method for identifying allele specific interactions

LoCAp is applicable in the case where parental strains have the same phenotype and this phenotype is loss in a fraction of progeny. It models the loss of the phenotype by loss of compatibility between two interacting loci as a result of genetic crossing (Figure 5.1). The model accounts for both between and within pathways interaction models. However, we note that the loss of compatibility for between pathway model and the asymmetric version of the within-pathway model have the same genotype-phenotype patterns. Specifically, if both loci are inherited from the same strain then the phenotype is the same as in the parental phenotype, while the phenotype of the 01 hybrid (a hybrid where first loci is inherited from strain 0 and the second loci is inherited from strain 1) is opposite to the phenotype in the 10 hybrid. This asymmetric case, termed by Litvin *et al.* “allele specific” is in fact more prevalent[120]. Therefore LoCAp focuses on uncovering pairs of loci consistent with this particular genotype phenotype pattern.

A brute force approach to identify putatively interacting locus pairs based on the genotype -phenotype pattern consisted with the proposed model would look at all possible pairs of loci and test if the inherence pattern is consistent with the above scenario. However, such brute force approach would have a limited statistical power and be computationally inefficient. To bypass these problems we used a graph theoretical approach to efficiently filter promising pairs. Simply put, we represented genotype and phenotype relation as a graph and applied an efficient graph searching algorithm to select locus pairs which, for a large enough set progenies, show phenotype and genotype combinations consistent with the LoCAp model. To assess significance, we randomly permuted data repeatedly and then applied the same method to search for locus pairs in random data.

We applied LoCAp to a QTL study conducted in *S. cerevisiae*, which analyzed 123 progenies resulting from a cross of a laboratory strain BY with a wild type isolate RM^{175}. For each progeny the genotype at 2956 genomic loci were measured along with their growth rate under four different DNA damaging agents methyl methane sulfonate (MMS), 4-nitroquinoline 1-oxide (4-NQO), bleomycin, and caffeine. Two parental strains have the same phenotype only in the case of 4-NQO, bleomycin and caffeine. Hence we only consider these three drugs. We identified four, three, and five interactions underlying sensitivity to 4-NQO, bleomycin, and caffeine, respectively. Below, we provide a detailed analysis for each of these predictions.

Separately, we considered symmetric model for compatibility loss where both 01 and 10 hybrids are associated with phenotype loss. The fact that we did not detect

any interaction based on loss of symmetric compatibility could indicate that asymmetric incompatibility is more prevalent – an observation that parallels a similar result in eQTL studies[120].

Chapter 5.3.2: RAD5 identified as a genetic hub underlying 4-NQO sensitivity

In the response to 4-NQO treatment reported by Demogines et al.[175], 31 progenies show sensitive phenotype and 53 progenies show resistant phenotype (see Methods for the definition of the phenotype). Applying LoCAp to this data (see Methods), we found four candidate locus pairs, each of which contained a locus that was in close proximity to the Rad5 open reading frame (Figure 5.2).

For each of the identified four loci interacting with Rad5 locus, we investigated nearby genes as possible causal genes (Table 5.1). Interestingly, using BioGRID database[176], we were able to confirm that gene Rdh54(YBR073W) and Pol31(YJR006W), which are close to two of the four loci respectively, genetically interact with Rad5. The probability of finding such overlapping by chance is about 0.005. In particular, evidence suggested that Pol31 and Rad5 could be in two parallel pathways. The work by Motlagh *et al.* indicated that Pol31 and Mgs1 could function in the same pathway for modulating replication fork movement[177]. This pathway is essential in cells with defective Rad6-dependent DNA damage tolerance pathway[178], where Rad5 resides on an error-free branch.

Next, we investigated if any of the above four loci interacting with Rad5 locus has small main effect individually since multiple locus interactions often occur

between loci that individually have some small main effects[26]. However, due to the limited number of progeny genotyped in this dataset, the identification of small main effects from each of these four loci proved difficult. Recently, Ehrenreich et al. developed a new experimental technique termed Extreme QTL (X QTL) technique, where they used ~107 unique BYxRM MAT**a** haploid segregants, giving the study a much higher power to detect loci with small effects[179]. They identified 14 loci, including Rad5 locus, as associated with sensitivity of yeast to 4-NQO. In their model, Rad5 had strong effect on phenotype and 13 other loci only had small effects, explaining less than 10% of phenotype variation individually[179]. We compared the four loci in our result to the 13 small effect loci identified by X-QTL. One of our loci (4023_s_at_x02) on chromosome 9 is in significant linkage disequilibrium (LD) with its closest X-QTL ($p < 0.05$; see Methods). This indicates that this locus is also likely to have a small main effect on the phenotype.

As another line of support, we performed double knock-out experiments under the presence of MMS (an alkylating agent which induces damage similar to 4-NQO) to test if Rad5 might act as a genetic hub. Our screen identified 332 genetic interactions containing Rad5 out of a total of 1252 tested interactions (Supplementary File 1). Recent large-scale genetic interaction screens have shown that on average less than 2% of tested interactions show a significant genetic interaction[146]. As Rad5 exhibits a much higher rate of genetic interactions (26.2%) it provides further evidence that the Rad5 locus may serve as a hub in the response to 4-NQO.

Chapter 5.3.3: Prediction and literature based support for locus pairs related to sensitivity to bleomycin

For bleomycin, 36 segregants displayed a sensitive phenotype while 43 segregants showed a resistant phenotype. We detected three significant locus pairs ($p < 0.005$). Interestingly, all three pairs included the marker located on chromosome 4 (around position 520kb).

We identified one gene Rad28 (YDR030C) close to this locus, and two genes Rfc4 (YOL094C) and Msh2 (YOL090W) close to locus 8666_at_x04 as potential underlying genes which could affect the phenotype. Their functional annotation suggests their possible involvement in DNA repair process. Rad28 is involved in transcription-coupled repair nucleotide excision repair[180]. Kim *et al.* showed that Rfc4 was required in both DNA replication and DNA repair checkpoints[181]. In addition, Rfc4 is known to function in the RFC-RAD24 complex loaded at DNA repair sites[182]. Msh2, along with Msh3 and Msh6, has important function in mismatch repair system[183].

For two loci in other two candidate locus pairs (7026_at_x11, 6453_at_x15) and (7222_at_x11, 6453_at_x15), both are close to a gene with function in DNA repair. Cdc28 is close to 7026_at_x11 and Mec1 is close to 7222_at_x11. Though Cdc28 is not required to initiate DNA damage checkpoint, it was showed that it helps cell viability during DNA damage[184]. Mec1p is a well-known essential DNA damage checkpoint protein that transduces signals in response to DNA damage.

Chapter 5.3.4: Prediction and literature based support for locus pairs related to sensitivity to caffeine

As a purine analog, caffeine affects many cellular processes and the specific mechanism by which it acts is still largely unclear[185]. In *S. cerevisiae* a recent study showed evidence that caffeine inhibited the target of rapamycin complex 1 (TORC1)[186]. In *S. pombe* it is known to influence DNA damage checkpoints[187].

Using LoCAp, we obtained five candidate locus pairs with $p < 0.01$. We noticed that some loci appear more than once in our result. For example, locus 4075_at_x02 appears twice in five locus pairs. We identified that Irc24 could be a putatively causal gene close to 4075_at_x02. The biological function of Irc24 is unknown. However, its deletion is known to cause increased level of Rad52 foci[188], indicating it could potentially affect homologous recombination for repairing DNA double strand break. Furthermore, the protein level of Irc24p increased at least three fold without change in the transcript level in response to DNA damaging chemical MMS[189]. Gene Tor2 is close to both 10428_at_x12 and 10803_at_x05, each of which appears in one locus pair respectively. Tor2 and Tor1 regulate cell growth in response to nutrient availability and cellular stresses.

We note that some loci in the above five locus pairs are close to loci related to the resistance to 4-NQO or bleomycin. Locus 4009_at_x00 is close to 4023_s_at_x01, which is related to the resistance to 4-NQO. Locus 6453_at_x15 itself shows up in locus pairs for bleomycin. This suggest that the genes underlying those loci could contribute to drug resistance to both chemical agents, highlighting the potentially

important role of Imp2' and Rad28 in DNA repair pathways. For locus 9198_s_at_x06 in the pair (4009_at_x00, 9198_s_at_x06), we identified two genes Ddi3 and Rpd3. It is known that the expression level of Ddi3 is induced over 10-fold by DNA damage caused MMS[190]. Though Rpd3 is not traditionally thought to directly impact DNA repair pathway, the newest evidence has shown that it and other deacetylases have a key role in DNA repair response[191].

For locus pair (2276_at_x03, 10044_at_x09), we identified two potential causal genes. Mcm3 is close to one locus and Mcm5 is close to the other locus of the pair. Both are subunits of six-member MCM2-7 complex, a ring-shaped heterohexamer, which binds chromosomal replication origins. It has been suggested that the MCM complex could be crucial S-phase checkpoint targets for fork stabilization, and may also mediate DNA damage repair and related signaling[192]. The protein-protein interaction between Mcm3 and Mcm5 may explain the observed interacting locus pair.

Chapter 5.4: Conclusions

In this work we developed a method that utilizes allele specific interaction compatibility to predict locus interactions. Such allele specific compatibility might arise when interacting loci, in the same pathway co-evolved within each parental strain in a way that maintains a given phenotype. By linking phenotype change to specific inheritance pattern, we could identify allele specific interaction between loci.

Focusing on asymmetric loss of compatibility has been particularly fruitful. In fact we have not detected any symmetric compatibility loss. This is consistent with the finding of Litvin *et al.*[120] in the context of eQTL analysis. In their expression Quantitative Trait Loci (eQTL) study where gene expression was phenotype, they found that majority of two-locus interactions are allele-specific and asymmetric with regard to the primary locus, where "the secondary locus exerts an influence on the phenotype only when the primary locus has a particular allele (and has little or no influence when the primary locus has another allele)"[120]. Thus asymmetric loss of compatibility seems to be common for both gene expression phenotypes as well as more macroscopic phenotypes such as cell growth under drugs. In addition, the genotype/phenotype pattern for between pathways compatibility model is identical to the asymmetric within pathway model. Without additional information we will not be able to distinguish between two underlying models.

Applying LoCAp to a recent QTL study on yeast DNA repair phenotype responding to chemical agents[175], we were able to detect candidate locus interaction pairs for 4-NQO, bleomycin and caffeine. In particular, we predicted that the Rad5 locus is an interaction hub responsive to DNA damage and supported our prediction through experimental validation. Our prediction for other interacting locus pairs also show promise to be biologically relevant based on known physical or functional interactions between genes underlying the loci. In addition, although it is difficult to distinguish between pathway and within pathway using only computational approach, we were able to identify possible cases for between pathway and within pathway

based on biological knowledge. For example, locus pair (11041_at_x12, 10260_at_x09) (with underlying gene Pol31 and Rad5) in response to 4-NQO could fit between pathway model. Locus pair (2276_at_x03, 10044_at_x09) (with underlying gene Mcm3 an Mcm5) in response to caffeine could be an example of within pathway model.

In this study, we utilized drug resistance as the phenotype under study. It is likely that our method benefited from the fact that the strong selective pressure related to such phenotype[193,194] can potentially make the co-evolution of the phenotype determining loci particularly pronounced and more easily detectable. If a phenotype of interest is a subject of selection, it is more readily possible to obtain strains that acquired independently the given phenotype, making our method particularly well suited in such setting. However, uncovering interactions underlying drug resistance is one of the fundamental problems in human health. In fact our approach suggests a general technique for uncovering interactions related to drug resistance or other phenotype that might be a subject of selection. Our method provides an important tool to achieve this goal.

Chapter 5.5: Materials and Methods

Chapter 5.5.1: Datasets

We obtained DNA repair phenotype data from the recent study of the response of 123 spore progenies derived from a BYxRM yeast cross to four DNA damaging chemicals[175].The phenotype was measured as the sensitivity of progenies to several

DNA damaging agents. Saturated progeny culture was put onto plates containing various concentrations of DNA damaging chemical agents. The four DNA damaging chemicals were methyl methane sulfonate (MMS), 4-nitroquinoline 1-oxide (4-NQO), bleomycin, and caffeine. MMS methylates DNA on N7-deoxyguanine and N3-deoxyadenine, which can cause double-strand breaks. Since one of parental strains shows sensitivity to MMS, which does not meet the assumption of our hypothesis, we did not consider MMS in our study. 4-NQO can lead to DNA lesion and often stimulate nucleotide excision repair. Bleomycin is a chemotherapeutical cancer agent, acting by causing DNA strand break. Though the mechanism of caffeine in DNA damage is not clear, it is thought to interfere with DNA damage checkpoints instead directly inducing DNA damage[187]. The phenotype values of strains were recorded as follows: “very sensitive”, “sensitive”, “slightly sensitive”, “wild-type resistance”, “increased resistance” and “not tested”. We only considered those relatively strong phenotype values. Unless otherwise noted, we consider “very sensitive” and “sensitive” as sensitive phenotype and “wild-type resistance” and “increased resistance” as resistant phenotype in our study.

We obtained yeast genotype data for the same BYxRM cross from the study by Brem *et al*[195], covering 2,956 genetic markers. A genotype 0/1 indicates the parental strand the marker is inherited from RM/BY. To reduce the computational burden, we binned adjacent markers if the Hamming distance between their genotype data (without considering missing data) did not exceed 8 as proposed by[120]. The leftmost

locus in each bin was taken as the representative locus of the bin. This way, we reduced the number of markers (or loci) to 558.

Chapter 5.5.2: Main idea of the computational method

To identify pairs of interacting loci consistent with our model, we developed a graph theory based method that allows us to identify a small number of likely candidates which are subsequently subjected to statistical test. We label the two parents as 0 and 1 respectively. Consistently, we label the genotype of a locus in a progeny with 0 or 1, indicating parent the locus is inherited from. As explained in the main text and Figure 1, we search for pairs of loci such that the genotype 01 correspond to opposite phenotypes than genotype 10 and whose genotype 00 and 11 correspond to the parental phenotype.

We encode the relation between genotypes of progenies and phenotype with a bipartite graph. In this graph for each locus *li* there are two nodes *li0* and *li1* representing locus inherited the parent 0 and parent 1, respectively. Each progeny strain, *s,* is also represented by a node. An edge between *li0*/*li1* and *s* indicates that the locus *li* has the genotype 0/1 in the progeny *s*. Bipartite cliques in such graph provide a set of progenies sharing genotype combinations of any subset of loci in the clique. Thus identification of such cliques allows to point to a pair of loci with consistent genotype-phenotype patterns over some set of progenies. We searched for locus pairs consistent with model in two steps. We first utilized bipartite cliques to find locus pairs such that majority of progenies having 01 genotype for the locus pair have

opposite phenotype from majority of progenies having 10 genotype with the help of bipartite cliques. Then from those pairs we selected the pairs whose genotype 00 and 11 correspond to resistance phenotype.

More specifically, we identified maximal bipartite cliques meeting the following three conditions: (i) Each clique contains one locus from BY and one locus from RM (ii) The total number of progenies is at least *t*. (iii) The majority of progenies in the bipartite clique have the same phenotype as formalized in the model parameter description section. Cliques satisfying the three conditions will be called predominantly sensitive/resistant depending on the phenotype of majority of progenies in the clique. The main idea is to search for a locus pair (*li*, *lj*), which is associated with two cliques: predominantly sensitive for joint genotype 01 (or 10) and predominantly resistant for the opposite genotype 10 (or 01). After getting a small number of locus pairs this way, the second step is a filtering step. We required that, for each locus pair, the number of resistant progenies is not smaller than the number of sensitive progenies for its genotype 00 and 11. This is to ensure that selected candidate locus pairs fit our model hypothesis.

To evaluate the significance of our result, we ran the method with the same parameters on random data. We calculated *p*-value by comparing the number of loci obtained from random data to the one from real data.

Chapter 5.5.3: Accounting for linkage disequilibrium

Since nearby loci are often in linkage disequilibrium (LD), to account for such fact in practice, we would search for locus pairs (*li, lj*) so that they have joint genotype 01 or 10 in predominantly sensitive cliques and there exists a locus pair (*li', lj'*) with genotype 10 or 01 in predominantly resistant cliques, where there are at most *h* loci between two ends on *i , i' l l* (and ' , *j j l l*), where ' , *i i l l* is used to indicate the genomic region between locus *li'* and *li* including themselves. If we could find such locus pair (*li, lj*) and (*li', lj'*), it indicates that there are two interacting loci, one on region ' , *i i l l* and the other on region ' , *j j l l* .

Chapter 5.5.4: Description of model parameters

Since the number of bipartite cliques in a bipartite graph can be exponential with respect to the size of the graph, we applied an efficient algorithm to generate bipartite cliques with the number of progenies in the cliques not smaller than a threshold. The algorithm has been successfully applied in our previous *P. falciparum* eQTL study[196]. For computation with each chemical agent, we had to determine the threshold, *t*, of the number of progenies, in each bipartite clique to consider locus pairs. There is a tradeoff between this threshold and the number of candidate locus pairs we are going to find and subsequently subject to statistical testing. With a low threshold we are likely to recover many candidate locus pairs with large false discovery rate. Since each such candidate pair is then tested for statistical significance, generating and testing too many cliques is neither feasible nor desirable. Additionally,

due to our manual validation process, we want parameters to be stringent enough so that we obtain a small set of candidate locus pairs first. However, making parameters too stringent, we could end up without detecting any candidate locus pairs even when less stringent parameters would recover statistically significant pairs. Hence, we determined the minimal number of progenies in bipartite cliques by testing multiple thresholds and choosing the appropriate one as described below.

To select big enough cliques where the majority of strains in the clique show the same phenotype, we set the fraction of progenies showing majority phenotype in a clique to be larger than or equal to a threshold r.

It is important to keep in mind that our goal is to discover small number of locus pairs with small false discovery on the expense of large number of false negative. It is possible that different parameters will lead to different pairs that are also statistically significant.

We set t approximately equal to one quarter of total number of progenies since there are four genotype value combinations (00, 10, 01, and 11) for a pair of loci. Additionally, we set r to be 0.7. For 4-NQO, there are 84 progenies showing either resistant or sensitive phenotype. Accordingly, we set t to be 22 and h to be 1 and obtain four candidate locus pairs. For drug bleomycin, applying the same parameters used with 4-NQO study (t = 22, r = 0.7 and h = 1), we detected three candidate locus pairs.

For caffeine treatment, 22 progenies have sensitive phenotype and 50 progenies have resistant phenotype. By setting t = 20, r = 0.7 and h = 1, we did not

find any predominantly sensitive cliques. Finally, by changing $r = 0.65$, we obtained five candidate locus pairs.

Chapter 5.5.5: Testing significance

To test the significance of candidate locus pairs, we performed the following test. We randomly permuted the phenotype of progenies and kept their genotype data. Then we ran our method on such randomized data 10 times with the same parameters that we selected through the procedure described above. Finally, we compared the vector containing the numbers of "interacting" locus pairs found in random data with the one found in real data using Wilcoxon signed rank test to obtain *p*-value.

Chapter 5.5.6: Genetic interaction profiling of Rad5 and Ies3 in MMS using Epistatic Mini-Array Profiles (EMAP)

The rad5Δ and ies3Δ single mutants were constructed as previously described[41]. Double mutants were generated and scored for genetic interaction through the E-MAP technique using a previously defined protocol[10,111]. Haploid double mutants were ultimately grown on selective media containing 0.01% MMS for 72 hours.

Chapter 5.6: Acknowledgements

This work was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine (Y.H. and T.P.), and the

National Institute of Environmental Health Sciences (R01 ES014811; R.S. and T.I.). We would like to thank Dr. Ehrenreich and Dr. Kruglyak for kindly providing us the genomic location of 14 loci identified in their X-QTL study.

Chapter 5, in full, is a reprint of a manuscript currently in submission: “Huang Y., Srivas R., Guénolé A, van Attikum H, Krogan NJ, Ideker T, Przytycka TM. *Allele specific compatibility of locus interactions underlying yeast DNA repair phenotypes*. Genome Biology. In submission”. The dissertation author was the second author on this work, responsible for generation of all experimental data. For the sake of brevity all Supplemental Tables and Datasets have not been included here. These items can be found online at the publication’s website.

Table 5.1: Candidate locus interaction pairs detected for sensitivity phenotype to three chemical agents. Potential candidate causal genes close to each locus are also given

	candidate locus interaction pair (l_i, l_j)	candidate genes close to l_i	candidate genes close to lj
4_NQO	(4023_s_at_x01, 10322_at_x15)	Imp2' (YIL154C) Rrd1 (YIL153W)	Rad5 (YLR032W) Ies3 (YLR052W)
4_NQO	(7292_at_x05, 10322_at_x15)	Rdh54 (YBR073W)	Rad5 (YLR032W) Ies3 (YLR052W)
4_NQO	(11041_at_x12, 10260_at_x09)	Pol31 (YJR006W)	Rad5 (YLR032W) Ies3 (YLR052W)
4_NQO	(5038_at_x05, 10268_at_x01)	Ckb1 (YGL019W) Alk1 (YGL021W)	Rad5 (YLR032W) Ies3 (YLR052W)
Bleomycin	(7026_at_x11, 6453_at_x15)	Cdc28 (YBR160W)	Rad28 (YDR030C)
Bleomycin	(7222_at_x11, 6453_at_x15)	Mec1 (YBR136W)	Rad28 (YDR030C)
Bleomycin	(6453_at_x15, 8666_at_x04)	Rad28 (YDR030C)	Rfc4 (YOL094C) Msh2 (YOL090W)
Caffeine	(4009_at_x00, 9198_s_at_x06)	Imp2' (YIL154C) Rrd1 (YIL153W)	Ddi3(YNL335W) Rpd3 (YNL330C)
Caffeine	(6453_at_x15, 10831_s_at_x15)	Rad28 (YDR030C)	Mgm101 (YJR144W)
Caffeine	(2276_at_x03, 10044_at_x09)	Mms21 (YEL019C) Mcm3 (YEL032W)	Mec3 (YLR288C) Mcm5 (YLR274W)
Caffeine	(4075_at_x02, 10428_at_x12)	Irc24 (YIR036C)	Tor2 (YKL203C) Doa1 (YKL213C)
Caffeine	(10803_at_x05, 8707_at_x08)	Tor2 (YKL203C) Doa1 (YKL213C)	Pms1 (YNL082W)

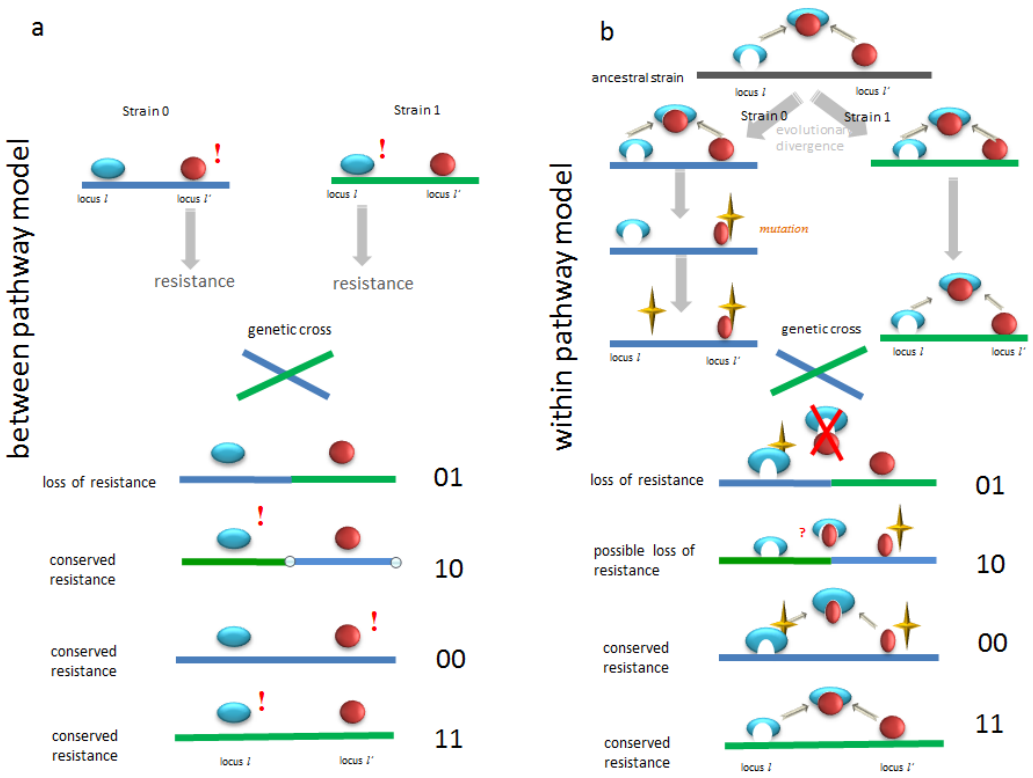


Figure 5.1: **Illustration of two models for allele specific locus interaction compatibility**

(A) Between pathway model: Two loci *l* and *l'* function in two parallel pathways, controlling a phenotype in a complementary way. In strain 0 (blue) the phenotype is mediated by one pathway involving locus *l'*. In strain 1 (green) the phenotype is mediated by the other pathway involving locus *l*. Then the progenies where *l* is from strain 0 and locus *l'* is from strain 1 will lose resistance (two loci are not complementary) (B) Within pathway model: Interacting loci *l* and *l'* in the same pathway have been a subject of compensatory mutations in strain 0. Mutated loci can still maintain their interaction in strain 0 (they are compatible in strain 0). However in a cross, where locus *l* comes from strain 0 and locus *l'* comes from strain 1 the interaction is not compatible. For the opposite case when locus *l* comes from strain 1 (green) and locus *l'* comes from strain 0 the loci may also be incompatible (symmetric case) or compatible (asymmetric case), indicated by a question mark.

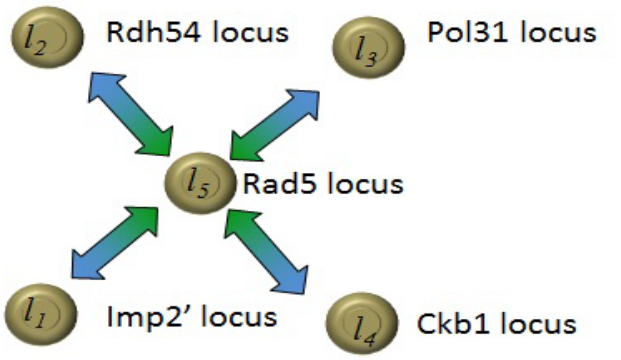


Figure 5.2: Four interacting locus pairs detected by LoCAp for yeast DNA repair phenotype in response to 4-NQO, where Rad5 locus appears in each pair.

Chapter 6. Conclusion

Despite nearly a decade's worth of progress in trying to map genetic variants responsible for complex human diseases such as diabetes, obesity, cardiovascular disease, and various neurological disorders, the variants identified, thus far, account for only a small fraction of the total heritability seen in these various diseases[197]. While there are multiple potential explanations for this so-called "missing heritability", epistatic interactions between genetic variants or loci have become one compelling explanation. However, the study of epistatic or genetic interactions in mammalian systems is complicated by the fact that, (i) generating deletions of genes remains a difficult task, rendering mammalian cell systems difficult to reverse genetic analysis[198], and (ii) mining forward genetic datasets (e.g., GWAS) is hampered by a lack of power due to limited sample sizes[199].

My work on the analysis of epistatic or genetic interactions in the model eukaryotic organism, *Saccharomyces cerevisiae*, suggests both potential challenges and potential suggestions for how to move forward in the analysis of genetic networks in humans. On the one hand, as described in Chapter 3, I observed that there was substantial re-wiring of genetic interactions in response to an external stimulus. More over, the changes in genetic interactions induced by the various genotoxic agents we examined were markedly different. Taken together, this suggests that genetic interactions are likely to be highly context dependent and will dramatically vary depending on the particular cell type (e.g. heart versus liver cells) or the microenvironment (e.g. hypoxic conditions) indicating a potentially large space of

interactions which will need to be tested in various mammalian cell lines. On the other hand, we described a computational method for organizing these differential genetic interactions in a map of protein modules and their inter-relationships (Chapter 2). While the relationships between modules and complexes were strikingly different, the modules identified in each condition were relatively stable. Thus, future combinatorial RNAi experiments may benefit by restricting the potential search space to include orthologs of the modules identified in this thesis.

In Chapters 4 and 5, I developed methods for analyzing genetic interaction data generated from forward genetic approaches. Our methods were able to boost our power for detecting genetic interactions as well provide a putative mechanism for the interaction by: (1) accounting for bi-cluster structure in the data and (2) by integrating genetic interactions derived from GWAS with protein complexes and functional annotations. Biologically and clinically, the clear and immediate application is towards the analysis of genome-wide association studies in humans. Many diseases, both common and rare, have so far been opaque to genome-wide association analysis[140]. The key question will be whether, using integrative maps such as those developed here, they can become less so.

References

1. Cusick, M.E., Klitgord, N., Vidal, M. & Hill, D.E. Interactome: gateway into systems biology. *Hum Mol Genet* **14 Spec No. 2**, R171-81 (2005).

2. Breitkreutz, A. *et al.* A global protein kinase and phosphatase interaction network in yeast. *Science* **328**, 1043-6.

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

4. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-6 (1989).

5. Gavin, A. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631-6 (2006).

6. Breitkreutz, B.J. *et al.* The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36**, D637-40 (2008).

7. Tong, A. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303**, 808-13 (2004).

8. Boone, C., Bussey, H. & Andrews, B. Exploring genetic interactions and networks with yeast. *Nat Rev Genet* **8**, 437-49 (2007).

9. Collins, S. *et al.* Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**, 806-10 (2007).

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

11. Pan, X. *et al.* A DNA integrity network in the yeast Saccharomyces cerevisiae. *Cell* **124**, 1069-81 (2006).

12. Breslow, D.K. *et al.* A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat Methods* **5**, 711-8 (2008).

13. Wilmes, G.M. *et al.* A genetic interaction map of RNA-processing factors reveals links between Sem1/Dss1-containing complexes and mRNA export and splicing. *Mol Cell* **32**, 735-46 (2008).

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

15. Fiedler, D. *et al.* Functional organization of the S. cerevisiae phosphorylation network. *Cell* **136**, 952-63 (2009).

16. Bandyopadhyay, S. *et al.* Rewiring of genetic networks in response to DNA damage. *Science* **330**, 1385-9 (2010).

17. Ideker, T. & Krogan, N.J. Differential network biology. *Mol Syst Biol* **8**, 565 (2012).

18. Bandyopadhyay, S., Kelley, R., Krogan, N. & Ideker, T. Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol* **4**, e1000065 (2008).

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

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

21. Luo, J. *et al.* A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell* **137**, 835-48 (2009).

22. Tischler, J., Lehner, B., Chen, N. & Fraser, A.G. Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. *Genome Biol* **7**, R69 (2006).

23. Harper, J.W. & Elledge, S.J. The DNA damage response: ten years after. *Mol Cell* **28**, 739-45 (2007).

24. Ziegler, A., König, I. & Thompson, J. Biostatistical aspects of genome-wide association studies. *Biom J* **50**, 8-28 (2008).

25. Carlborg, O. & Haley, C.S. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* **5**, 618-25 (2004).

26. Evans, D.M., Marchini, J., Morris, A.P. & Cardon, L.R. Two-stage two-locus models in genome-wide association. *PLoS Genet* **2**, e157 (2006).

27. Marchini, J., Donnelly, P. & Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* **37**, 413-7 (2005).

28. Storey, J., Akey, J. & Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **3**, e267 (2005).

29. Brem, R., Storey, J., Whittle, J. & Kruglyak, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**, 701-3 (2005).

30. Rockman, M.V. & Kruglyak, L. Genetics of global gene expression. *Nat Rev Genet* **7**, 862-72 (2006).

31. Brem, R.B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A* **102**, 1572-7 (2005).

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

33. Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nat Biotechnol* **24**, 537-44 (2006).

34. Tu, Z., Wang, L., Arbeitman, M.N., Chen, T. & Sun, F. An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics* **22**, e489-96 (2006).

35. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037-48 (1994).

36. Zuk, O., Hechter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* **109**, 1193-8 (2012).

37. Suter, B., Auerbach, D. & Stagljar, I. Yeast-based functional genomics and proteomics technologies: the first 15 years and beyond. *Biotechniques* **40**, 625-44 (2006).

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

39. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat Genet* **39**, 1217-24 (2007).

40. Bandyopadhyay, S. *et al.* Rewiring of genetic networks in response to DNA damage. *Science* **330**, 1385-9.

41. Schuldiner, M., Collins, S.R., Weissman, J.S. & Krogan, N.J. Quantitative genetic analysis in Saccharomyces cerevisiae using epistatic miniarray profiles (E-MAPs) and its application to chromatin functions. *Methods* **40**, 344-52 (2006).

42. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425-31.

43. Schlabach, M.R. *et al.* Cancer proliferation gene discovery through functional genomics. *Science* **319**, 620-4 (2008).

44. Bakal, C. *et al.* Phosphorylation networks regulating JNK activity in diverse genetic backgrounds. *Science* **322**, 453-6 (2008).

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

46. Sharma, V.M., Tomar, R.S., Dempsey, A.E. & Reese, J.C. Histone deacetylases RPD3 and HOS2 regulate the transcriptional activation of DNA damage-inducible genes. *Mol Cell Biol* **27**, 3199-210 (2007).

47. Jaimovich, A., Rinott, R., Schuldiner, M., Margalit, H. & Friedman, N. Modularity and directionality in genetic interaction maps. *Bioinformatics* **26**, i228-36.

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

49. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**, 2366-82 (2007).

50. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-2.

51. Roguev, A. *et al.* Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* **322**, 405-10 (2008).

52. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).

53. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448-9 (2005).

54. Ashkenazi, M., Bader, G.D., Kuchinsky, A., Moshelion, M. & States, D.J. Cytoscape ESP: simple search of complex biological networks. *Bioinformatics* **24**, 1465-6 (2008).

55. van Iersel, M.P. *et al.* The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* **11**, 5.

56. Saeed, A.I. *et al.* TM4 microarray software suite. *Methods Enzymol* **411**, 134-93 (2006).

57. Collins, S.R. *et al.* Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* **6**, 439-50 (2007).

58. Pu, S., Vlasblom, J., Emili, A., Greenblatt, J. & Wodak, S.J. Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics* **7**, 944-60 (2007).

59. Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S.J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* **37**, 825-31 (2009).

60. Hang, M. & Smith, M.M. Genetic Analysis Implicates the Set3/Hos2 Histone Deacetylase in the Deposition and Remodeling of Nucleosomes Containing H2A.Z. *Genetics*.

61. Kato, D. *et al.* Phosphorylation of human INO80 is involved in DNA damage tolerance. *Biochem Biophys Res Commun* **417**, 433-8 (2012).

62. Haider, S. *et al.* BioMart Central Portal--unified access to biological data. *Nucleic Acids Res* **37**, W23-7 (2009).

63. Jackson, S.P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461**, 1071-8 (2009).

64. Ciccia, A. & Elledge, S.J. The DNA damage response: making it safe to play with knives. *Mol Cell* **40**, 179-204 (2010).

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

66. Paulsen, R.D. *et al.* A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability. *Mol Cell* **35**, 228-39 (2009).

67. Gasch, A.P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**, 4241-57 (2000).

68. Zhou, B.B. & Elledge, S.J. The DNA damage response: putting checkpoints in perspective. *Nature* **408**, 433-9 (2000).

69. Longhese, M.P. DNA damage response at functional and dysfunctional telomeres. *Genes Dev* **22**, 125-40 (2008).

70. Norbury, C.J. & Zhivotovsky, B. DNA damage-induced apoptosis. *Oncogene* **23**, 2797-808 (2004).

71. Bjergbaek, L., Cobb, J.A., Tsai-Pflugfelder, M. & Gasser, S.M. Mechanistically distinct roles for Sgs1p in checkpoint activation and replication fork maintenance. *EMBO J* **24**, 405-17 (2005).

72. Tong, A.H. & Boone, C. Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods Mol Biol* **313**, 171-92 (2006).

73. Ulukan, H. & Swaan, P.W. Camptothecins: a review of their chemotherapeutic potential. *Drugs* **62**, 2039-57 (2002).

74. Lundin, C. *et al.* Methyl methanesulfonate (MMS) produces heat-labile DNA damage but no detectable in vivo DNA double-strand breaks. *Nucleic Acids Res* **33**, 3799-811 (2005).

75. Kats, E.S., Enserink, J.M., Martinez, S. & Kolodner, R.D. The Saccharomyces cerevisiae Rad6 postreplication repair and Siz1/Srs2 homologous recombination-inhibiting pathways process DNA damage that arises in asf1 mutants. *Mol Cell Biol* **29**, 5226-37 (2009).

76. Wu, M., Zhang, Z. & Che, W. Suppression of a DNA base excision repair gene, hOGG1, increases bleomycin sensitivity of human lung cancer cell line. *Toxicol Appl Pharmacol* **228**, 395-402 (2008).

77. Travesa, A. *et al.* DNA replication stress differentially regulates G1/S genes via Rad53-dependent inactivation of Nrm1. *EMBO J* (2012).

78. Caba, E., Dickinson, D.A., Warnes, G.R. & Aubrecht, J. Differentiating mechanisms of toxicity using global gene expression analysis in Saccharomyces cerevisiae. *Mutat Res* **575**, 34-46 (2005).

79. Shinohara, M., Sakai, K., Ogawa, T. & Shinohara, A. The mitotic DNA damage checkpoint proteins Rad17 and Rad24 are required for repair of double-strand breaks during meiosis in yeast. *Genetics* **164**, 855-65 (2003).

80. Usui, T., Ogawa, H. & Petrini, J.H. A DNA damage response pathway controlled by Tel1 and the Mre11 complex. *Mol Cell* **7**, 1255-66 (2001).

81. Liakopoulos, D., Doenges, G., Matuschewski, K. & Jentsch, S. A novel protein modification pathway related to the ubiquitin system. *EMBO J* **17**, 2208-14 (1998).

82. Bekker-Jensen, S. & Mailand, N. The ubiquitin- and SUMO-dependent signaling response to DNA double-strand breaks. *FEBS Lett* **585**, 2914-9 (2011).

83. Rabut, G. & Peter, M. Function and regulation of protein neddylation. 'Protein modifications: beyond the usual suspects' review series. *EMBO Rep* **9**, 969-76 (2008).

84. Laplaza, J.M., Bostick, M., Scholes, D.T., Curcio, M.J. & Callis, J. Saccharomyces cerevisiae ubiquitin-like protein Rub1 conjugates to cullin proteins Rtt101 and Cul3 in vivo. *Biochem J* **377**, 459-67 (2004).

85. Soucy, T.A., Dick, L.R., Smith, P.G., Milhollen, M.A. & Brownell, J.E. The NEDD8 Conjugation Pathway and Its Relevance in Cancer Biology and Therapy. *Genes Cancer* **1**, 708-16 (2010).

86. Ben-Aroya, S. *et al.* Proteasome nuclear activity affects chromosome stability by controlling the turnover of Mms22, a protein important for DNA repair. *PLoS Genet* **6**, e1000852 (2010).

87. Xirodimas, D.P. *et al.* Ribosomal proteins are targets for the NEDD8 pathway. *EMBO Rep* **9**, 280-6 (2008).

88. Harrison, J.C. & Haber, J.E. Surviving the breakup: the DNA damage checkpoint. *Annu Rev Genet* **40**, 209-35 (2006).

89. van Attikum, H., Fritsch, O. & Gasser, S.M. Distinct roles for SWR1 and INO80 chromatin remodeling complexes at chromosomal double-strand breaks. *EMBO J* **26**, 4113-25 (2007).

90. Zhang, H., Myshkin, E. & Waskell, L. Role of cytochrome b5 in catalysis by cytochrome P450 2B4. *Biochem Biophys Res Commun* **338**, 499-506 (2005).

91. Lisby, M., Rothstein, R. & Mortensen, U.H. Rad52 forms DNA repair and recombination centers during S phase. *Proc Natl Acad Sci U S A* **98**, 8276-82 (2001).

92. Huh, W.K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686-91 (2003).

93. Dyavaiah, M., Rooney, J.P., Chittur, S.V., Lin, Q. & Begley, T.J. Autophagy-dependent regulation of the DNA damage response protein ribonucleotide reductase 1. *Mol Cancer Res* **9**, 462-75 (2011).

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

95. Bandyopadhyay, S., Kelley, R., Krogan, N.J. & Ideker, T. Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol* **4**, e1000065 (2008).

96. Ulitsky, I., Shlomi, T., Kupiec, M. & Shamir, R. From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. *Mol Syst Biol* **4**, 209 (2008).

97. Srivas, R. *et al.* Assembling global maps of cellular function through integrative analysis of physical and genetic networks. *Nat Protoc* **6**, 1308-23 (2011).

98. Redon, C., Pilch, D.R. & Bonner, W.M. Genetic analysis of Saccharomyces cerevisiae H2A serine 129 mutant suggests a functional relationship between H2A and the sister-chromatid cohesion partners Csm3-Tof1 for the repair of topoisomerase I-induced DNA damage. *Genetics* **172**, 67-76 (2006).

99. Han, J. *et al.* Rtt109 acetylates histone H3 lysine 56 and functions in DNA replication. *Science* **315**, 653-5 (2007).

100. Prakash, S., Johnson, R.E. & Prakash, L. Eukaryotic translesion synthesis DNA polymerases: specificity of structure and function. *Annu Rev Biochem* **74**, 317-53 (2005).

101. Andersen, P.L., Xu, F. & Xiao, W. Eukaryotic DNA damage tolerance and translesion synthesis through covalent modifications of PCNA. *Cell Res* **18**, 162-73 (2008).

102. Masumoto, H., Hawke, D., Kobayashi, R. & Verreault, A. A role for cell-cycle-regulated histone H3 lysine 56 acetylation in the DNA damage response. *Nature* **436**, 294-8 (2005).

103. Collins, S.R. *et al.* Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**, 806-10 (2007).

104. Han, J., Zhou, H., Li, Z., Xu, R.M. & Zhang, Z. Acetylation of lysine 56 of histone H3 catalyzed by RTT109 and regulated by ASF1 is required for replisome integrity. *J Biol Chem* **282**, 28587-96 (2007).

105. Chang, D.J. & Cimprich, K.A. DNA damage tolerance: when it's OK to make mistakes. *Nat Chem Biol* **5**, 82-90 (2009).

106. Wei, D. *et al.* Radiosensitization of human pancreatic cancer cells by MLN4924, an investigational NEDD8-activating enzyme inhibitor. *Cancer Res* **72**, 282-93 (2012).

107. Lin, J.J., Milhollen, M.A., Smith, P.G., Narayanan, U. & Dutta, A. NEDD8-targeting drug MLN4924 elicits DNA rereplication by stabilizing Cdt1 in S phase, triggering checkpoint activation, apoptosis, and senescence in cancer cells. *Cancer Res* **70**, 10310-20 (2010).

108. Hartwell, L.H., Szankasi, P., Roberts, C.J., Murray, A.W. & Friend, S.H. Integrating genetic approaches into the discovery of anticancer drugs. *Science* **278**, 1064-8 (1997).

109. Perez-Enciso, M. Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. *Genetics* **163**, 1497-510 (2003).

110. Hastbacka, J. *et al.* The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell* **78**, 1073-87 (1994).

111. Collins, S.R., Roguev, A. & Krogan, N.J. Quantitative genetic interaction mapping using the E-MAP approach. *Methods Enzymol* **470**, 205-31 (2010).

112. Hoppins, S. *et al.* A mitochondrial-focused genetic interaction map reveals a scaffold-like complex required for inner membrane organization in mitochondria. *J Cell Biol* **195**, 323-40 (2011).

113. Zheng, J. *et al.* Epistatic relationships reveal the functional organization of yeast transcription factors. *Mol Syst Biol* **6**, 420 (2010).

114. van Attikum, H., Fritsch, O., Hohn, B. & Gasser, S.M. Recruitment of the INO80 complex by H2A phosphorylation links ATP-dependent chromatin remodeling with DNA double-strand break repair. *Cell* **119**, 777-88 (2004).

115. Chen, C. & Kolodner, R.D. Gross chromosomal rearrangements in Saccharomyces cerevisiae replication and recombination defective mutants. *Nat Genet* **23**, 81-5 (1999).

116. Johnson, R.E. *et al.* Identification of APN2, the Saccharomyces cerevisiae homolog of the major human AP endonuclease HAP1, and its role in the repair of abasic sites. *Genes Dev* **12**, 3137-43 (1998).

117. Hannum, G. *et al.* Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLoS Genet* **5**, e1000782 (2009).

118. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).

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

120. Litvin, O., Causton, H.C., Chen, B.J. & Pe'er, D. Modularity and interactions in the genetics of gene expression. *Proc Natl Acad Sci U S A* **106**, 6441-6 (2009).

121. Wall, J.D. & Pritchard, J.K. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* **4**, 587-97 (2003).

122. Mewes, H. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**, 31-4 (2002).

123. Christman, M.F., Dietrich, F.S. & Fink, G.R. Mitotic recombination in the rDNA of S. cerevisiae is suppressed by the combined action of DNA topoisomerases I and II. *Cell* **55**, 413-25 (1988).

124. Kressler, D., Linder, P. & de La Cruz, J. Protein trans-acting factors involved in ribosome biogenesis in Saccharomyces cerevisiae. *Mol Cell Biol* **19**, 7897-912 (1999).

125. Koehler, C.M. *et al.* Tim9p, an essential partner subunit of Tim10p for the import of mitochondrial carrier proteins. *Embo J* **17**, 6477-86 (1998).

126. Jungmann, J. & Munro, S. Multi-protein complexes in the cis Golgi of Saccharomyces cerevisiae with alpha-1,6-mannosyltransferase activity. *Embo J* **17**, 423-34 (1998).

127. Sacher, M., Barrowman, J., Schieltz, D., Yates, J.R., 3rd & Ferro-Novick, S. Identification and characterization of five new subunits of TRAPP. *Eur J Cell Biol* **79**, 71-80 (2000).

128. Iung, A.R. *et al.* Mitochondrial function in cell wall glycoprotein synthesis in Saccharomyces cerevisiae NCYC 625 (Wild type) and [rho(0)] mutants. *Appl Environ Microbiol* **65**, 5398-402 (1999).

129. Shimada, K. *et al.* Ino80 chromatin remodeling complex promotes recovery of stalled replication forks. *Curr. Biol.* **18**, 566-75 (2008).

130. Shen, X., Mizuguchi, G., Hamiche, A. & Wu, C. A chromatin remodelling complex involved in transcription and DNA processing. *Nature* **406**, 541-4 (2000).

131. Papamichos-Chronakis, M. & Peterson, C.L. The Ino80 chromatin-remodeling enzyme regulates replisome function and stability. *Nat. Struct. Mol. Biol.* **15**, 338-45 (2008).

132. Schwabish, M.A. & Struhl, K. Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. *Mol. Cell Biol.* **24**, 10111-7 (2004).

133. Ford, J., Odeyale, O., Eskandar, A., Kouba, N. & Shen, C.H. A SWI/SNF- and INO80-dependent nucleosome movement at the INO1 promoter. *Biochem. Biophys. Res. Commun.* **361**, 974-9 (2007).

134. Klopf, E. *et al.* Cooperation between the INO80 complex and histone chaperones determines adaptation of stress gene transcription in the yeast Saccharomyces cerevisiae. *Mol Cell Biol* **29**, 4994-5007 (2009).

135. Jonikas, M.C. *et al.* Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science* **323**, 1693-7 (2009).

136. Warner, J.R. Synthesis of ribosomes in Saccharomyces cerevisiae. *Microbiol Rev* **53**, 256-71 (1989).

137. Brauer, M.J. *et al.* Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Mol Biol Cell* **19**, 352-67 (2008).

138. Schadt, E.E. & Lum, P.Y. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *J Lipid Res* **47**, 2601-13 (2006).

139. Suthram, S., Beyer, A., Karp, R.M., Eldar, Y. & Ideker, T. eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol Syst Biol* **4**, 162 (2008).

140. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**, 241-51 (2009).

141. Cherry, J. *et al.* Genetic and physical maps of Saccharomyces cerevisiae. *Nature* **387**, 67-73 (1997).

142. Cohen, B.A., Mitra, R.D., Hughes, J.D. & Church, G.M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* **26**, 183-6 (2000).

143. Sahai, H. & Ageel, M.I. *The analysis of variance : fixed, random, and mixed models*, xxxv, 742 (Birkhäuser, Boston, 2000).

144. Storey, J.D. A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 479-498 (2002).

145. Schork, N.J. Genetically complex cardiovascular traits. Origins, problems, and potential solutions. *Hypertension* **29**, 145-9 (1997).

146. Koh, J.L. *et al.* DRYGIN: a database of quantitative genetic interaction networks in yeast. *Nucleic Acids Res* **38**, D502-7 (2010).

147. Baryshnikova, A. *et al.* Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat Methods* **7**, 1017-24 (2010).

148. Sinha, H., Nicholson, B.P., Steinmetz, L.M. & McCusker, J.H. Complex genetic interactions in a quantitative trait locus. *PLoS Genet* **2**, e13 (2006).

149. Ayyadevara, S. *et al.* Genetic loci modulating fitness and life span in Caenorhabditis elegans: categorical trait interval mapping in CL2a x Bergerac-BO recombinant-inbred worms. *Genetics* **163**, 557-70 (2003).

150. Wiltshire, S. *et al.* Epistasis between type 2 diabetes susceptibility Loci on chromosomes 1q21-25 and 10q23-26 in northern Europeans. *Ann Hum Genet* **70**, 726-37 (2006).

151. Ma, D.Q. *et al.* Association and gene-gene interaction of SLC6A4 and ITGB3 in autism. *Am J Med Genet B Neuropsychiatr Genet* **153B**, 477-83 (2010).

152. Abou Jamra, R. *et al.* The first genomewide interaction and locus-heterogeneity linkage scan in bipolar affective disorder: strong evidence of epistatic effects between loci on chromosomes 2q and 6q. *Am J Hum Genet* **81**, 974-86 (2007).

153. Zeng, Z.B., Wang, T. & Zou, W. Modeling quantitative trait Loci and interpretation of models. *Genetics* **169**, 1711-25 (2005).

154. Zhao, J., Jin, L. & Xiong, M. Test for interaction between two unlinked loci. *Am J Hum Genet* **79**, 831-45 (2006).

155. Nelson, M.R., Kardia, S.L., Ferrell, R.E. & Sing, C.F. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* **11**, 458-70 (2001).

156. Ritchie, M.D., Hahn, L.W. & Moore, J.H. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* **24**, 150-7 (2003).

157. Young, S.S. & Ge, N. Recursive partitioning analysis of complex disease pharmacogenetic studies. I. Motivation and overview. *Pharmacogenomics* **6**, 65-75 (2005).

158. Chen, X., Liu, C.T., Zhang, M. & Zhang, H. A forest-based approach to identifying gene and gene gene interactions. *Proc Natl Acad Sci U S A* **104**, 19199-203 (2007).

159. Zhang, Y. & Liu, J.S. Bayesian inference of epistatic interactions in case-control studies. *Nat Genet* **39**, 1167-73 (2007).

160. Park, M.Y. & Hastie, T. Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30-50 (2008).

161. Zhang, Z., Zhang, S., Wong, M.Y., Wareham, N.J. & Sha, Q. An ensemble learning approach jointly modeling main and interaction effects in genetic association studies. *Genet Epidemiol* **32**, 285-300 (2008).

162. Wan, X. *et al.* MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. *BMC Bioinformatics* **10**, 13 (2009).

163. Michaelson, J.J., Alberts, R., Schughart, K. & Beyer, A. Data-driven assessment of eQTL mapping methods. *BMC Genomics* **11**, 502 (2010).

164. Huang, Y., Siwo, G., Wuchty, S., Ferdig, M.T. & Przytycka, T.M. Symmetric Epistasis Estimation (SEE) and its application to dissecting interaction map of Plasmodium falciparum. *Mol Biosyst* **8**, 1544-52 (2012).

165. Phillips, P.C. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**, 855-67 (2008).

166. Cordell, H.J. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* **10**, 392-404 (2009).

167. Moore, J.H. & Williams, S.M. Epistasis and its implications for personal genetics. *Am J Hum Genet* **85**, 309-20 (2009).

168. Pattin, K.A. & Moore, J.H. Role for protein-protein interaction databases in human genetics. *Expert Rev Proteomics* **6**, 647-59 (2009).

169. Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D. & Cohen, F.E. Co-evolution of proteins with their interaction partners. *J Mol Biol* **299**, 283-93 (2000).

170. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* **14**, 609-14 (2001).

171. Juan, D., Pazos, F. & Valencia, A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci U S A* **105**, 934-9 (2008).

172. Pazos, F. & Valencia, A. Protein co-evolution, co-adaptation and interactions. *EMBO J* **27**, 2648-55 (2008).

173. Kann, M.G., Shoemaker, B.A., Panchenko, A.R. & Przytycka, T.M. Correlated evolution of interacting proteins: looking behind the mirrortree. *J Mol Biol* **385**, 91-8 (2009).

174. Heck, J.A. *et al.* Negative epistasis between natural variants of the Saccharomyces cerevisiae MLH1 and PMS1 genes results in a defect in mismatch repair. *Proc Natl Acad Sci U S A* **103**, 3256-61 (2006).

175. Demogines, A., Smith, E., Kruglyak, L. & Alani, E. Identification and dissection of a complex DNA repair sensitivity phenotype in Baker's yeast. *PLoS Genet* **4**, e1000123 (2008).

176. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-9 (2006).

177. Vijeh Motlagh, N.D., Seki, M., Branzei, D. & Enomoto, T. Mgs1 and Rad18/Rad5/Mms2 are required for survival of Saccharomyces cerevisiae mutants with novel temperature/cold sensitive alleles of the DNA polymerase delta subunit, Pol31. *DNA Repair (Amst)* **5**, 1459-74 (2006).

178. Hishida, T., Ohno, T., Iwasaki, H. & Shinagawa, H. Saccharomyces cerevisiae MGS1 is essential in strains deficient in the RAD6-dependent DNA damage tolerance pathway. *EMBO J* **21**, 2019-29 (2002).

179. Ehrenreich, I.M. *et al.* Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**, 1039-42 (2010).

180. Bhatia, P.K., Verhage, R.A., Brouwer, J. & Friedberg, E.C. Molecular cloning and characterization of Saccharomyces cerevisiae RAD28, the yeast homolog of the human Cockayne syndrome A (CSA) gene. *J Bacteriol* **178**, 5977-88 (1996).

181. Kim, H.S. & Brill, S.J. Rfc4 interacts with Rpa1 and is required for both DNA replication and DNA damage checkpoints in Saccharomyces cerevisiae. *Mol Cell Biol* **21**, 3725-37 (2001).

182. Majka, J. & Burgers, P.M. Yeast Rad17/Mec3/Ddc1: a sliding clamp for the DNA damage checkpoint. *Proc Natl Acad Sci U S A* **100**, 2249-54 (2003).

183. Antony, E. & Hingorani, M.M. Mismatch recognition-coupled stabilization of Msh2-Msh6 in an ATP-bound state at the initiation of DNA repair. *Biochemistry* **42**, 7682-93 (2003).

184. Enserink, J.M., Hombauer, H., Huang, M.E. & Kolodner, R.D. Cdc28/Cdk1 positively and negatively affects genome stability in S. cerevisiae. *J Cell Biol* **185**, 423-37 (2009).

185. Bode, A.M. & Dong, Z. The enigmatic effects of caffeine in cell cycle and cancer. *Cancer Lett* **247**, 26-39 (2007).

186. Wanke, V. *et al.* Caffeine extends yeast lifespan by targeting TORC1. *Mol Microbiol* **69**, 277-85 (2008).

187. Osman, F. & McCready, S. Differential effects of caffeine on DNA damage and replication cell cycle checkpoints in the fission yeast Schizosaccharomyces pombe. *Mol Gen Genet* **260**, 319-34 (1998).

188. Alvaro, D., Lisby, M. & Rothstein, R. Genome-wide analysis of Rad52 foci reveals diverse mechanisms impacting recombination. *PLoS Genet* **3**, e228 (2007).

189. Lee, M.W. *et al.* Global protein expression profiling of budding yeast in response to DNA damage. *Yeast* **24**, 145-54 (2007).

190. Jelinsky, S.A. & Samson, L.D. Global response of Saccharomyces cerevisiae to an alkylating agent. *Proc Natl Acad Sci U S A* **96**, 1486-91 (1999).

191. Robert, T. *et al.* HDACs link the DNA damage response, processing of double-strand breaks and autophagy. *Nature* **471**, 74-9 (2011).

192. Bailis, J.M. & Forsburg, S.L. MCM proteins: DNA damage, mutagenesis and repair. *Curr Opin Genet Dev* **14**, 17-21 (2004).

193. Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev* **74**, 417-33 (2010).

194. Aminov, R.I. & Mackie, R.I. Evolution and ecology of antibiotic resistance genes. *FEMS Microbiol Lett* **271**, 147-61 (2007).

195. Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752-5 (2002).

196. Huang, Y., Wuchty, S., Ferdig, M.T. & Przytycka, T.M. Graph theoretical approach to study eQTL: a case study of Plasmodium falciparum. *Bioinformatics* **25**, i15-20 (2009).

197. Eichler, E.E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446-50 (2010).

198. Aagaard, L. & Rossi, J.J. RNAi therapeutics: principles, prospects and challenges. *Adv Drug Deliv Rev* **59**, 75-86 (2007).

199. McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-69 (2008).